# Optimal radiological gallbladder lesion characterization by combining visual assessment with CT-based radiomics

Yin, Yunchao; Yakar, Derya; Slangen, Jules J G; Hoogwater, Frederik J H; Kwee, Thomas C; de Haas, Robbert J

*Published in:*
European Radiology

*DOI:*
10.1007/s00330-022-09281-6

European Radiology (2023) 33:2725–2734
https://doi.org/10.1007/s00330-022-09281-6

HEPATOBILIARY-PANCREAS

# Optimal radiological gallbladder lesion characterization by combining visual assessment with CT-based radiomics

Yunchao Yin[1] · Derya Yakar[1] · Jules J. G. Slangen[1] · Frederik J. H. Hoogwater[2] · Thomas C. Kwee[1] · Robbert J. de Haas[1]

Received: 28 May 2022 / Revised: 30 October 2022 / Accepted: 4 November 2022 / Published online: 25 November 2022


## Abstract

**Objectives** Differentiating benign gallbladder diseases from gallbladder cancer (GBC) remains a radiological challenge because they can appear very similar on imaging. This study aimed at investigating whether CT-based radiomic features of suspicious gallbladder lesions analyzed by machine learning algorithms could adequately discriminate benign gallbladder disease from GBC. In addition, the added value of machine learning models to radiological visual CT-scan interpretation was assessed.

**Methods** Patients were retrospectively selected based on confirmed histopathological diagnosis and available contrast-enhanced portal venous phase CT-scan. The radiomic features were extracted from the entire gallbladder, then further analyzed by machine learning classifiers based on Lasso regression, Ridge regression, and XG Boosting. The results of the best-performing classifier were combined with radiological visual CT diagnosis and then compared with radiological visual CT assessment alone.

**Results** In total, 127 patients were included: 83 patients with benign gallbladder lesions and 44 patients with GBC. Among all machine learning classifiers, XG boosting achieved the best AUC of 0.81 (95% CI 0.72–0.91) and the highest accuracy rate of 73% (95% CI 65–80%). When combining radiological visual interpretation and predictions of the XG boosting classifier, the highest diagnostic performance was achieved with an AUC of 0.98 (95% CI 0.96–1.00), a sensitivity of 91% (95% CI 86–100%), a specificity of 93% (95% CI 90–100%), and an accuracy of 92% (95% CI 90–100%).

**Conclusions** Machine learning analysis of CT-based radiomic features shows promising results in discriminating benign from malignant gallbladder disease. Combining CT-based radiomic analysis and radiological visual interpretation provided the most optimal strategy for GBC and benign gallbladder disease differentiation.

**Key Points**

- *Radiomic-based machine learning algorithms are able to differentiate benign gallbladder disease from gallbladder cancer.*
- *Combining machine learning algorithms with a radiological visual interpretation of gallbladder lesions at CT increases the specificity, compared to visual interpretation alone, from 73 to 93% and the accuracy from 85 to 92%.*
- *Combined use of machine learning algorithms and radiological visual assessment seems the most optimal strategy for GBC and benign gallbladder disease differentiation.*

**Keywords** Gallbladder neoplasms · Machine learning · Medical oncology · Tomography, spiral computed

**Abbreviations**

| | |
|---|---|
| GBC | Gallbladder cancer |
| IBSI | Image biomarker standardization initiative |
| XG | Extreme gradient |

## Introduction

Gallbladder cancer (GBC) is known to have a poor prognosis, with overall 5-year survival rates of only up to 13% [1–4]. This dismal prognosis of GBC can be explained by its non-

✉ Robbert J. de Haas
r.j.de.haas@umcg.nl

[1] Department of Radiology, Medical Imaging Center Groningen, University of Groningen, University Medical Center Groningen, PO Box 30001, 9700, RB Groningen, The Netherlands

[2] Department of Surgery, Section Hepato-Pancreato-Biliary Surgery and Liver Transplantation, University of Groningen, University Medical Center Groningen, PO Box 30001, 9700, RB Groningen, The Netherlands

specific symptoms, leading to a high number of patients in whom GBC is diagnosed at an advanced stage when surgery is not an option anymore [5]. However, in patients with T1b/T2 tumors undergoing radical resection, 5-year survival rates can be increased to 53% [4]. Thus, early detection of GBC is crucial to improve the survival rates of these patients. Furthermore, adequate characterization of gallbladder lesions (i.e., correct differentiation between benign and malignant entities) is very important, because GBC patients should be treated at specialized hepatobiliary hospitals.

Ultrasound is the primary imaging modality for gallbladder disease diagnosis, and CT and MRI have been used as additional imaging modalities to evaluate gallbladder lesions. However, differentiating benign gallbladder diseases such as chronic or xanthogranulomatous cholecystitis and adenomyomatosis from GBC remains a challenge because they can appear very similar on imaging [6–10]. In a recent study investigating the radiologist's ability to visually discriminate benign gallbladder disease from GBC based on CT scans, a relatively high sensitivity of 90% was achieved. However, the specificity was relatively low, merely approximately 60% [11]. In that study, irregular lesion aspect, absence of fat stranding, and locoregional lymphadenopathy were identified as predictors of GBC [11].

Radiomics is an emerging method for quantitative medical image analysis, which uses a high number of automatically extracted radiomic features for analysis. Radiomic features are thought to represent CT-based radiological lesion characteristics more accurately and objectively than a radiological judgment [12]. Liu et al used various machine learning methods to model the radiomic features for predicting survival outcomes of GBC patients [13]. Their study showed that CT-based radiomic features extracted from the gallbladder could distinguish high-risk patients with lower long-term survival from low-risk patients with better survival rates [13]. Given these results, we hypothesized that radiomics-based machine learning models could provide a more automatic and quantified differentiation between benign gallbladder disease and GBC that can be complementary to standard visual assessment by the radiologist.

The primary aim of the current study was to determine whether CT-based radiomic features of suspicious gallbladder lesions analyzed by machine learning algorithms, could adequately discriminate between benign gallbladder disease and GBC. The secondary aim was to investigate the additional value of machine learning models to radiological visual interpretation of CT scans in the same patient group.

## Materials and methods

### Study population

All patients referred to our hospital (which is a tertiary referral center) between January 2007 and October 2020 for suspicion of GBC or because of an incidentally found GBC after cholecystectomy, were included in the study. The patient exclusion diagram is shown in Fig. 1, including the following exclusion criteria: no contrast-enhanced portal venous phase CT scan available (for incidentally found GBC, CT had to be performed prior to cholecystectomy), and missing histopathological confirmation of the diagnosis. Reasons for suspicion of GBC

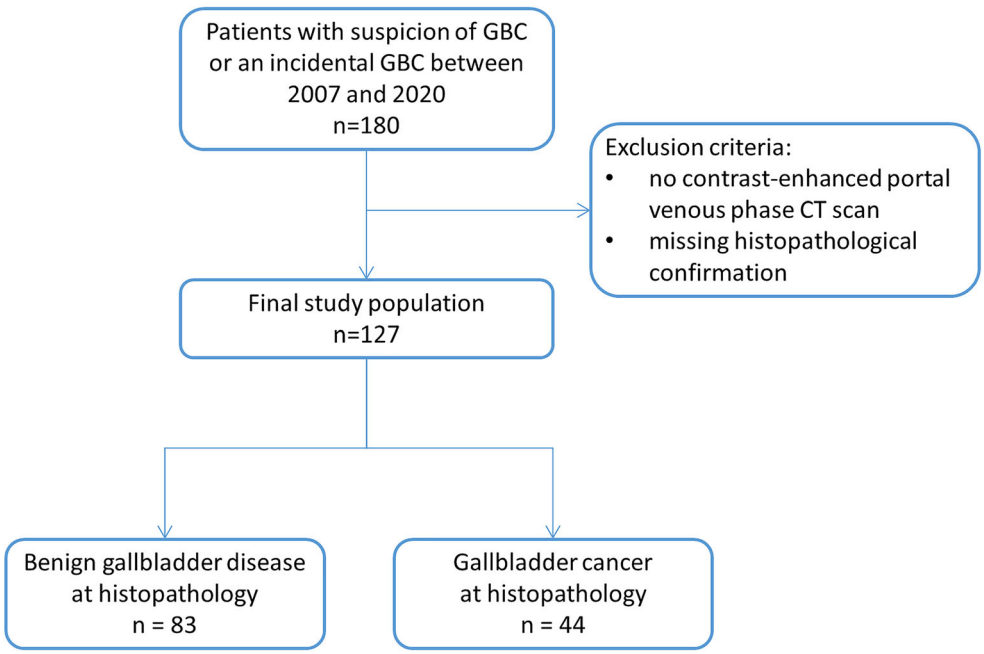


**Fig. 1** Flowchart study population

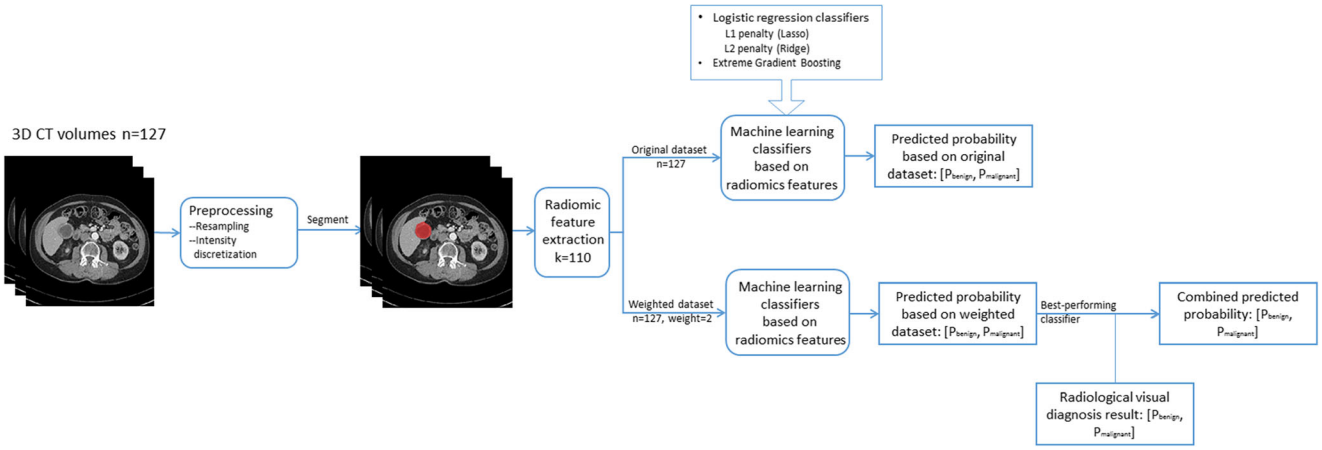


**Fig. 2** Workflow of the radiomic analysis for gallbladder disease characterization

and subsequent referral to our hospital were: a polyp with a diameter > 10 mm, a focal or diffuse wall thickening without obvious signs of benign disease, a mass lesion, or a porcelain gallbladder that has been considered to increase the risk of GBC [1]. Used CT systems were of multivendor origin. However, scan parameters were harmonized between our hospital and surrounding referring hospitals as follows: automatic tube current modulation and tube voltage selection, slice thickness of 1 mm, delay of 75 s after IV injection of 90–100 mL of contrast medium at a flow rate of 3.6–4.0 mL/s followed by 32 mL of saline solution. All patients were identified in a prospectively maintained surgical institutional database and analyzed retrospectively. In addition, to ensure the inclusion of all eligible cases, multidisciplinary team meeting lists were manually searched. Approval of the Institutional Review Board was obtained, and the need for written informed consent was waived. Part of the study population of the current study ($N$ = 118, 93%) was also the subject of a previous study [11].

Collected data included: patient age, gender, date, and type of surgery, date of CT, and histopathology results. Each resection and biopsy specimen underwent routine histopathological examination, performed by a specialized hepatobiliary pathologist.

## Image processing and radiomic feature extraction

The workflow of the radiomic analysis for gallbladder disease characterization is shown in Fig. 2. The 3D portal venous phase CT scans were used as initial input. To enhance the contrast among abdominal organs, a soft tissue window centering at 50 HU with a width of 400 HU was applied to the segmented gallbladder image. To normalize the CT scans, the images were resampled to the same spacing (0.7 mm, 0.7 mm, 1.0 mm) by BSpline interpolator, and the gray-level of the scan was discretized with a fixed bin width of 1.

The entire gallbladder was used for analysis, which was manually segmented by an abdominal radiologist using ITK-SNAP software, blinded to the final diagnosis [14]. Examples of benign gallbladder disease and GBC are shown in Figs. 3 and 4. In total, 110 radiomic features were extracted from the segmented 3D CT volume of the gallbladder according to the image biomarker standardization initiative (IBSI) [15], which included 18 first-order statistical features, 24 gray-level co-occurrence matrix features, 16 gray-level size zone matrix features, 16 gray level run length matrix features, 5 neighboring gray-tone difference matrix features, 14 gray level dependence matrix features, and 17 3D shape-based features. To minimize differences between CT scans and to optimize the

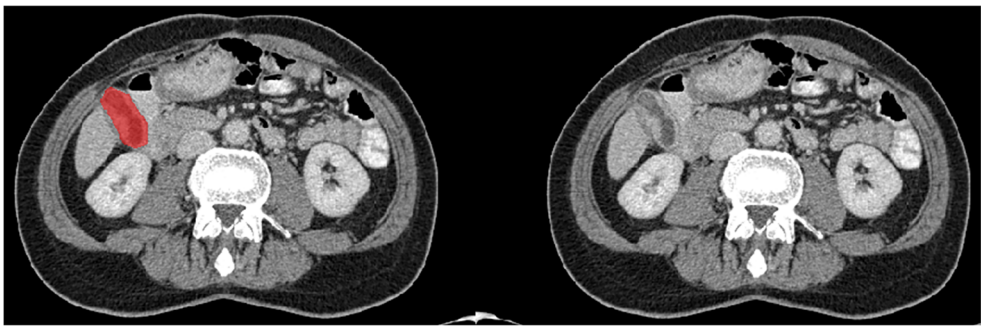

**Fig. 3** Example of a segmented gallbladder with gallbladder cancer on computed tomography in the axial direction

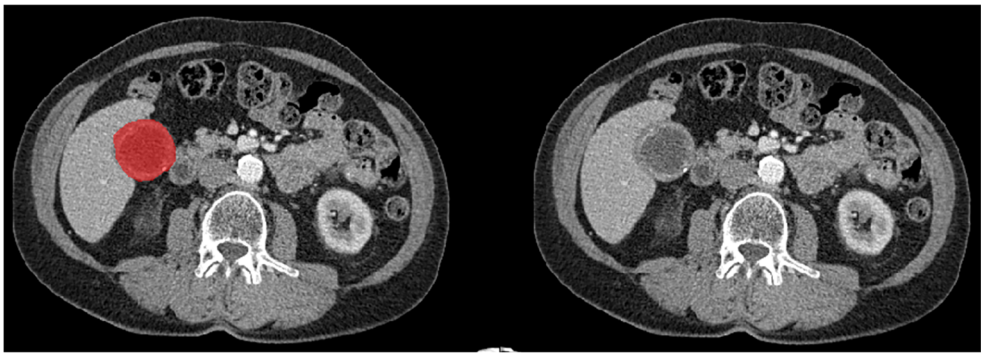

**Fig. 4** Example of a segmented gallbladder with benign gallbladder disease (chronic cholecystitis) on computed tomography in the axial direction

learning efficiency of machine learning classifiers, radiomic features were standardized by the standard scaler. Feature extraction was performed with Python 3.7.9 in the open-source library with Pyradiomics 3.0 [16].

## Machine learning classifiers

Due to the rarity of GBC, the size of the database was not sufficient for deep learning, requiring large numbers of images to be trained. Therefore, logistic regression with L1 penalties (Lasso regression) and L2 penalties (Ridge regression) was used, which are suitable for small-scale data analysis tasks [17]. No preselection of the radiomic features was performed, because the L1 and L2 penalties were used as the regularization of the logistic regression classifier, which can penalize high-valued regression coefficients to eliminate features that are redundant and reduce multi-collinearity in feature sets. The radiomic features that were correlated with benign gallbladder disease and GBC were automatically selected by logistic regression classifiers with L1 or L2 penalty during training. The feature importance for Lasso regression and Ridge regression was measured by the corresponding feature weights in the trained classifiers. Meanwhile, Extreme Gradient Boosting (XG boost) was also used. The XG boost classifier is constructed by decision trees, enabling the selection of the most powerful features to discriminate benign diseases from GBC at each split node [18]. The feature importance of the XG boost classifier was measured by Gini importance (mean decrease in impurity) [19].

The dataset was randomly split into the training set and test set by open-source library scikit-learn [20], thereby ensuring that the positive-negative ratio resembled the original dataset. The machine learning classifiers were trained by 80% of the patients based on five-fold cross-validation. The modulated machine learning models were then tested in the remaining 20% of the patients and the discrimination ability was represented by the AUC of the ROC curve, accuracy, sensitivity, and specificity values.

The datasets of patients with benign gallbladder diseases and GBC were imbalanced, which could influence machine learning classifiers' ability to learn from the minority class. To address this problem, we increased the class weight of the minority. Increasing the weight of patients with GBC can force the classifier to take the asymmetry of cost error between benign gallbladder disease and GBC into consideration. The model will be penalized more when misclassifying GBC during training. The trained model was tested by a dataset mimicking the prevalence of GBC in the original dataset so that the evaluation of performance is unbiased. The model was developed by the open-source library scikit-learn 0.23.2 with Python 3.7.9 [20].

## Combining machine learning results with radiological visual interpretation

In a previous study investigating the radiologists' ability to visually discriminate benign gallbladder disease from GBC, the specificity was relatively low [11]. Combining predictions provided by the machine learning classifier with the radiological diagnosis could possibly improve the specificity rates. As a result, the overall performance of discriminating between GBC and benign gallbladder disease could possibly also be improved.

To test this hypothesis, the five-point scale of the radiological visual interpretation (Appendix 1 illustrates the CT characteristics on which the radiological visual interpretation was based) performed by two radiologists after consensus reading [11] was converted into the probability of GBC, namely definitely benign = 0.0, probably benign = 0.25, equivocal = 0.5, probably GBC = 0.75, and definitely GBC = 1.0. This converted probability of the radiological visual interpretation and the predicted probability of GBC by the machine learning classifier were summed up with an equal weight of 0.5 as the combined prediction. The accuracy, sensitivity, specificity, and AUC were calculated accordingly from the combined prediction (a probability score $\geq 0.5$ was considered GBC).

## Radiomic analysis of the gallbladder combined with surrounding liver parenchyma

A recent study found that suspicion of invasion of adjacent liver parenchyma at CT was more frequently observed in

patients with GBC, compared to benign gallbladder disease [11]. Therefore, a subanalysis was performed including the radiomic features extracted from both the gallbladder and a surrounding rim of liver parenchyma to discriminate benign gallbladder disease from GBC. One hundred and ten hepatic radiomic features were extracted from a rim of 2 cm of liver parenchyma surrounding the gallbladder, which was automatically segmented and subsequently checked by an abdominal radiologist, blinded to the final diagnosis. The radiomic features extracted from the surrounding liver parenchyma were combined with the radiomic features extracted from the gallbladder for further analysis. Used machine learning models and performance metrics were similar to the other analyses in this study.

## Results

### Study population characteristics

In total, 127 patients were included in the study (Table 1 and Fig. 1). The median age of the patients was 66 years (interquartile range: 58–73 years), and among them, there were 80 women (63%) and 47 men (37%). Details on types of (surgical) treatment and results of histopathological examination are summarized in Table 1.

### Radiomics model

The XG boost classifier trained by the weighted dataset outperformed Lasso regression and Ridge regression, resulting in the highest AUC of 0.81 (95% CI 0.72–0.91) and an accuracy rate of 73% (95% CI 65–80%) in the test set for the differentiation between benign and malignant gallbladder disease. The results of the test set for different classifiers, the radiological visual interpretation, and the combined results of the radiological visual interpretation and the XG boosting classifier are shown in Table 2, and the corresponding ROC curves are plotted in Fig. 5.

Although the AUC of the classifiers trained by the original dataset seem decent, the sensitivity rates varied between 18 (95% CI 0–29%) and 36% (95% CI 22–57%). After applying techniques to address the class imbalance problem, the sensitivity rates of all classifiers largely improved, varying between 55 (95% CI 38–75%) and 64% (95% CI 50–83%) (Table 2).

**Table 1** Demographics table

| Characteristic | | Patient number |
|---|---|---|
| Median age, IQR | 66 (58–73) | 127 (100%) |
| Gender | Female | 80 (63%) |
| | Male | 47 (37%) |
| Benign gallbladder disease | Acute cholecystitis | 1 (1%) |
| | Chronic cholecystitis | 49 (39%) |
| | Xanthogranulomatous cholecystitis | 6 (5%) |
| | Adenoma | 4 (3%) |
| | Adenomyomatosis | 15 (12%) |
| | Porcelain gallbladder | 2 (2%) |
| | Other benign entities | 6 (5%) |
| Gallbladder cancer | Adenocarcinoma | 37 (29%) |
| | Adenosquamous carcinoma | 3 (2%) |
| | High-grade dysplasia | 2 (2%) |
| | Other types of malignancy | 2 (2%) |
| Types of (surgical) treatment | Open cholecystectomy | 42 (33%) |
| | Laparoscopic cholecystectomy | 13 (10%) |
| | Cholecystectomy combined with resection of liver segment 4/5 | 5 (4%) |
| | Cholecystectomy combined with a wedge resection of the liver parenchyma | 41 (32%) |
| | Cholecystectomy combined with extensive surgery* | 4 (3%) |
| | Cholecystectomy combined with lymphadenectomy | 6 (5%) |
| | Open-closure procedure | 10 (8%) |
| | Biopsy without any further operation | 6 (5%) |

Abbreviation: *IQR* interquartile range

* e.g. ≥ 3 liver segments, and/or pancreaticoduodenectomy

**Table 2** Radiomic analysis results

| Classifier | Methods to deal with class imbalance | AUC (95% CI) | Accuracy (95% CI) | Sensitivity (95% CI) | Specificity (95% CI) |
|---|---|---|---|---|---|
| Logistic Regression (L1 penalty)* | Original dataset | 0.77 (0.67, 0.88) | 0.65 (0.55, 0.75) | 0.27 (0.13, 0.43) | 0.93 (0.90, 1.00) |
| | Adding class weight | 0.76 (0.66, 0.88) | 0.65 (0.55, 0.75) | 0.64 (0.50, 0.83) | 0.67 (0.55, 0.82) |
| Logistic Regression (L2 penalty)# | Original dataset | 0.77 (0.68, 0.88) | 0.62 (0.50, 0.75) | 0.18 (0.00, 0.29) | 0.93 (0.90, 1.00) |
| | Adding class weight | 0.75 (0.64, 0.86) | 0.65 (0.55, 0.75) | 0.55 (0.38, 0.75) | 0.73 (0.60, 0.90) |
| XG Boosting | Original dataset | 0.72 (0.62, 0.86) | 0.69 (0.60, 0.80) | 0.36 (0.22, 0.57) | 0.93 (0.90, 1.00) |
| | Adding class weight | 0.81 (0.72, 0.91) | 0.73 (0.65, 0.80) | 0.64 (0.50, 0.83) | 0.80 (0.70, 0.92) |
| Combined results of XG Boosting and radiological diagnosis | Radiological diagnosis on weighted dataset | 0.94 (0.90, 0.95) | 0.85 (0.80, 0.95) | 1.00 (1.00, 1.00) | 0.73 (0.60, 0.90) |
| | Weighted XG boost with radiological diagnosis | 0.98§ (0.96, 1.00) | 0.92^ (0.90 1.00) | 0.91 (0.86, 1.00) | 0.93 (0.90, 1.00) |

*Lasso regression; #Ridge regression; § $p < 0.001$ when comparing the AUC of the radiological diagnosis with the combined diagnostic results; ^ $p < 0.001$ when comparing the accuracy of the radiological diagnosis with the combined diagnostic results

Abbreviations: *AUC* area under the receiver operating characteristic curve, *CI* confidence interval

The top 5 most important features and the total number of features used by the corresponding classifiers are shown in Table 3. The Ridge regression and XG boosting classifier made use of various features, while Lasso regression used only few shape-based features. The full list of radiomic features with corresponding weights assigned by each classifier can be found in Appendices 2–4 (weighted dataset).

## Combined radiomic analysis of the gallbladder and surrounding liver parenchyma

When using the XG boosting classifier trained by the weighted dataset for discriminating between benign gallbladder disease and GBC, an AUC of 0.77 (95% CI 0.67–0.88), an accuracy rate of 72% (95% CI 63–80%), sensitivity rate of 71% (95% CI 57–83%), and specificity rate of 73% (95% CI 60–90%) were achieved. Lasso regression achieved an AUC of 0.71 (95% CI 0.62–0.88) and an accuracy of 69% (95% CI 60–80%), and Ridge regression achieved an AUC of 0.70 (95% CI 0.62–0.86) and an accuracy of 0.58 (95% CI 0.46–0.70). These results did not outperform the radiomic analysis based solely on the gallbladder (Table 2).

## Combined prediction of the machine learning model and radiological visual interpretation

Using the five-point scale, the radiological visual interpretation resulted in a sensitivity of 100%, a specificity of 73% (95% CI 60–90%), an accuracy of 85% (95% CI 80–95%), and an AUC of 0.94 (95% CI 0.90–0.95) (Table 2).

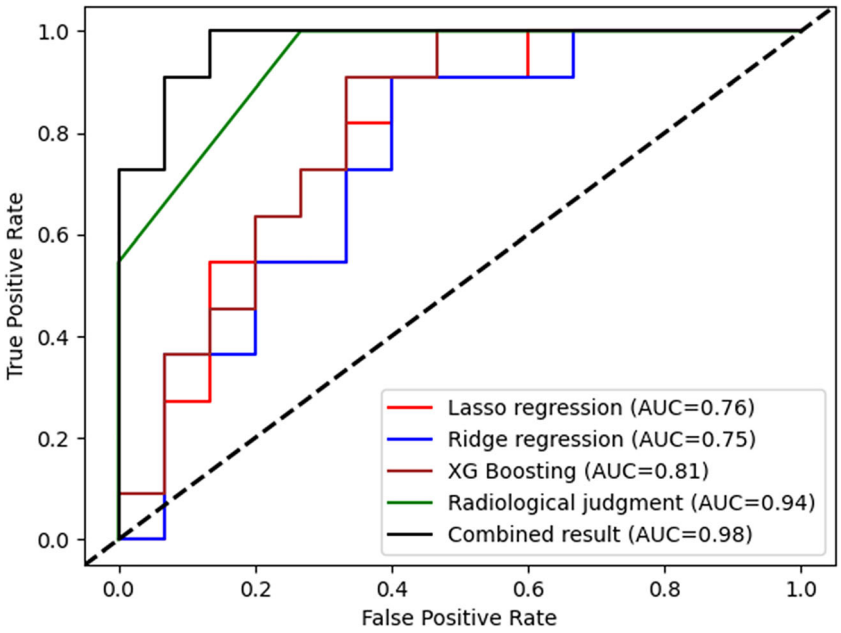


**Fig. 5** Receiver operating characteristic (ROC) curves of machine learning classifiers, radiological diagnosis by visual interpretation, and the combined results of the radiological judgment and XG boosting classifier. The area under the ROC curve (AUC) was significantly higher when combining radiological judgment with the XG boosting classifier, compared with radiological judgment alone ($p < 0.001$)

**Table 3** Top 5 important features of machine learning classifiers based on the weighted dataset

| Weight ranking | Lasso regression | | | | Ridge regression | | | | XG Boosting | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Benign gallbladder disease | | Malignant gallbladder disease | | Benign gallbladder disease | | Malignant gallbladder disease | | | |
| | Feature category | Feature name | Feature category | Feature name | Feature category | Feature name | Feature category | Feature name | Feature category | Feature name |
| 1st | Shape | Surface volume ratio | Shape | Compactness2 | Shape | Surface Volume Ratio | Shape | Voxel volume | Shape | Surface Volume Ratio |
| 2nd | - | - | Shape | Voxel volume | NGTDM | Strength | GLRLM | Run length non-uniformity | Shape | Voxel Volume |
| 3rd | - | - | Shape | Maximum 2D diameter column | NGTDM | Contrast | Shape | Maximum 2D diameter column | GLSZM | Low Gray Level Zone Emphasis |
| 4th | - | - | - | - | NGTDM | Coarseness | GLSZM | Size zone non-uniformity | GLSZM | Small Area Low Gray Level Emphasis |
| 5th | - | - | - | - | GLCM | Contrast | GLDM | Dependence non-uniformity | GLCM | Joint Average |
| Total number | 1 | | 3 | | 59 | | 50 | | 49 | |

Abbreviations: *NGTDM* neighbouring gray-tone difference matrix, *GLRLM* gray-level run length matrix, *GLSZM* gray-level size zone matrix, *GLCM* gray-level co-occurrence matrix, *GLDM* Gray Level Dependence Matrix

Given the performance results of the XG boosting classifier trained by radiomic features of the gallbladder compared to the other machine learning classifiers, we combined the radiological visual interpretation with the results of the XG boosting classifier trained by the weighted dataset. This way, the diagnostic accuracy of the radiological visual interpretation increased from 85 to 92% (95% CI 90–100%), and the AUC increased from 0.94 to 0.98 (95% CI 0.96–1.00) (both $p < 0.001$; Table 2). In addition, the specificity of the radiological visual interpretation increased from 73 to 93% (95% CI 0.90–1.00) after combining the results of the XG boosting classifier.

## Discussion

In the current study, our hypothesis that radiomics-based machine learning models provide more automatic and quantified differentiation between benign gallbladder disease and GBC and that it can be complementary to standard visual radiological assessment was tested.

In a total of 127 patients, 83 patients with benign gallbladder lesions and 44 patients with GBC, XG boosting achieved the best AUC of 0.81 and the highest accuracy rate of 73%. In addition, when combining radiological visual interpretation and CT-based radiomic features, the highest diagnostic performance was achieved with an AUC of 0.98, a sensitivity of 91%, a specificity of 93%, and an accuracy of 92%. Of note, the performance is based on a relatively small test set, namely 20% of our dataset.

To our knowledge, this is the first study using radiomic features analyzed by machine learning methods to discriminate between benign gallbladder diseases and GBC, as well as the first study combing radiomics with standard visual assessment. Compared with the visual interpretation of experienced radiologists [11, 21–23], radiomic-based machine learning evaluated the difference between benign gallbladder disease and GBC in a more quantitative way by using automatically calculated radiographic features from the CT scan [12]. However, the results of the radiomic-based machine learning classifiers trained by the original dataset showed high specificity rates and low sensitivity rates due to an imbalance of the data between benign gallbladder disease and GBC. After increasing the class weight of the minority during model training, the XG boost classifier obtained decent and balanced sensitivity and specificity rates.

In this study, the entire gallbladder including the gallbladder wall and bile was used as a volume of interest. As shown in Fig. 3, some GBC patients only have a small amount of bile in their gallbladder, which could introduce a difference between benign gallbladder disease and GBC. However, this can also be the case in patients with chronic cholecystitis. The same holds true for the presence of gallstones. Thus, the exact influence on the discriminating ability of the radiomics analysis when including the bile (and sludge and/or stones) in the region of interest, in other words, their potential contribution to the intensity, gray value distribution, and/or shape of the gallbladder on which radiomics analysis is based, is unclear. Besides, a precise segmentation of only the gallbladder wall and the lesion is difficult and very time-consuming.

The visual discrimination of benign gallbladder disease and GBC at CT scans of patients with suspicious gallbladder lesions by a radiologist resulted in high sensitivity rates of 100% but a relatively low specificity of 73%. After combining the results of the radiomic-based machine learning methods with the radiological diagnosis, the specificity rate increased to 93%. The accuracy and AUC also improved by 7% and 4%, respectively, compared with the radiological judgment alone. Therefore, at this moment, the most ideal interpretation of CT scans of patients with suspicious gallbladder lesions seems to be a combination of visual interpretation by a radiologist and radiomic-based machine learning analysis. This might lead to the more adequate characterization of gallbladder lesions, and thereby, an improvement of patient survival and better use of specialized hepatobiliary healthcare.

Often, the liver parenchyma surrounding the gallbladder is involved in the case of GBC, which might be an important clue in the differential diagnosis when a gallbladder lesion is present [9, 24]. Therefore, a subanalysis was performed in which a rim of 2 cm of liver parenchyma around the gallbladder was included in the CT-based radiomic analysis. However, adding surrounding liver parenchyma to the radiomic-based models did not improve the performance significantly compared to the results based only on the gallbladder. Perhaps, the parenchymal invasion of GBC might be too small to be reflected as a difference in texture features. Another explanation could be that inflammatory conditions such as cholecystitis can lead to infiltration of the surrounding liver parenchyma resulting in a similar aspect as can be seen in GBC [25].

Regarding the important features used by machine learning classifiers, Lasso regression only used a small number of features, which could indicate that Lasso regression mainly uses morphological changes to differentiate GBC from benign gallbladder diseases. Besides, various features are used by Ridge regression and XG boosting classifier as illustrated by Table 3 and Appendices 2–4, and the XG boosting classifier achieved better performance than the other radiomic-based machine learning methods. This could indicate that including different radiographic information can be helpful and complicated models such as XG boosting classifier are more capable to exploit useful information from high dimensional features for discriminating benign gallbladder disease from GBC.

Due to the fact that GBC is a relatively rare disease and the single-center study design, the current study was based on a relatively small population. To overcome this problem, future

studies should be multicentric in design. Ideally, deep learning methods should also be used when evaluating gallbladder lesions in large study populations. In addition, the impact of including bile (and sludge and/or stones) in the gallbladder segmentation on the discriminating ability of machine learning methods should be subject of future research, as this is currently unknown.

In conclusion, machine learning analysis of radiomic features shows promise to discriminate benign gallbladder lesions from malignant gallbladder disease. In addition, the combination of CT-based radiomic analysis and radiological visual interpretation provided the best results. More specifically, the radiomic models seem to recompense the low specificity of radiological visual assessment, thereby optimizing the ability to differentiate between benign and malignant gallbladder lesions.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-09281-6.

**Acknowledgements** We would like to thank Kim B. Mouridsen, MSc, PhD, for technical support.

**Funding** The authors state that this work has not received any funding.

## Declarations

**Guarantor** The scientific guarantor of this publication is Dr Robbert J. de Haas.

**Conflict of interest** The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

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

**Informed consent** Written informed consent was waived by the Institutional Review Board.

**Ethical approval** Institutional Review Board approval was obtained.

**Methodology**
- retrospective
- observational
- performed at one institution

## References

1. Lazcano-Ponce EC, Miquel JF, Muñoz N et al (2001) Epidemiology and molecular pathology of gallbladder cancer. CA Cancer J Clin 51:349–364
2. Siegel RL, Miller KD, Jemal A (2017) Cancer Statistics, 2017. CA Cancer J Clin 67:7–30
3. Lau CSM, Zywot A, Mahendraraj K, Chamberlain RS (2017) Gallbladder Carcinoma in the United States: a population based clinical outcomes study involving 22,343 patients from the surveillance, epidemiology, and end result database (1973-2013). HPB Surg 2017:1532835
4. de Savornin LE, de Bitter T, Verhoeven R et al (2020) Trends in treatment and survival of gallbladder cancer in the Netherlands; identifying gaps and opportunities from a nation-wide cohort. Cancers (Basel) 12:918
5. Duffy A, Capanu M, Abou-Alfa GK et al (2008) Gallbladder cancer (GBC): 10-year experience at Memorial Sloan-Kettering Cancer Centre (MSKCC). J Surg Oncol 98:485–489
6. Kimura K, Fujita N, Noda Y et al (2004) Localized wall thickening of the gallbladder mimicking a neoplasm. Dig Endosc 16:54–57
7. Zemour J, Marty M, Lapuyade B, Collet D, Chiche L (2014) Gallbladder tumor and pseudotumor: diagnosis and management. J Visc Surg 151:289–300
8. Elsayes KM, Oliveira EP, Narra VR, El-Merhi FM, Brown JJ (2007) Magnetic resonance imaging of the gallbladder: spectrum of abnormalities. Acta Radiol 48:476–482
9. Chang BJ, Kim SH, Park HY et al (2010) Distinguishing xanthogranulomatous cholecystitis from the wall-thickening type of early-stage gallbladder cancer. Gut Liver 4:518–523
10. Liang JL, Chen MC, Huang HY et al (2009) Gallbladder carcinoma manifesting as acute cholecystitis: clinical and computed tomographic features. Surgery 146:861–868
11. Kuipers H, Hoogwater FJH, Holtman GA, Slangen JJG, de Haas RJ, de Boer MT (2021) Diagnostic performance of preoperative CT in differentiating between benign and malignant origin of suspicious gallbladder lesions. Eur J Radiol 138:109619
12. Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441–446
13. Liu Z, Zhu G, Jiang X et al (2020) Survival prediction in gallbladder cancer using CT based machine learning. Front Oncol 10:604288
14. Yushkevich PA, Piven J, Hazlett HC et al (2006) User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31:1116–1128
15. Zwanenburg A, Vallières M, Abdalah MA et al (2020) The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295:328–338
16. van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. Cancer Res 77:e104–e107
17. Panesar SS, D'Souza RN, Yeh FC, Fernandez-Miranda JC (2019) Machine learning versus logistic regression methods for 2-year mortality prognostication in a small, heterogeneous glioma database. World Neurosurg X 2:100012
18. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016.
19. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinformatics 9:307
20. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
21. Ching BH, Yeh BM, Westphalen AC, Joe BN, Qayyum A, Coakley FV (2007) CT differentiation of adenomyomatosis and gallbladder cancer. AJR Am J Roentgenol 189:62–66
22. Lee ES, Kim JH, Joo I, Lee JY, Han JK, Choi BI (2015) Xanthogranulomatous cholecystitis: diagnostic performance of

US, CT, and MRI for differentiation from gallbladder carcinoma. Abdom Imaging 40:2281–2292
23. Jang JY, Kim SW, Lee SE et al (2009) Differential diagnostic and staging accuracies of high resolution ultrasonography, endoscopic ultrasonography, and multidetector computed tomography for gallbladder polypoid lesions and gallbladder cancer. Ann Surg 250: 943–949
24. Bang SH, Lee JY, Woo H et al (2014) Differentiating between adenomyomatosis and gallbladder cancer: revisiting a comparative study of high-resolution ultrasound, multidetector CT, and MR imaging. Korean J Radiol 15:226–234
25. Ratanaprasatporn L, Uyeda JW, Wortman JR, Richardson I, Sodickson AD (2018) Multimodality imaging, including dual-energy CT, in the evaluation of gallbladder disease. Radiographics 38:75–89