World Journal of Emergency Surgery

RESEARCH

Open Access

An explainable predictive machine learning model of gangrenous cholecystitis based on clinical data: a retrospective single center study

Ying Ma[1†], Man Luo[2†], Guoxin Guan[1], Xingming Liu[1], Xingye Cui[1] and Fuwen Luo[1*]

Abstract

Background Gangrenous cholecystitis (GC) is a serious clinical condition associated with high morbidity and mortality rates. Machine learning (ML) has significant potential in addressing the diverse characteristics of real data. We aim to develop an explainable and cost-effective predictive model for GC utilizing ML and Shapley Additive explanation (SHAP) algorithm.

Results This study included a total of 1006 patients with 26 clinical features. Through 5-fold CV, the best performing integrated learning model, XGBoost, was identified. The model was interpreted using SHAP to derive the feature subsets WBC, NLR, D-dimer, Gallbladder width, Fibrinogen, Gallbladder wallness, Hypokalemia or hyponatremia, these subsets comprised the final diagnostic prediction model.

Conclusions The study developed a explainable predictive tool for GC at an early stage. This could assist doctors to make quick surgical intervention decisions and perform surgery on patients with GC as soon as possible.

Key Summary

- Using clinical data from 1006 cholecystitis patients, we developed a machine learning-based diagnostic prediction model to help identify patients at high risk for acute gangrenous cholecystitis.
- During the study, the deficiency and imbalance of actual clinical data were directly addressed, leading to the ultimate selection of the integrated learning model XGBoost as the predictive model exhibiting superior performance and stability on a novel, unidentified validation set and compared to preoperative clinical diagnosis.
- The model employs variables that are non-specific, readily available, reasonably priced, and appropriate for clinical generalization.

†Ying Ma and Man Luo contributed equally to this work and co-first-authors.

*Correspondence:
Fuwen Luo
fuwenluo@aliyun.com

Full list of author information is available at the end of the article

Keywords Gangrenous cholecystitis, Machine learning, Integrated learning, Data imbalance, Diagnostic predictive model

Introduction

Gangrenous cholecystitis (GC) is a prevalent variety of complicated cholecystitis, which is characterized by progressive ischemia, necrosis, and even perforation of the gallbladder wall, and is the most often encountered of complicated cholecystitis [1–3]. GC is distinguished by a quick advancement of the disease, a high fatality rate and a poor prognosis [4]. It is reported that the incidence rate of GC is about 10-40% of all acute cholecystitis, and the morbidity and mortality rates range from about 15-50% [5]. Since this disease can only be diagnosed through pathology evidence, it is challenging to make a preoperative diagnosis for patients with GC, research indicates that only approximately 9% of individuals with GC are accurately diagnosed before surgery [6]. The current clinical Tokyo guideline (2018 edition) defines acute cholecystitis into three levels, with GC being designated as a second level. However, it does not provide a comprehensive summary of the diagnosis and therapy specifically connected to GC [7]. Hence, it is crucial to develop a prognostic diagnostic model to assist doctors in diagnosing and making informed decisions.

Multiple researches have been done to gather clinical data for the purpose of analyzing the risk factors of GC and attempting to develop an early prediction model. In a retrospective study, Raffee et al. discovered that GC patients were predominantly male, they also observed an increased risk of GC in patients with raised levels of erythrocyte sedimentation rate (ESR), leukocytes, and neutrophil ratio (neutrophil/leukocyte), as well as a low lymphocyte ratio (lymphocyte/leukocyte) [8]. Binit et al. found that GC had higher Hu values of gallbladder wall and bile in plain CT, which combined with the Hu values of the gallbladder wall and bile lumen, was used for the prediction of GC with 100% sensitivity and 75% specificity when the threshold value of 35 Hu was reached, but the sample size of this study was only 23 [9]. In contrast, Mok et al. primarily investigated the predictive capacity of C-reactive protein, they found that when C-reactive protein levels exceeded 200 mg/dL, the positive predictive value was approximately 50%. Additionally, the sensitivity and specificity were 100% and 87.9%, respectively [10]; KeeHwan Kim et al. constructed a prognostic model utilizing age, gender, blood cell count, liver function, and gallbladder wall thickness in CT scans, the model achieved a sensitivity and specificity of 74% [11]. The majority of previous studies employed conventional logistic regression models to predict diseases and had limited sample sizes, posing a difficulty for developing effective models. Simultaneously, the real-world gathering of clinical data on such severely ill patients is incomplete and imbalanced; past studies have similarly neglected these problems.

Artificial intelligence (AI), particularly machine learning (ML), has become extensively utilized in the medical fields in recent times due to its robust computational capacity. This allows it to not only analyze the correlation between numerous predictor variables and outcome variables, but also determine the nonlinear relationship between each predictor [12–14]. The current prediction models for GC are constructed using typical logistic regression models, this restricts the investigation of more complex interactions among multivariate inputs and illness outcomes [15]. The ML model can effectively utilize the various clinical data with multiple parameters to extract crucial information. This enables the construction of a highly accurate model with a low rate of false negatives. Consequently, clinicians can promptly identify and monitor patients suspected of having GC in an early stage. This early detection allows for timely intervention, such as surgery, which ultimately reduces the mortality rate associated with the disease.

Therefore, the objective of this research was to create an cost-effective predictive model for GC utilizing a ML technique and to design a clinical decision support tool by leveraging the model's interpretability. The study can enhance the care and management of GC patients by offering physicians more precise and evidence-based clinical decision support.

Materials and methods

Study design and patient selection

This retrospective cohort study was approved by the Ethics Committee of the Second Affiliated Hospital of Dalian Medical University (KY2024-006-02) and followed the Declaration of Helsinki. The consent form was waived by the review board due to the retrospective nature and deidentification of the data. This was a single-center, retrospective, and observational study on patients admitted to our center who were diagnosed with cholecystitis through ICD-9 code recognition and underwent cholecystectomy from January 2015 to May 2023, and the data from January 2015 to December 2022 was used for model training, while the data from January 2023 to May 2023 was used as an unknown external test set to test the performance of the model. This retrospective cohort study was registered with the ClinicalTrials (https://register.clinicaltrials.gov/prs/app/action/LoginUser?ts=1&cx=-jg9qo4). Data has been reported in line with STARD [16] and STROCSS [17] criteria.

Identification of research variables and data collection

To alleviate the impact of invalid variables on model calculations, it is imperative to find variables that are scientifically legitimate. Initially, we employed the subsequent search algorithm on the pubmed: (((((((predict) OR predictive factors) OR risk assessment) OR diagnosis) AND gangrene) OR gangrenous) AND cholecystitis) AND (("1980/01/01"[PDAT]: "2023/09/31"[PDAT])) to search for relevant literature over the years, remove duplicates to select statistically significant variables in each study, and then combine with the real-time existing items in the testing system of our center to delete and eventually incorporate the following studies Variables: Baseline parameters included: gender, age, body mass index (BMI), history of hypertension, diabetes mellitus, history of cardiovascular and cerebrovascular diseases (such as coronary heart disease and stroke), history of anticoagulants use, and history of abdominal general anesthesia. The laboratory indicators were: WBC, Neutrophil(NEU), Lymphocyte(LYM), Platelet, Neutrophil to Lymphocyte ratio(NLR), Platelet to lymphocyte ratio(PLR), Aspartate aminotransferase (AST), Alanine transaminase(ALT), Gamma-glutamyltransferase(GGT), Total bilirubin, D-dimer, Fibrinogen, and the presence of hypokalemia or hyponatremia upon admission. The imaging indices assessed in the abdominal CT or ultrasound report conducted after admission are the length, width, and wall thickness of the gallbladder. Other pertinent clinical information includes the body temperature and heart rate recorded upon admission. In order to compare the results between the GC and No-GC groups, we also collected the duration of surgery, intraoperative blood loss, and days of hospitalization. All the patients did not use antibiotics before admission, all the blood-related tests and imaging data were completed before the use of antibiotics.

The main reference standard for the preoperative diagnosis of GC by doctors at our center is from the Tokyo Guidelines 2018 edition [7]. GC was defined by the presence of surgical indications such as gangrene or perforation of the gallbladder wall that could be observed without the use of magnification, as well as confirmation through pathological examination.

The following are the inclusion and exclusion criteria for data collection:

Inclusion criteria: (1) patients diagnosed with acute cholecystitis or acute exacerbation of chronic cholecystitis in our hospital and receiving complete clinical treatment in our hospital; (2) performing cholecystectomy; (3) having complete and searchable clinical data, such as patient's age, surgical records, and hospitalization days.

Exclusion criteria: (1) patients who have previously diagnosed with chronic cholecystitis, this time for elective surgical treatment; (2) patients who have previously been diagnosed with acute cholecystitis and have undergone Percutaneous transhepatic gallbladder drainage (PTGBD), this time for elective laparoscopic cholecystectomy; (3) patients who have other acute biliary and pancreatic system-related diseases, such as obstructive jaundice caused by choledochal stones, acute cholangitis, acute pancreatitis, etc. (4) patients who underwent additional surgeries such as choledochotomy and lithotripsy, choledochoscopic exploration and lithotripsy, bile-intestinal anastomosis, appendectomy, etc.; (5) those with incomplete data; A total of 1006 patients were included in the study, with 109 of them used as an external test set.

Data transformation and normalization

The data utilized for constructing ML models must exhibit data integrity, with no missing values, and standardization in terms of scale. The predictor variables with missing values in this study are D-dimer, BMI, gallbladder imaging data. The missing values for D-dimer and gallbladder imaging data are primarily due to variations in the practices of different doctors in recommending medical tests. Additionally, some patients were referred from emergency rooms or other hospitals, and during their second visit to our hospital, relevant tests were not conducted to avoid duplication. Furthermore, the results of tests conducted in other hospitals are not recorded in our hospital's system. The missing values for BMI are a result of some patients being bedridden and unable to have their body weight measured. The aforementioned causes for missing data are random and varied, make it unsuitable to use mean interpolation or plurality interpolation, and multiple interpolation is currently one of the predominant methods extensively applied, which utilized multi-chain equations to interpolate missing values [18]. We used the MICE package in R4.2.3 [19] to interpolate the raw data and produce the complete dataset according to the approved criterias. Furthermore, the recently developed MIDASpy package by Lall, a researcher from the UK, utilizes deep learning to interpolate missing values. The authors of the study demonstrated the superiority of this technology by comparing its interpolation results with those of other multiple interpolation programs, such as MICE [20]. As a result, we also used interpolation to estimate the missing data based on the approved criteria. The missing values are all continuous numerical variables, to measure the disparity between the interpolated complete dataset and the original dataset, we emloyed the Normalized Root Mean Square Error (NRMSE) and selected the complete dataset with the smallest difference for subsequent model construction. Finally, the data standardization was performed using z-score standardization, a regularly used and widely applied method. This was done to remove the impact of scale differences between variables and to enhance the

computational convergence of the model [21]. The formula was as follows: where x is the original value of the sample, μ is the score mean of the overall sample, and σ is the standard deviation of the overall sample:

$$z = (x - \mu)/\sigma \quad (1)$$

Model construction and feature selection

This study used Stratified 5-fold cross validation to verify that the distribution of data in the training and validation sets is consistent with that of the original data [22]. Based on the correlation analysis results from the preceding part, Decision Tree, SVM, Random Forest(RF), XGBoost, AdaBoost were constructed by excluding the covariates and assessing the models' fit using them. A combination of Grid Search and manual parameterization was used to optimize the model's parameters. Additionally, the learning curve of the model was plotted to assess the fit of models, the feature variables were ranked using RFECV to rank the tuned models [23]. The SHAP value is a technique employed to interpret deep learning models by quantifying the feature variables within these models, which are often considered as "black boxes" [24]. The SHAP module contained interpreters capable of interpreting tree models like RF and XGBoost, so we generated the SHAP values for the XGBoost model to determine the importance ranking of the variables in the model. We then combined this ranking with the importance ranking of the feature variables in the RF and SVM models to synthesize and select a subset of features. The aforementioned procedures were executed utilizing the scikit-learn package in Python (version 3.9, Python Software Foundation, https://www.python.org/).

Data enhancement and model evaluation metrics

This study is a binary classification problem where the number of observations in the No-GC group is significantly greater than that in the GC group. This situation, where one class has more observations than the other, is referred to as data imbalance [25]. In fact, the medical field has encountered data imbalance issues, such as cancer detection, identification of rare symptoms, etc., which requires picking out only positive samples from a huge pool of normal values. This work involved the construction of the Random Under-Sampling, Random Over-Sampling, and Synthetic Minority Over-Sampling Technique (SMOTE) algorithms, along with it respective versions: Borderline-SMOTE, SVMSMOTE to resample the data. For the imbalanced dataset, we used Balanced accuracy [26] and the standard metrics of the confusion matrix: Recall, Precision, F1 score for evaluation. The specific formulas for these metrics are provided below (See Supplementary Table 1, Additional file 1). Given that our model's primary objective is to accurately detect positive samples, often known as GC, our main focus should be on maximizing the Recall value. Furthermore, we also generated the precision-recall (PR) curve for each model. This metric is particularly responsive to the minority class of positive samples when dealing with imbalanced data [27].

To facilitate the comparison between the final model and the preoperative diagnosis, Accuracy, Specificity, Sensitivity, and AUROC (area under the receiver operating characteristic curve) were calculated.

Statistical analysis

Continuous numerical variables were first analyzed by Shapiro to determine if they conformed to the normal distribution, if they did, the mean±standard deviation was used, otherwise, the median (interquartile spacing) was used to describe them. Percentages were used to characterize subtypes of variables; Chi-square tests were employed to assess the disparities between two groups for all categorical variables; Two-tailed t-tests were used for continuous numerical variables if they were normally distributed, and the Kruskal-Wallis rank-sum test was used for them otherwise; To exclude the effects of covariates in constructing the model, the correlation test between variables was used as follows: spearman correlation coefficient was used between continuous variables; Cramer's V correlation coefficient was used between categorical variables; point-biserial correlation coefficient was used between continuous variables and categorical variables. The above statistical analyses were performed using Rstudio version 4.2.3. A p value of <0.05 was considered to denote statistical significance.

Results

Clinical characteristics of patients

The workflow of this study was shown in Fig. 1. 897 patients were included in the study, divided into No-GC group ($n=689$) and GC group ($n=208$), with a total of 468 (52.2%) male patients and 429 (47.8%) female patients. In the GC group, patients were found to be older (64 vs. 56 years, $p<0.001$), predominantly male (68.3% vs. 31.7%), and had higher WBC (13.49 vs. 6.82, $p<0.001$) and NLR (13.01 vs. 2.56, $p<0.001$); elevated D-dimer (1.02 vs. 0.54, $p<0.001$) and higher NLR (13.01 vs. 2.56, $p<0.001$) and increased fibrinogen (5.61 vs. 3.71, $p<0.001$) demonstrating that coagulation was also affected; liver function was also elevated but only the difference in total bilirubin was statistically significant (24.30 vs. 15.28, $p<0.001$); admission was more likely to be associated with hypokalemia or hyponatremia (58.1% vs. 41.9%, $p<0.001$); and admission to the hospital was more likely to be associated with hypokalemia (58.1% vs. 41.9%, $p<0.001$); In addition, the gallbladder was likely to have an overall enlarged

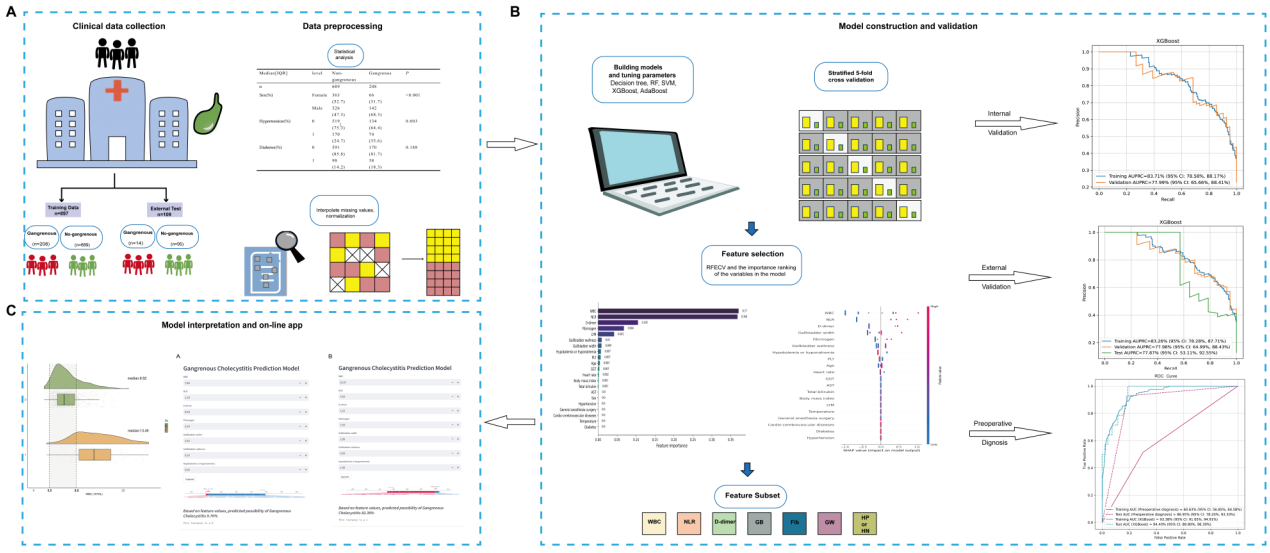

Fig. 1 The workflow of this study and workflow of the data analysis

gallbladder wall thickening on imaging (0.50 vs. 0.30 cm, $p<0.001$) (Table 1). Past medical history revealed that the No-GC group had a greater propotion of patients with a previous occurrence of cardiovascular disease and who were now using anticoagulant medications. Furthermore, some additional variables related to an outcome that we collected showed that the GC group had a longer operative time (110 vs. 75 min, $p<0.001$), more intraoperative blood loss (10 vs. 5 mL, $p<0.001$), and a longer hospital stay (8 vs. 7 days, $p<0.001$). These findings emphasize the importance of promptly and accurately predicting the occurrence of GC.

Removing covariates and interpolating missing values

A correlation analysis was performed on the predictive variables of the model, which included age, WBC, NEU, LYM, NLR, PLT, PLR, ALT, AST, GGT, Total bilirubin, D-dimer, Fibrinogen, BMI, Temperature, Heart rate, Gallbladder length, Gallbladder width, Gallbladder wallness, Sex, Hypertension, Diabetes, Cardio cerebrovascular diseases, Anticoagulant drugs, General anesthesia surgery, Hypokalemia or hyponatremia and the dependent variable Gangrenous. The results demonstrated a strong correlation between WBC and NEU, NLR, ALT and AST, and between cardiovascular disease and anticoagulant drug history (Fig. 2A). Considering that the NLR has demonstrated high research value in several previous studies and is highly correlated with the dependent variable, and based on the p value magnitude of the previous statistical analyses, we decided to keep the NLR and removed NEU, ALT, PLR, Gallbladder length, Anticoagulant drugs.The correlations between the remaining variables were also confirmed using heatmaps (Fig. 2B). Following that, we interpolated the missing values by referring to the approved criterias of MICE, MIDASpy two program packages, each program package generating 10 complete datasets and calculating the mean of NRMSE of each missing variable, as shown in the following table (See Supplementary Table 2, Additional file 1) we chose MIDASpy2 to construct the subsequent ML model.

Model performance

The models achieved a Balanced accuracy ranging from 77.49% (95% CI: 70.67-85.78%) to 83.20% (95% CI: 76.31-90.14%), while the Recall values were 59.63% (95% CI: 47.22-75.02%)-88.00% (95% CI. 84.00-100.00%) (Table 2). These models have excellent performance on different metrics respectively, SVM achieves the highest Balanced accuracy, RF has the highest Recall value and XGBoost model has strong performance across multiple metrics.

The RF model had moderate performance with a PR curve and AUPRC of (75.51%, 95% CI: 62.42-86.02%), on the other hand, SVM exhibited the smallest disparity between the training and validation sets (79.24% vs. 78.48%). This suggests that the SVM model may not be suffering from either overfitting or underfitting. The XGBoost and AdaBoost models showed good performance on the training set, but there was a noticeable difference in performance compared to the validation set (83.71% vs. 77.99%, 83.25% vs. 77.34%), suggesting that the models may be at danger of overfitting (Fig. 3). Consequently, we graphed the learning curves of several models mentioned above to detect the current level of model fitting (See Supplementary Fig. 1, Additional file 2). To avoid excessive false positives, we used F1 scores as the learning curve scores, and most of the

Table 1 Comparison of clinical characteristics of cholecystitis patients between different cohorts

	level	No-gangrenous	Gangrenos	*P*
n		689	208	
Sex(%)	Female	363 (52.7)	66 (31.7)	< 0.001
	Male	326 (47.3)	142 (68.3)	
Hypertension(%)	0	519 (75.3)	134 (64.4)	0.003
	1	170 (24.7)	74 (35.6)	
Diabetes(%)	0	591 (85.8)	170 (81.7)	0.188
	1	98 (14.2)	38 (18.3)	
CCD(%)	0	630 (91.4)	164 (78.8)	< 0.001
	1	59 (8.6)	44 (21.2)	
Anticoagulant drugs(%)	0	669 (97.1)	186 (89.4)	< 0.001
	1	20 (2.9)	22 (10.6)	
General Anesthesia surgery(%)	0	437 (63.4)	126 (60.6)	0.507
	1	252 (36.6)	82 (39.4)	
Hypokalemia Or hyponatremia(%)	0	604 (87.7)	100 (48.1)	< 0.001
	1	85 (12.3)	108 (51.9)	
Age, median[IQR], year		56.00 [45.00, 65.00]	64.00 [55.00, 71.00]	< 0.001
WBC, median[IQR], 10^9/L		6.82 [5.30, 9.21]	13.49 [10.13, 17.53]	< 0.001
NEU, median[IQR], 10^9/L		4.32 [2.99, 7.11]	11.95 [8.36, 15.39]	< 0.001
LYM, median[IQR], 10^9/L		1.59 [1.16, 2.05]	0.98 [0.71, 1.30]	< 0.001
NLR, median[IQR]		2.56 [1.67, 5.27]	13.01 [7.21, 18.70]	< 0.001
PLT, median[IQR], 10^9/L		231.00 [190.50, 274.00]	204.50 [161.75, 253.00]	< 0.001
PLR, median[IQR]		148.86 [112.11, 196.67]	208.76 [148.75, 288.90]	< 0.001
ALT, median[IQR], U/L		26.92 [18.00, 44.90]	31.15 [22.08, 56.15]	0.800
AST, median[IQR], U/L		23.80 [18.62, 33.20]	29.20 [21.08, 45.23]	0.490
GGT, median[IQR], U/L		29.70 [18.00, 76.78]	47.05 [26.00, 125.27]	0.055
Total bilirubin median[IQR], μmol/L		15.28 [11.07, 22.40]	24.30 [16.13, 37.00]	< 0.001
D-dimer median[IQR], ug/mL		0.54 [0.39, 0.77]	1.02 [0.66, 1.65]	< 0.001
Fibrinogen, median[IQR], ug/mL		3.71 [3.12, 4.52]	5.61 [4.28, 7.24]	< 0.001

Table 1 (continued)

level	No-gangrenous	Gangrenos	*P*
BMI, median[IQR], kg/m2	24.80 [22.86, 27.68]	25.72 [23.44, 27.77]	0.060
Temperature, median[IQR],℃	36.40 [36.10, 36.50]	36.50 [36.20, 36.90]	< 0.001
Heart rate, median[IQR], bpm	78.00 [72.00, 85.00]	82.00 [73.00, 96.00]	< 0.001
Operation time, median[IQR], min	75.00 [55.00, 100.00]	110.00 [80.00, 150.75]	< 0.001
Intraoperative Bloodloss, median[IQR], mL	10.00 [5.00, 20.00]	30.00 [10.00, 50.00]	< 0.001
Length of Stay, median[IQR], day	7.00 [5.00, 9.00]	8.00 [6.00, 10.00]	< 0.001
Gallbladder Length, median[IQR], cm	7.80 [6.00, 8.30]	8.90 [7.80, 10.00]	< 0.001
Gallbladder width, median[IQR], cm	3.30 [2.40, 4.00]	3.90 [3.30, 4.50]	< 0.001
Gallbladder wallness, median[IQR], cm	0.35 [0.30, 0.50]	0.50 [0.40, 0.60]	< 0.001

CCD: Cardio cerebrovascular diseases

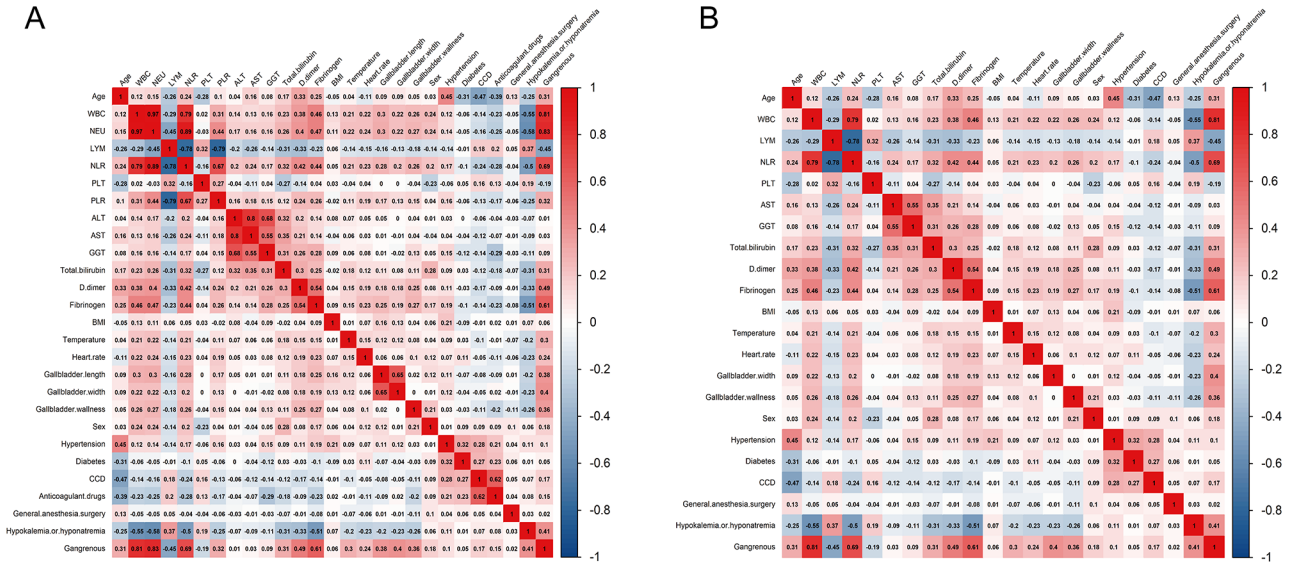

Fig. 2 The correlation heatmap between the variables. (**A**) Correlation Heatmap of all variables. (**B**) Correlation heatmap after removing collinear variables. CCD, Cardio cerebrovascular diseases

Table 2 The evaluation indicators for each model

Model	Balanced accuracy	Recall	Precision	F1 score
Decision Tree	80.46%(78.57–89.92%)	87.04%(80.00-98.12%)	50.31%(43.74–67.23%)	63.73%(57.73–78.26%)
SVM	83.20%(76.31–90.14%)	82.66%(67.65–91.67%)	60.79%(51.77–78.44%)	69.94%(59.99–80.41%)
Random Forest	82.75%(81.94–91.81%)	88.00%(84.00-100.00%)	54.25%(47.83–73.02%)	67.09%(63.41–81.63%)
XGBoost	82.49%(78.79–91.90%)	82.69%(70.58–94.12%)	58.85%(55.17–80.44%)	68.61%(63.63–83.73%)
AdaBoost	77.49%(70.67–85.78%)	59.63%(47.22–75.02%)	79.39%(66.67–92.59%)	68.07%(56.60–80.00%)

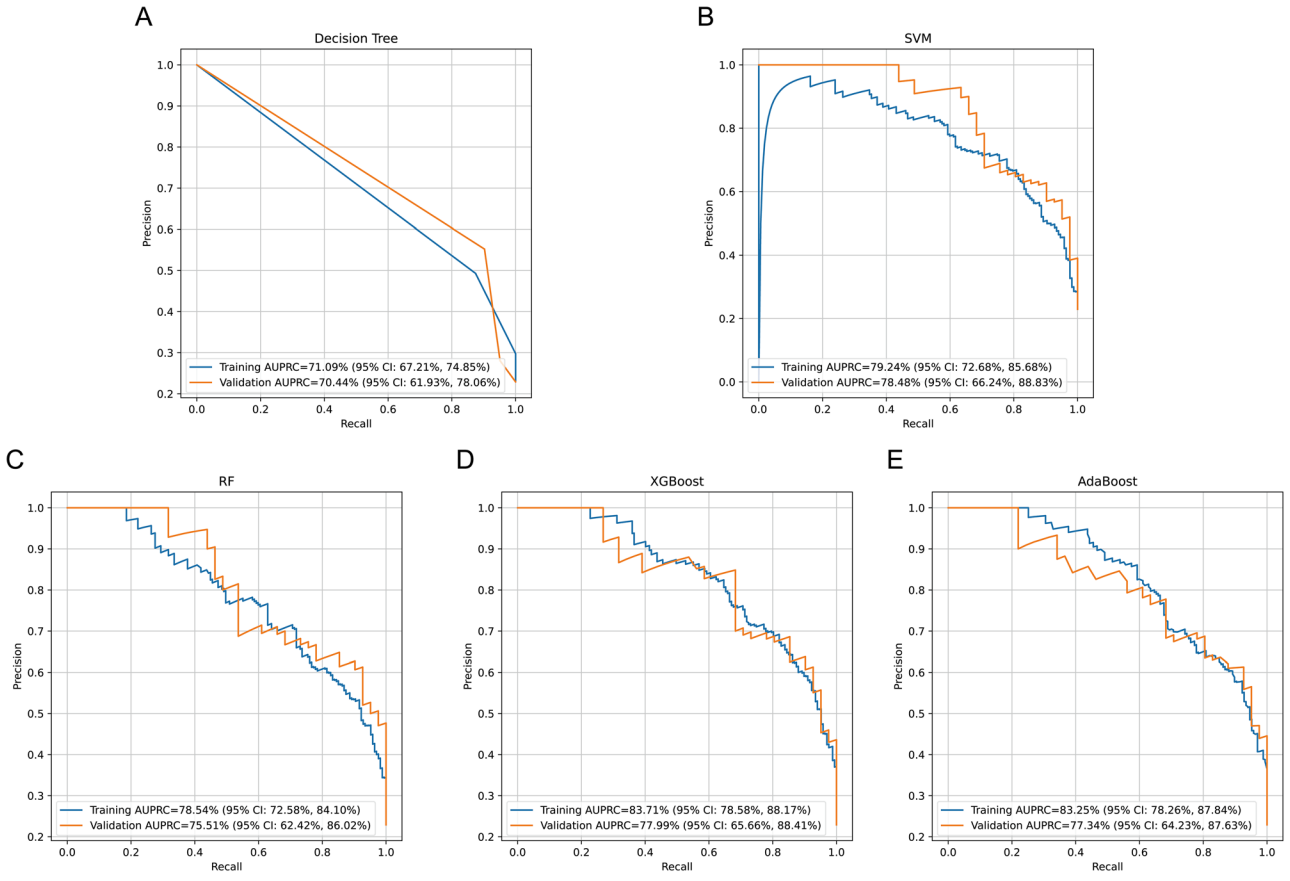

Fig. 3 The PR curves of each model plotted using all variables.The Training set and Validation set were generated using a standardized k-fold cross validation (k = 5)

models showed good fitting, except for the AdaBoost model, which exhibited slight overfitting.

Construct and validate the feature subsets

In this study, we implemented the RFECV algorithm (See Supplementary Table 3, Additional file 1) for each model and merged the importance of the feature variables that obtained from the RF and SVM models (See Supplementary Fig. 2A, B, Additional file 2), along with the SHAP values from the XGBoost model (Fig. 4). We identified the following seven variables as a feature subset for the final construction of the decision tool: WBC, NLR, D-dimer, Gallbladder width, Fibrinogen, Gallbladder wallness, Hypokalemia or hyponatremia. NLR is the ratio of neutrophils to lymphocytes and, similar to WBC, is significantly correlated with inflammatory disorders. D-dimer is a byproduct of fibrin degradation, whereas fibrinogen serves as a precursor to fibrin; both are key components of the body's coagulation process. The width and wall thickness of the gallbladder indicate if it is enlarged or edematous. Sodium and potassium are the principal ions essential for cellular electrophysiological activity in the body, while hyponatremia and hypokalemia signify environmental disruptions within the human organism. To test the validity of the feature subset, the model was re-evaluated using the above variables as presented in Table 3, and the evaluation of the various Indicators showed a slight decrease or no change, which confirms the validity of the feature subset, this conclusion was further supported by the results of the PR curves (Fig. 5). In the external test set, the indicators have decreased but still perform well (Table 4), except for RF, the AUPRC of the other models has dramatically fallen (Fig. 5). XGBoost, on the other hand, still performs well, the highest among the five models 77.87% (95% CI: 53.11-92.55%).

Date resampling

We resampled the imbalanced data and re-trained the models according to the five aforementioned algorithms, which proved that these methods either did not improve the models significantly as in the case of XGBoost, or enhanced the predictive ability of the models on the training and validation sets but deteriorated their performance on a new external test set, exacerbating the overfitting of the models. We therefore deemed these

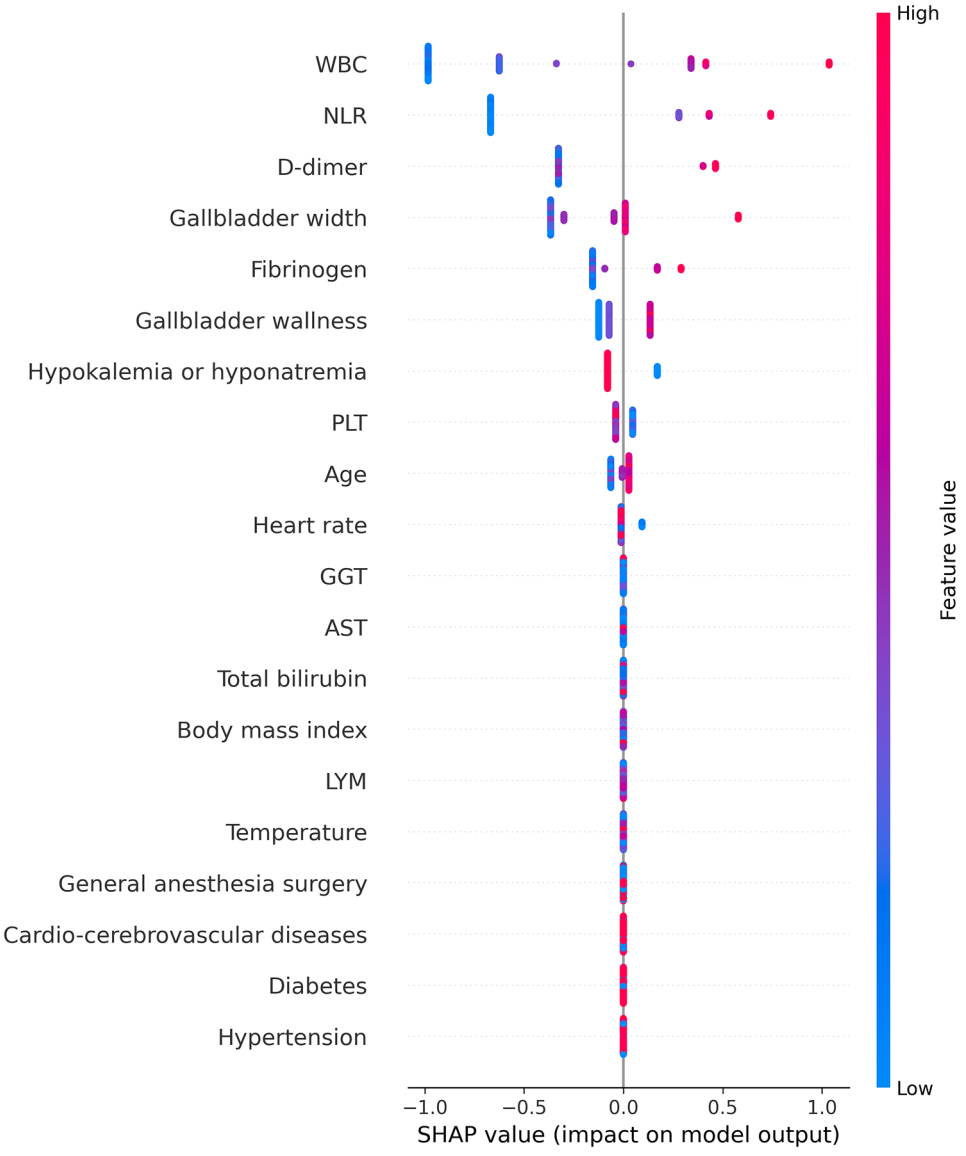

Fig. 4 The SHAP values about XGBoost

Table 3 The evaluation metrics for each model using feature subsets

Model	Balanced accuracy	Recall	Precision	F1 score
Decision Tree	79.66%(78.13–89.56%)	85.59%(80.49–97.73%)	49.80%(43.55–66.68%)	62.91%(58.00-77.19%)
SVM	82.83%(75.11–88.61%)	82.64%(65.21–90.63%)	59.71%(48.94–76.09%)	69.24%(58.53–79.67%)
Random Forest	81.09%(81.36–91.42%)	85.12%(84.44–100.00%)	53.03%(46.55–71.15%)	65.31%(62.50-81.08%)
XGBoost	82.54%(78.14–91.04%)	82.21%(67.57–91.90%)	59.61%(54.90–81.40%)	68.93%(53.11–92.55%)
AdaBoost	77.18%(70.12–85.44%)	0.5913(81.36–91.42%)	0.7884(81.36–91.42%)	0.6758(81.36–91.42%)

algorithms to be unsuitable for use in the construction of this predictive model. Additionally, we generated the PR curves for the model using various resampling procedures to confirm our findings (See Supplementary Fig. 3, Additional file 2).

Compare with preoperative diagnosis

Furthermore, we conducted a comparison between the model and the preoperative diagnosis made by the clinician at our center. Since the preoperative diagnosis only provided the final predictive label, we generated the corresponding confounding matrix and calculated the corresponding index (Table 5). XGBoost demonstrated higher accuracy and specificity compared to the preoperative

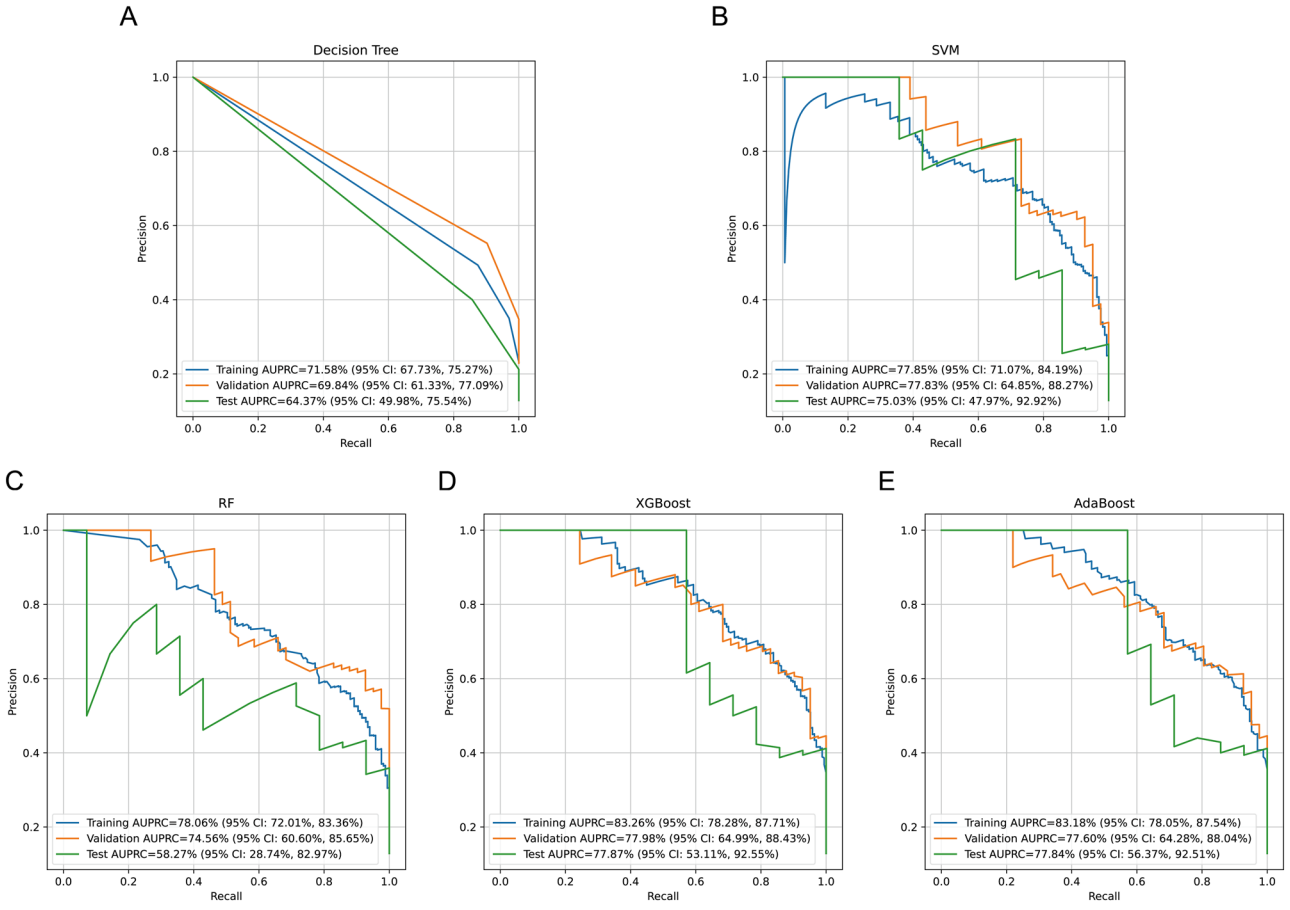

Fig. 5 The PR curves of each model were used feature subsets on Training, Validation, and Testing sets

Table 4 The evaluation metrics for each model using feature subsets on external test set

Model	Balanced accuracy	Recall	Precision	F1 score
Decision Tree	83.38%(71.68–92.42%)	85.71%(62.50–100.00%)	40.00%(22.73–58.34%)	54.55%(35.29-70.00%)
SVM	79.92%(67.63–92.28%)	71.43%(44.44–93.77%)	47.62%(25.00-69.23%)	57.14%(33.33–75.57%)
Random Forest	84.43%(72.83–92.86%)	85.71%(64.70–100.00%)	42.86%(25.92-60.00%)	57.14%(37.50-74.42%)
XGBoost	81.50%(68.90–93.50%)	71.43%(44.43–92.86%)	55.56%(30.75-80.00%)	62.50%(37.02–78.79%)
AdaBoost	78.05%(63.99–91.39%)	57.14%(31.25–81.86%)	88.89%(64.27–100.00%)	69.57%(42.11–88.01%)

diagnosis. Furthermore, XGBoost achieves an AUROC of 94.40%, indicating that this model has superior predictive capability over traditional diagnostic methods (Fig. 6).

Model interpretation and on-line app

To better understand the ability of the model as an early predictive tool and its clinical application, we demonstrated the local interpretability of the model by creating a force plot. This plot was generated using the SHAP values in XGBoost and showcased three samples when the model made accurate predictions (See Supplementary Fig. 4, Additional file 2). We identified the seven best predictors by XGBoost as WBC, NLR, D-dimer, Gallbladder width, Fibrinogen, Gallbladder wallness, Hypokalemia or hyponatremia. We labeled the distribution of eigenvalues for each group in Table 2 and GC and No-GC are statistically significantly different ($p<0.001$). In addition, we plotted the statistics data for each variable within the specific set of features (See Supplementary Fig. 5, Additional file 2): the WBC was 3.50–9.50 (10^9/L) in normal people, while it was 13.49 [10.13, 17.53] (10^9/L) in the GC group and 6.82 [5.30, 9.21] (10^9/L) in the No-GC group; About the normal range of NLR Currently, there is no reference value for large healthy population samples in China, so we used the reference value of 0.4-3.1930 for ethnically similar Korean population samples, which was 13.01 [7.21, 18.70] in the GC group, and 2.56 [1.67, 5.27] in the No-GC group; the D-dimer was about 0–1 (ug/mL) in normal subjects, and 1.02 [0.66, 1.65](ug/mL) in the GC group and 0.54 [0.39, 0.77] (ug/mL) in the No-GC

Table 5 Various model indicators in preoperative diagnosis and XGBoost model

	Training Set		Test Set	
	Preoperative diagnosis	XGBoost	Preoperative diagnosis	XGBoost
Accuracy	65.55% (62.54–69.01%)	83.61% (81.38–85.84%)	82.57% (76.15–88.99%)	88.07% (81.65–93.58%)
Sensitivity	51.44% (44.83–58.53%)	87.02% (82.52–91.88%)	92.86% (75.00-100.00%)	71.43% (45.45–93.34%)
Specificity	69.81% (66.37–73.15%)	82.58% (79.82–85.48%)	81.05% (72.91–88.42%)	90.53% (84.54–95.92%)
AUROC	69.81% (56.85–64.58%)	93.38% (91.65–94.91%)	86.95% (78.20–93.30%)	94.40% (88.80-98.39%)

The values were shown as mean (95% confidence interval)

AUROC = area under receiver operating characteristic curve

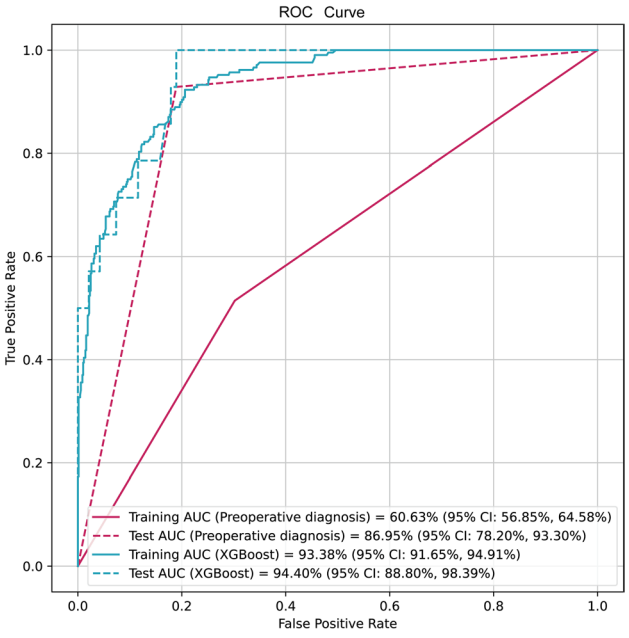

Fig. 6 The AUROC curve graphically represents the performance of XGBoost and preoperative diagnosis. AUROC = area under receiver operating characteristic curve

group; Fibrinogen is about 2–4 (g/L) in normal subjects, and in this study it was 5.61 [4.28, 7.24] (g/L) in the GC group and 3.71 [3.12, 4.52] (g/L) in the No-GC group; And Gallbladder width was about 3–5 (cm) in normal subjects, 3.30 [2.40, 4.00] (cm) in the GC group, and 3.90 [3.30, 4.50] (cm) in the No-GC group, all of which were within normal limits; Gallbladder wallness was 0.3–0.5 (cm) in normal subjects, and 0.50 [0.40, 0.60] (cm) in GC groups and 0.35 [0.30, 0.50] (cm) in the No-GC group. The prevalence of Hypokalemia or hyponatremia was 87.7% in the GC group and 48.1% in the No-GC group, making it 1.82 times higher than in the No-GC group.

In order to enhance accessibility for patients and clinicians at different centers, we had transformed the final model into a user-friendly predictive on-line app. By inputting the values of the aforementioned predictive variables, the app could accurately determine the likelihood of having gangrenous cholecystitis. Additionally, the contribution of every variable to the prediction outcome was clearly displayed in Fig. 7 (https://gangrenous-cholecystitis-prediction-model.streamlit.app/).

Discussion

We developed a predictive tool for early detection of GC by utilizing various ML models and clinically significant data obtained from patients at our medical facility. Based on the actual data characteristics of missing and imbalanced clinical information, we initially performed data interpolation and subsequently assessed the models using measures such as Balanced accuracy and PR curve, which were not influenced by imbalanced data. The XGBoost integrated learning model demonstrated the highest effectiveness among the five ML models trained using low-cost clinical examination methods, and the Balanced accuracy on the validation set could reach up to 82.54% (95% CI: 78.14-91.04%), and the AUPRC was 77.98% (95% CI: 64.99- 88.43%), and the prediction ability was also optimal on the external test set: Balanced accuracy was 81.50% (95% CI: 68.90-93.50%) and AUPRC was 77.87% (95% CI: 53.11-92.55%). Utilizing SHAP values, we employed XGBoost to analyze and extract a subset of characteristics, namely WBC, NLR, D-dimer, Gallbladder width, Fibrinogen, Gallbladder wallness, Hypokalemia or hyponatremia. Additionally, we conducted an interpretability analysis and on-line app based on the good performance of XGBoost, we advised users to utilize this model to anticipate as soon as possible and take surgical intervention measures as soon as possible. The study showcased that our model is a practical, cost-effective, and high-performing AI model for making healthcare decisions.

In clinical research, it is inevitable to encounter missing and imbalanced data in real-world studies, the deliberate omission may result "easy data," which can lead to biased model evaluations. This bias occurs because the model tends to select samples from the majority class without considering other factors, thus producing overly optimistic predictive scores [28]. Among the numerous current studies on risk factors associated with GC, for example, Yacoub et al. constructed a logistic regression model that considered factors such as sex, WBC count, heart rate, Gallbladder wallness, and age, and the model achieved a precision of approximately 90% for patients with scores higher than 4.5. However, it is important to note that

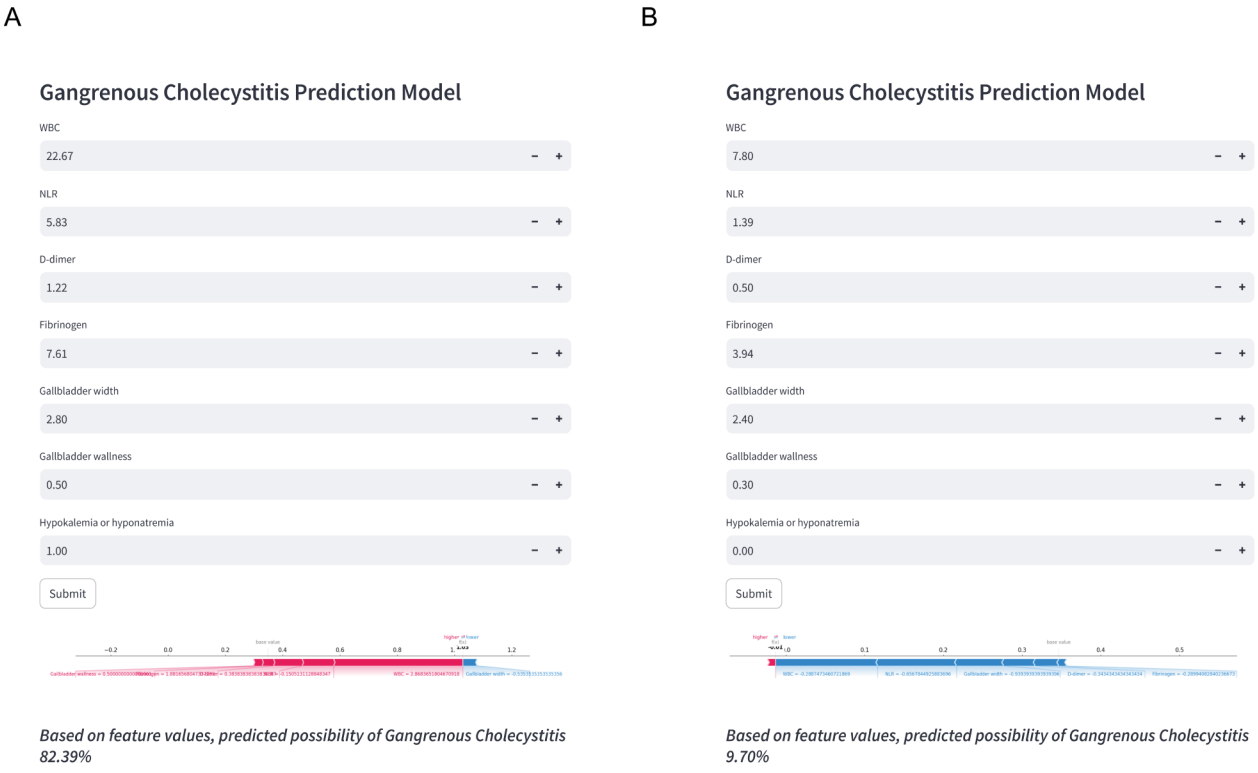

Fig. 7 The actual screen of predicting on the on-line app after inputting the true values of feature subset variables. (**A**) No-GC group, (**B**) GC group. GC = Gangrenous cholecystitis

this study had a small sample size of 245 specimens, with only 68 of them being GCs [29]. Wu et al. conducted a study with a sample size of 5243 patients, they developed a prediction model using four variables: WBC count, heart rate, Gallbladder wallness, age and had an AUROC of 0.77. However, despite increasing the sample size, the study only had 351 cases of GC and did not consider the issue of data imbalance [6]. Similarly Mahdi et al. utilized the American Society of Anesthesiology (ASA) score, temperature, duration of symptoms, WBC, male gender, and pericholecystic fluid had an AUC of 0.84 (95% CI: 0.78–0.90) in a single-center sample of 587 cases. This performance was superior to the first two studies mentioned, but it still suffered from the limitation that the model evaluation metrics may not accurately reflect the model's performance (24.7% GC) [30].

All of the above studies used traditional logistic regression statistical models, which further compromised the accuracy and reliability of evaluating the importance of predictor variables and model performance in the presence of data imbalance. Furthermore, ML diverges from classical statistics by prioritizing prediction accuracy over hypothesis validation, making it better suited for developing predictive diagnostic models. Describing complex clinical realities using a limited number of mathematical formulas is inherently challenging. ML does not make any assumptions about the data and departs from the traditional statistical process of validating assumptions with *p* values and using cross-validation to confirm the final results [15, 31]. ML may integrate clinical data, including text and images, with various data sources like genomes to create adaptable models, they can be continuously enhanced to achieve more precise predicting outcomes in the future [32]. In this study, we addressed the limitations of a previous study by considering missing and imbalanced data, as well as a small sample size, and proposed a novel approach using ML models to develop an early predictive diagnostic model for GC. We rigorously evaluate the performance of our model, and the results from an external test set validate its effectiveness. Furthermore, the model exhibits superior performance in comparison to preoperative diagnoses made by clinical doctors, achieving an AUROC of 94.40% (95% CI: 88.80-98.39%) as opposed to 86.95% (95% CI: 88.80-98.39%).

Several studies have grouped acute purulent cholecystitis with GC as severe cholecystitis in subsequent analyses [33–36], which confirms the view of some researchers that purulent cholecystitis is equally "important" as GC in some cases, as can be seen from the quartile spacing and statistical plots of the previous indicators. The disparity between the GC group and the normal range is merely 0.02ug/mL for D-dimer, the distinction between

Gallbladder width and Gallbladder wallness falls within the normal range, despite the presence of a discrepancy between the two indicators. This may be due to the propensity of many patients in the sample to seek medical attention within 6 h of symptom onset, during which local inflammation has not yet triggered a systemic reaction, and changes in blood tests remain negligible. It is evident that in many instances, the performance of the two diseases is comparable, leading to confusion and consequently making the diagnosis of GC more challenging. Nevertheless, based on our expertise, certain instances of purulent cholecystitis can be managed with conservative approaches involving anti-inflammatory therapies. On the other hand, GC always necessitates immediate surgical intervention, either through PTGBD or surgery. Among the 1006 patients in this study, there were 3 cases of malignant deaths during this admission, yielding a mortality rate of about 0.3%, all of which were GC without exception. The causes of death were analyzed by reviewing the death discussion records, one case was due to acute myocardial infarction, and the other two cases were multiple organ dysfunction syndrome (MODS) resulting from infectious shock. The surgery of the former was performed on the 13th day after the onset of the disease, and the latter two on the 4th and 11th day after the onset of the disease, respectively. we analyzed that the late surgical intervention might be one of the underlying causes of the patients' deaths. Hence, the diagnostic prediction tool developed in this study offers theoretical backing for determining the need for surgical intervention in patients with GC, thereby providing significant clinical value in terms of minimizing postoperative complications and death rates, and enhancing overall patient outcomes. Furthermore, with the clinical applicaiton of on-line app, clinicians have the ability to enhance preoperative preparation, which includes psychological readiness, and minimize the incidence of surgical incidents.

In the subset of features chosen by those ML models in this study, both inflammatory markers such as WBC and coagulation system-related markers such as D-dimer played a significant role in constructing the prediction models. However, the variables of diabetes mellitus and history of coronary artery disease, which have been previously identified as independent risk factors for GC in other studies, were found to be more prevalent in the No-GC group in this study [2, 37]. We attribute this discrepancy to the larger sample size in the No-GC group, which resulted in a statistically significant difference. Previous studies have demonstrated the significance of WBC has in the constructing GC models [8, 38, 39], while D-dimer plays a crucial role in the diagnosis and treatment of acute abdominal conditions, e.g. Hizir et al. discovered that D-dimer had a sensitivity of up to 95.7% in detecting nontraumatic abdominal pain patients who require surgical intervention [40]; Cayrol et al. used D-dimer to predict an AUC of 0.93 for acute gangrenous appendicitis in children [41]. There is a lack of recorded studies on the relationship between acute gangrenous cholecystitis and gangrenous cholecystitis, which raises questions about their connection. Many studies have confirmed the correlation between inflammation and coagulation [42–44], the interaction between many inflammatory factors and coagulation pathways forms a comprehensive feedback loop, and the relationship between them is highly intricate. Patients with sepsis and COVID-19 disease typically exhibit with elevated levels of fibrinogen and D-dimer, along with thrombocytopenia [45]. A study by De Simone et al. further revealed that the incidence of GC even doubled in patients with COVID-19 [46]. Based on the above studies, one might hypothesize whether the treatment methodologies for these two disorders can be interchanged. Can treatment procedures be shared or adapted from one another? Further verification in future investigations is required.

This study involves conducting research using authentic clinical data and directly addressing the challenges of incomplete and imbalanced clinical information. Through addressing these practical obstacles, our objective is to develop a GC prediction model that is more aligned with real-world scenarios. This strategy, which utilizes actual patient data, seeks to improve the dependability and flexibility of the model in order to effectively handle intricate clinical settings. Hence, our emphasis lies not only on the theoretical optimality but also on the model's resilience in handling actual clinical settings. Our study focuses on resolving the issue of incomplete and imbalanced clinical information. The objective is to develop clinical decision support systems that are more actionable and practical, so offering physicians more trustworthy aid in real clinical practice. It is essential to prioritize the safeguarding of patient privacy when utilizing this paradigm. Medical institutions and researchers can implement numerous protective measures, including data encryption, access control, user and developer privacy training, frequent security audits, and strict adherence to applicable laws and regulations to secure patient medical data privacy.

This study has several limitations: (1) This study is a single-center retrospective study, and it can be seen that the model has a certain instability on new data, so it is still necessary to repeatedly train the model to enhance the stability of the model in the future with multicenter prospective studies. (2) The ML models used in this study are the prevailing ones. In recent years, deep learning models, such as the CNN, have demonstrated their exceptional performance [47] in classifying imbalanced data, it is recommended to construct future experiments to evaluate their performance. (3) Due to the limitations

of the clinical laboratory system in this clinical center and the retrospective cohort study, there may be biases in patient collection. We have established clear and specific inclusion and exclusion criteria. Ensure that all research subjects meet these standards in order to improve sample consistency and representativeness. (4) In patients with comorbidities that may induce changing aberrations in the model, the likelihood of misdiagnosis may escalate when employing the model. (5) There may be a bias in the collection of effective variables in this model. The variables of interest in this paper are the affordable and readily available universal indicators in the clinic that have been studied before. The model built using these variables has shown a high level of accuracy, indicating better performance. Future studies can consider incorporating additional input variables such as the ASA score, peripheral fat of gallbladder, mucosal interruption sign in gallbladder ultrasound [48]and transient hepatic attenuation differences (THADs) [49] in CT arterial phase. This will allow for further exploration of clinical features related to GC.

Conclusion

In conclusion, our study has successfully developed a XGBoost ML model for the early diagnosis of GC, achieving high classification accuracy of 81.50% and AUROC of 94.40% over traditional diagnostic methods. The SHAP value was utilized to interpret the model and an convenient on-line predictive app including WBC, NLR, D-dimer, Gallbladder width, Fibrinogen, Gallbladder wallness, Hypokalemia or hyponatremia was developed. Overall, our research highlighted the potential of ML in advancing early detection strategies for GC, advocating for prompt surgical interventions, and offering valuable support to healthcare professionals in optimizing patient care and outcomes.

Abbreviations

GC	Gangrenous Cholecystitis
No-GC	No-Gangrenous Cholecystitis
SVM	Support Vector Machine
RF	Random Forest
XGBoost	Extreme Gradient Boosting
AdaBoost	Adaptive Boosting
SHAP	Shapley Additive explanation
AUPRC	Area Under the Precision-Recall Curve
AI	Artificial intelligence
ML	Machine Learning
BMI	Body Mass Index
WBC	White Blood Cell
NEU	Neutrophil
LYM	Lymphocyte
PLT	Platelet
NLR	Neutrophil to Lymphocyte Ratio
PLR	Platelet to Lymphocyte Ratio
AST	Aspartate aminotrasferase
ALT	Alanine transaminase
GGT	Gamma-glutamyltransferase
SMOTE	Synthetic Minority Over-Sampling

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13017-024-00571-6.

Supplementary Material 1

Supplementary Material 2

Author contributions

Author's ContributionYing Ma, Man Luo and Fuwen Luo were responsible for the overall study design. Ying Ma, Man Luo, Guoxin Guan, Xingming Liu and Xingye Cui supervised the data collection. Ying Ma, Man Luo and Xingming Liu performed data analysis. Ying Ma, Man Luo, Guoxin Guan completed manuscript drafting. Xingye Cui and Fuwen Luo were responsible for manuscript editing. All authors read, discussed, and approved the final version of the manuscript.

Funding

This study is supported by the National Natural Science Foundation of China.

Data availability

The data where our results derived from were from the Second Hospital of Dalian Medical University. The original data were not publicly available and could only be shared with the permission of the Ethics Committee of the Second Hospital of Dalian Medical University.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of the Second Affiliated Hospital of Dalian Medical University (KY2024-006-02) and followed the Declaration of Helsinki.

Consent for publication

Not applicable.

Provenance and peer review

Not commissioned, externally peer-reviewed.

Assistance with the study

none.

Presentation

None.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of General Surgery, The Second Hospital of Dalian Medical University, Zhongshan Road, Shahekou District, Dalian City, Liaoning Province 116023, China
[2]Center on Frontiers of Computing Studies, School of Compter Science, Inst. for Artificial Intelligence, Peking University, Beijing 100871, China

Received: 15 October 2024 / Accepted: 16 December 2024
Published online: 06 January 2025

References

1. Morfin E, Ponka JL, Brush BE. Gangrenous cholecystitis. Arch Surg. 1968;96:567–73. https://doi.org/10.1001/archsurg.1968.01330220083015.
2. Ganapathi AM, Speicher PJ, Englum BR, Perez A, Tyler DS, Zani S. Gangrenous cholecystitis: a contemporary review. J Surg Res. 2015;197:18–24. https://doi.org/10.1016/j.jss.2015.02.058.
3. Maddu K, Phadke S, Hoff C. Complications of cholecystitis: a comprehensive contemporary imaging review. Emerg Radiol. 2021;28:1011–27. https://doi.org/10.1007/s10140-021-01944-z.

4. Shirah BH, Shirah HA, Saleem MA, Chughtai MA, Elraghi MA, Shams ME. Predictive factors for gangrene complication in acute calculous cholecystitis. Ann Hepatobiliary Pancreat Surg. 2019;23:228–33. https://doi.org/10.14701/ahbps.2019.23.3.228.
5. Safa R, Berbari I, Hage S, Dagher GA. Atypical presentation of gangrenous cholecystitis: a case series. Am J Emerg Med. 2018;36:e21351–2135. https://doi.org/10.1016/j.ajem.2018.08.039.
6. Wu B, Buddensick TJ, Ferdosi H, Narducci DM, Sautter A, Setiawan L, Shaukat H, Siddique M, Sulkowski GN, Kamangar F, et al. Predicting gangrenous cholecystitis. HPB (Oxford). 2014;16:801–6. https://doi.org/10.1111/hpb.12226.
7. Mayumi T, Okamoto K, Takada T, Strasberg SM, Solomkin JS, Schlossberg D, Pitt HA, Yoshida M, Gomi H, Miura F, et al. Tokyo guidelines 2018: management bundles for acute cholangitis and cholecystitis. J Hepatobiliary Pancreat Sci. 2018;25:96–100. https://doi.org/10.1002/jhbp.519.
8. Raffee L, Kuleib S, Kewan T, Alawneh K, Beovich B, Williams B. Utility of leucocytes, inflammatory markers and pancreatic enzymes as indicators of gangrenous cholecystitis. Postgrad Med J. 2020;96:134–. https://doi.org/10.1136/postgradmedj-2019-137095.
9. Sureka B, Jha S, Rodha MS, Chaudhary R, Elhence P, Khera PS, Garg PK, Yadav T, Goel A. Combined hyperdense gallbladder wall-lumen sign: new computed tomography sign in acute gangrenous cholecystitis. Pol J Radiol. 2020;85:e183–7. https://doi.org/10.5114/pjr.2020.94337.
10. Mok KWJ, Reddy R, Wood F, Turner P, Ward JB, Pursnani KG, Date RS. Is C-reactive protein a useful adjunct in selecting patients for emergency cholecystectomy by predicting severe/gangrenous cholecystitis? Int J Surg. 2014;12:649–53. https://doi.org/10.1016/j.ijsu.2014.05.040.
11. Kim K-H, Kim S-J, Lee SC, Lee SK. Risk assessment scales and predictors for simple versus severe cholecystitis in performing laparoscopic cholecystectomy. Asian J Surg. 2017;40. https://doi.org/10.1016/j.asjsur.2015.12.006.
12. Hood SP, Cosma G, Foulds GA, Johnson C, Reeder S, McArdle SE, Khan MA, Pockley AG. Identifying prostate cancer and its clinical risk in asymptomatic men using machine learning of high dimensional peripheral blood flow cytometric natural killer cell subset phenotyping data. Elife. 2020;9:e50936. https://doi.org/10.7554/eLife.50936.
13. Gould MK, Huang BZ, Tammemagi MC, Kinar Y, Shiff R. Machine learning for early Lung Cancer Identification using Routine Clinical and Laboratory Data. Am J Respir Crit Care Med. 2021;204:445–53. https://doi.org/10.1164/rccm.202007-2791OC.
14. Yavuz E, Eyupoglu C. An effective approach for breast cancer diagnosis based on routine blood analysis features. Med Biol Eng Comput. 2020;58:1583–601. https://doi.org/10.1007/s11517-020-02187-9.
15. Wang P, Li Y, Reddy CK. (2017). Machine Learning for Survival Analysis: A Survey. Preprint at arXiv, https://doi.org/10.48550/arXiv.1708.04649
16. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin Chem. 2015;61:1446–52. https://doi.org/10.1373/clinchem.2015.246280.
17. Mathew G, Agha R, Albrecht J, Goel P, Mukherjee I, Pai P, D'Cruz AK, Nixon IJ, Roberto K, Enam SA, et al. STROCSS 2021: strengthening the reporting of cohort, cross-sectional and case-control studies in surgery. Int J Surg. 2021;96:106165. https://doi.org/10.1016/j.ijsu.2021.106165.
18. Schafer JL, Olsen MK. Multivar Behav Res. 1998;33:545–71. https://doi.org/10.1207/s15327906mbr3304_5. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective.
19. Buuren SV, Groothuis-Oudshoorn K. (2011). MICE: Multivariate Imputation by Chained Equations in R. Journal of statistical software *45*. https://doi.org/10.18637/jss.v045.i03
20. Lall R, Robinson T. Polit Anal. 2022;30:179–96. https://doi.org/10.1017/pan.2020.49. The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning.
21. Shalabi LA, Shaaban Z, Kasasbeh B. Data Mining: a Preprocessing Engine. J Comput Sci. 2006;2. https://doi.org/10.3844/jcssp.2006.735.739.
22. Thölke P, Mantilla-Ramos Y-J, Abdelhedi H, Maschke C, Dehgan A, Harel Y, Kemtur A, Mekki Berrada L, Sahraoui M, Young T, et al. Class imbalance should not throw you off balance: choosing the right classifiers and performance metrics for brain decoding with imbalanced data. NeuroImage. 2023;277:120253. https://doi.org/10.1016/j.neuroimage.2023.120253.
23. Plante TB, Blau AM, Berg AN, Weinberg AS, Jun IC, Tapson VF, Kanigan TS, Adib AB. Development and External Validation of a machine Learning Tool to Rule out COVID-19 among adults in the Emergency Department using routine blood tests: a large, Multicenter, Real-World Study. JMIR Publications Inc; 2020. https://doi.org/10.2196/24048.
24. Lundberg SM, Lee S-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems NIPS'17. (Curran Associates Inc.), pp. 4768–4777.
25. He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng. 2009;21:1263–84.
26. He H, Ma Y. (2013). Imbalanced learning. Foundations, algorithms, and applications (Imbalanced learning. Foundations, algorithms, and applications).
27. Davis J. (2006). The Relationship Between Precision-Recall and ROC Curves. In Proceedings of the 23th International Conference on Machine Learning, 2006.
28. Tasci E, Zhuge Y, Camphausen K, Krauze AV. Bias and Class Imbalance in Oncologic Data—towards Inclusive and transferrable AI in large scale Oncology Data sets. Cancers. 2022;14. https://doi.org/10.3390/cancers14122897.
29. Wn Y, M, P., I, S., Y, M., P, C., and, Rj M. (2010). Prediction of patients with acute cholecystitis requiring emergent cholecystectomy: a simple score. Gastroenterology research and practice *2010*. https://doi.org/10.1155/2010/901739
30. Bouassida M, Madhioub M, Kallel Y, Zribi S, Slama H, Mighri MM, Touinsi H. Acute gangrenous cholecystitis: proposal of a score and comparison with previous published scores. J Gastrointest Surg. 2021;25:1479–86. https://doi.org/10.1007/s11605-020-04707-2.
31. Liu W, Laranjo L, Klimis H, Chiang J, Yue J, Marschner S, Quiroz JC, Jorm L, Chow CK. Machine-learning versus traditional approaches for atherosclerotic cardiovascular risk prognostication in primary prevention cohorts: a systematic review and meta-analysis. Eur Heart J Qual Care Clin Outcomes. 2023;9:310–22. https://doi.org/10.1093/ehjqcco/qcad017.
32. Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. Nat Biotechnol. 2018;36:829–38. https://doi.org/10.1038/nbt.4233.
33. Borzellino G, Sauerland S, Minicozzi AM, Verlato G, Di Pietrantonj C, De Manzoni G, Cordiano C. Laparoscopic cholecystectomy for severe acute cholecystitis. A meta-analysis of results. Surg Endosc. 2008;22:8–15. https://doi.org/10.1007/s00464-007-9511-6.
34. Borzellino G, Steccanella F, Mantovani W, Genna M. Predictive factors for the diagnosis of severe acute cholecystitis in an emergency setting. Surg Endosc. 2013;27:3388–95. https://doi.org/10.1007/s00464-013-2921-8.
35. Chen J, Gao Q, Huang X, Wang Y. Prognostic clinical indexes for prediction of acute gangrenous cholecystitis and acute purulent cholecystitis. BMC Gastroenterol. 2022;22:491. https://doi.org/10.1186/s12876-022-02582-6.
36. Portinari M, Scagliarini M, Valpiani G, Bianconcini S, Andreotti D, Stano R, Carcoforo P, Occhionorelli S. Do I need to operate on that in the Middle of the night? Development of a Nomogram for the diagnosis of severe Acute Cholecystitis. J Gastrointest Surg. 2018;22:1016–25. https://doi.org/10.1007/s11605-018-3708-y.
37. Bourikian S, Anand RJ, Aboutanos M, Wolfe LG, Ferrada P. Risk factors for acute gangrenous cholecystitis in emergency general surgery patients. Am J Surg. 2015;210:730–3. https://doi.org/10.1016/j.amjsurg.2015.05.003.
38. Alghamdi KA, Rizk HA, Jamal WH, Bakhshween AA, Basourrah MK. Risk factors of gangrenous cholecystitis in general surgery patient admitted for Cholecystectomy in King Abdul-Aziz University Hospital (KAUH), Saudi Arabia. Materia Socio Med. 2019;31:286. https://doi.org/10.5455/msm.2019.31.286-289.
39. Siada S, Jeffcoach D, Dirks RC, Wolfe MM, Davis JW. (2019). A predictive grading scale for acute cholecystitis. Trauma Surgery Acute Care Open *4*. https://doi.org/10.1136/tsaco-2019-000324
40. Akyildiz HY, Erdoan Szüer, Akcan A, Can Küük, Yilmaz, N. The value of D-dimer test in the diagnosis of patients with nontraumatic acute abdomen. Ulusal Travma ve acil Cerrahi Dergisi = Turkish J Trauma Emerg Surgery: TJTES. 2010;16:22–6. https://doi.org/10.1016/j.resuscitation.2009.10.016.
41. Julie C. Maria, Concepción, Miguez, Gloria, Guerrero, Cristina, Tomatis, and Isabel (2016). Diagnostic accuracy and prognostic utility of D Dimer in acute appendicitis in children. European Journal of Pediatrics. https://doi.org/10.1007/s00431-015-2632-3
42. Wu C, Lu W, Zhang Y, Zhang G, Shi X, Hisada Y, Grover SP, Zhang X, Li L, Xiang B, et al. Inflammasome activation triggers blood clotting and host death through pyroptosis. Immunity. 2019;50:1401–e14114. https://doi.org/10.1016/j.immuni.2019.04.003.
43. Zhang H, Zeng L, Xie M, Liu J, Zhou B, Wu R, Cao L, Kroemer G, Wang H, Billiar TR, et al. TMEM173 drives Lethal Coagulation in Sepsis. Cell Host Microbe. 2020;27:556–e5706. https://doi.org/10.1016/j.chom.2020.02.004.
44. Yang X, Cheng X, Tang Y, Qiu X, Wang Y, Kang H, Wu J, Wang Z, Liu Y, Chen F, et al. Bacterial endotoxin activates the Coagulation Cascade through Gasdermin D-Dependent Phosphatidylserine exposure. Immunity. 2019;51:983–e9966. https://doi.org/10.1016/j.immuni.2019.11.005.

45. Tang D, Comish P, Kang R. The hallmarks of COVID-19 disease. PLoS Pathog. 2020;16:e1008536. https://doi.org/10.1371/journal.ppat.1008536.
46. De Simone B, Abu-Zidan FM, Chouillard E, et al. The ChoCO-W prospective observational global study: does COVID-19 increase gangrenous cholecystitis? World J Emerg Surg. 2022;17:61. https://doi.org/10.1186/s13017-022-00466-4.
47. Valova I, Harris C, Mai T, Gueorguieva N. Optimization of convolutional neural networks for Imbalanced Set classification. Procedia Comput Sci. 2020;176:660–9. https://doi.org/10.1016/j.procs.2020.09.038.
48. Tse JR, Gologorsky R, Shen L, Bingham DB, Jeffrey RB, Kamaya A. Evaluation of early sonographic predictors of gangrenous cholecystitis: mucosal discontinuity and echogenic pericholecystic fat. Abdom Radiol (NY). 2022;47:1061–70. https://doi.org/10.1007/s00261-021-03320-4.
49. Uemura S, Higuchi R, Yazawa T, Izumo W, Sugishita T, Morita S, Yamamoto M. Impact of transient hepatic attenuation differences on computed tomography scans in the diagnosis of acute gangrenous cholecystitis. J Hepatobiliary Pancreat Sci. 2019;26:348–53. https://doi.org/10.1002/jhbp.637.

Publisher's note