BMC Gastroenterology

RESEARCH

Open Access

Risk prediction and effect evaluation of complicated appendicitis based on XGBoost modeling

Sunmeng Chen[1†], Jianfu Xia[1†], Beibei Xu[2], Yi Huang[3], Miaomiao Teng[4] and Juyi Pan[2*]

Abstract

Purpose The distinction between complicated appendicitis (CAP) and uncomplicated appendicitis (UAP) remains challenging. The purpose of this study was to construct a safe and economical diagnostic model that can accurately and rapidly differentiate between CAP and UAP.

Methods Patient data from 773 appendectomies were retrospectively collected, important features were selected using random forests, and the data were divided into training and test sets in a 3:1 ratio. An integrated learning algorithm, Extreme Gradient Boosting (XGBoost), was introduced to predict the risk of CAP and compared with Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (CART) algorithms. A comprehensive comparison of the four algorithms was performed using model performance metrics such as the area under the receiver's operating characteristic curve (AUC), sensitivity, specificity, accuracy, precision, negative predictive value(NPV), positive predictive value(PPV),calibration curves, and clinical decision curve analysis (DCA).

Result The results show that all four prediction models exhibit some predictive ability. The XGBoost model showed the best prediction with AUC, accuracy, sensitivity, specificity,NPV and PPV of 0.914, 0.855, 0.865, 0.846,0.848 and 0.897, respectively, followed by the SVM model with results of AUC, accuracy, sensitivity, specificity,NPV and PPV of 0.882, 0.819, 0.865, 0.779, 0.770 and 0.871, respectively. XGBoost and SVM models show very good calibration. The XGBoost model showed better net clinical benefit compared to the DCA curves of the other models.

Conclusion Predictive models based on the XGBoost algorithm have good performance in predicting the risk of acute appendicitis progressing to complicated appendicitis, which helps to optimize clinical decision making.

Keywords Disease risk prediction, Machine learning, Complicated appendicitis, XGBoost model, Predictive modeling

†Sunmeng Chen and Jianfu Xia contributed equally to this work.

*Correspondence:
Juyi Pan
panjuyi@126.com
[1] Department of General Surgery, The Dingli Clinical College of Wenzhou Medical University (Wenzhou Central Hospital), Wenzhou, Zhejiang 325000, China
[2] Department of Gastroenterology, Wenzhou Third Clinical Institute Affiliated to Wenzhou Medical University, The Third Affiliated Hospital of Shanghai University, No 57, Cang Hou Street, Wenzhou, Zhejiang 325000, China
[3] Department of General Surgery, Wenzhou Third Clinical Institute Affiliated to Wenzhou Medical University, The Third Affiliated Hospital of Shanghai University, Wenzhou, Zhejiang 325000, China
[4] Department of Gastroenterology, Postgraduate Training Base Alliance of Wenzhou Medical University, Wenzhou, Zhejiang 325000, China

Introduction

Acute appendicitis is one of the most common acute abdominal surgical disorders globally, and the lifetime risk of appendicitis in the United States is estimated to be approximately nine percent, with a prevalence of 16–40% for complicated appendicitis [1]. In uncomplicated appendicitis, conservative antibiotic therapy is both effective and safe; in contrast, most patients with complicated appendicitis require emergency appendectomy, and delayed removal of complicated appendicitis leads to additional complications [1, 2]. Although the mortality rate for uncomplicated appendicitis is about 0.1%, the risk of death for complicated appendicitis caused by a perforated or gangrenous appendix is significantly higher, up to five to six percent [3, 4]. Therefore, it is very important to differentiate between these two diseases in time.

The preoperative diagnosis of appendicitis is mainly based on medical history, physical examination, laboratory tests, and imaging. Through these diagnostic methods, more than 90% of acute appendicitis patients can achieve early and accurate diagnosis [5], However, these methods will be challenging to differentiate between complicated appendicitis and uncomplicated appendicitis. The AIRS and Alvarado scoring systems used to diagnose acute appendicitis are constructed on the basis of clinical presentation and laboratory findings [6, 7].These scoring systems are simple to use but are highly subjective and do not effectively differentiate between complicated and uncomplicated appendicitis [8]. There are also several CT-based scoring systems that improve sensitivity and specificity to about 80% [9, 10], CT examinations increase radiation exposure and economic burden for patients, and the diagnostic results obtained in exchange for increased radiation exposure still need to be further improved.The scoring systems of Appendix Severity Index(APSI) and Atema et al. combined clinical use of CT imaging findings for the prediction of complicated appendicitis, and although the results of the two studies possessed a high specificity of 93% and a sensitivity of 90.2%, respectively [11, 12], a study validating these two studies showed that both scoring systems performed poorly overall [13].

In recent years, machine learning techniques have been widely used in clinical medicine, especially in predictive models, which improve the accuracy of predictions by learning from data and experience [14]. Some researchers, tried to apply machine learning to the prediction of complicated appendicitis.Tuong-Anh Phan-Mai et al. combined clinical data and ultrasound results to predict the risk of complicated appendicitis using various machine learning algorithms.The gradient boosting model showed a strong performance (AUC 0.890, accuracy 0.810) [15]. Hilmi Yazici et al., on the other hand, evaluated the performance of multiple machine learning algorithms for the prediction of complex appendicitis in conjunction with clinical data and abdominal CT results, with multiple algorithms demonstrating more than 90% accuracy and specificity, but only 60% sensitivity [16]. Currently relevant predictive models are very dependent on imaging and there is still room to increase their performance. Therefore, a new approach to rule out complicated appendicitis is urgently needed. Therefore, an urgent need exists for a new method to rule out complicated appendicitis.

XGBoost is a relatively novel and efficient machine learning method that combines multiple decision trees in a boosted manner, with the advantages of saving resources, less training time, and higher accuracy, developed by Chen and Guestrin in 2016, and was first used in fields such as transportation and electronics [17–19] and is now gradually being applied to many aspects of the medical field, e.g., the prediction of early-stage lung cancer, type 2 diabetes, myocardial infarction, etc., all of which have demonstrated good prediction results [20–22]. In this study we attempted to use the XGBoost algorithm for risk assessment of complicated appendicitis using only clinical data with the aim of improving the accuracy of identifying non-complicated appendicitis from complex appendicitis, which is an innovative application of the XGBoost algorithm in the field, and to compare it with other currently dominant machine learning models.The model is finally interpreted through the SHapley's method of additive interpretation (SHAP).

Methods

Patient and data

This retrospective study included all patients who underwent appendectomy between January 1, 2020, and December 30, 2023, at Wenzhou Central Hospital (Zhejiang Province, China). Inclusion criteria were all patients with histologically reported appendicitis, and patients who met one of the following criteria were excluded from the study: 1. patients who were not ultimately diagnosed with acute appendicitis (e.g., histology reported as normal appendix as well as appendiceal tumor); 2. samples with obvious errors in the data (e.g., erythrocyte count of 0.1×10^{12}/L); and 3. samples with greater than 50% missing data. After the above inclusion and exclusion criteria, a total of 773 patients were finally enrolled in the study (Supplementary Fig. 1). For the small amount of missing data in this group of patients, we utilized the missForest package to fill in the missing values, the median to fill in the continuous variables, and the plurality to fill in the categorical variables.Data on all patients were collected through electronic medical records, including demographic data (gender, age, area of residence,

underlying disease), vital signs (temperature, heart rate, systolic blood pressure, diastolic blood pressure), preoperative clinical manifestations (appetite, nausea or vomiting, diarrhea, right lower abdominal tenderness, rebound tenderness and myalgia, metastatic right lower abdominal pain, whether the time between pain and surgery was greater than 24 h, and the degree of pain), laboratory findings (white blood cell, red blood cell count, neutrophil, lymphocyte, monocyte, and eosinophil percentages, hemoglobin, erythrocytes, platelets, C-reactive protein, glucose level, blood urea nitrogen, blood creatinine, and total bilirubin), and the presence or absence of fecaliths. Underlying diseases include diabetes, hypertension, myocardial infarction, heart failure, chronic kidney disease, chronic liver disease, and chronic lung disease. Based on the VRS score, the pain level was categorized as mild, moderate, and severe. Based on the surgical reports and histopathologic findings of the appendix, the patients were divided into two groups: the uncomplicated appendicitis group and the complicated appendicitis group. Patients were considered to have CAP if they had (1) perforated appendicitis, (2) gangrenous appendicitis, or (3) complications such as diffuse peritonitis and pathological findings of abscess formation [2, 5]. Simple and suppurative appendicitis were categorized as UAP [2, 5]. The study follows the revised Declaration of Helsinki and was approved by the the ethics committee of Wenzhou CentralHospital.

Feature selection

For predictive models, too many features increase the risk of model overfitting and can predict diseases more cumbersome in clinical work. To reduce the number of features to be included, we also investigate the importance of predictor variables, in this paper, we use the Mean Decrease Accuracy function provided by RF for feature selection, which ranks the importance of the features by directly measuring the impact of each feature on the accuracy of the model, and use a bar chart to visualize the obtained importance ranking of the features, and then exclude unimportant, This operation does not significantly affect the accuracy of the model, and we finally include the top 15 features for model training.

Model development

Machine learning models were developed and validated using R software (version 4.1.4, https://www.rproject.org/). The XGBoost model and three other common machine learning predictive models were constructed by dividing the data into training and test datasets in a 3:1 ratio, with "whether or not it is complicated appendicitis" as the main output of the model, and accuracy and AUC value as the endpoint variables. The models were also evaluated by calculating AUC, accuracy, sensitivity, specificity, precision, recall, positive predictive value(PPV) and negative predictive value(NPV).

XGBoost model: XGBoost is a powerful integrated learning algorithm, a modification of gradient-boosted Decision Trees (GBDT), which integrates many weak classifiers to form a single strong classifier. The algorithm has good learning results and efficient training speeds and is able to produce prediction accuracies that match many state-of-the-art supervised learning techniques. In this study, we use the "XGBoost" package for model training, and use cross-validation and grid search to optimize the XGBoost model parameters in the training set, and finally we obtain the optimal model parameters, eta, nrounds, max_depth, gamma. subsample, which are 0.3, 1000, 2, 0.001 and 0.7 respectively, subsample are 0.3, 1000, 2, 0.001, and 0.7, respectively, which control the learning rate of the model, the number of decision trees, the maximum depth of the decision tree, the smallest sub-weight in the decision tree, the minimum loss reduction (gamma) required for partitioning, and the number of columns to be subsampled when building the tree. Finally, the model is retrained using the optimal hyperparameters described above and final predictions are made on the test set.

SVM model: SVM is a widely used supervised machine learning algorithm with the goal of creating a hyperplane that can effectively partition a given dataset into different classes. The"e1071"package was used for training, and the Gaussian kernel function was used to establish the nonlinear decision boundary, and the hyperparameters cost and gamma were set to 1 and 0.1, respectively. Cost is a penalty parameter in SVM that controls how much the model is penalized for misclassification and indirectly controls the generalization ability of the model. Gamma controls the complexity of the model by regulating the distribution of the data in the feature space.

RF model: Random forest is a commonly used integrated learning algorithm in the field of machine learning, which consists of a number of decision trees, by combining the predictions of multiple decision trees and thus improving the performance of the model. The random forest model is trained using the"randomForest"package in the R software, using two hyperparameters, ntree, and mtry. ntree is used to set the tree of the decision tree in the random forest, and mtry is used to set the number of variables that can be selected at each node, and the original parameters are set to 500 and 6, respectively, and then determined the values of ntree and mtry when the error value is minimized.

Decision tree model: A decision tree consists of root nodes, branch nodes, and leaf nodes, which can progressively partition the data set based on the input features,

and ultimately be able to predict the class of the target variable based on the values of the features. We use the"rpart"package to train the decision tree model, using two hyperparameters, cp, and split. cp is the complexity parameter of the tree, which can be used as a penalty factor to control the size of the tree, and split controls the splitting rule of the tree."gini"is set as the splitting rule and the result will be constructed as a CART decision tree model. The initial cp value is set to 0.001, then the optimal cp value is found and pruned based on the optimal cp value to finally output the decision tree model.

Model evaluation

To evaluate the predictive models using the test dataset, we plotted receiver operating characteristic curves(ROC), calculated the corresponding AUC, accuracy, sensitivity, specificity, precision, recall, PPV and NPV to evaluate the model discrimination. Brier scores and calibration curves were used to evaluate the calibration of the model, and clinical decision curve analysis (DCA) was used to evaluate the net clinical benefit of the model.

Model interpretation

Understandability of predictive models is important in clinical practice.SHAP theory, derived from cooperative game theory, provides a rigorous and highly interpretable tool. We analyze and discuss models with the highest overall performance using SHAP-based explanatory methods to reveal their predictive mechanisms and characteristic contributions.

Statistical analysis

All statistical analyses were performed using R software (version 4.1.4, https://www.rproject.org/),Python(version 3.9.12) software,and SPSS (version 24.0, IBM).

Continuous variables: different representations are used depending on the type of their distribution. For normally distributed data, mean ± standard deviation (SD) is used; for non-normally distributed data, median and interquartile range (IQR) are used.

Categorical variables: expressed as frequencies and percentages. In the table, these data will be presented as "n (%)", where "n" denotes the number of samples in a category and "%" denotes the proportion of the total sample in that category.

Between-group comparisons: for normally distributed data, the t-test was used for between-group comparisons; for non-normally distributed data, the Mann–Whitney U-test was used for between-group comparisons. Between-group comparisons for categorical variables were tested using the chi-square test. The original hypothesis (H0) was that the distribution of the two groups of data was the same, and the alternative hypothesis (H1) was that the distribution of the two groups of data was different. The two-sided significance level for all tests was set at five percent ($p < 0.05$).

Results

Baseline characteristics of participants

A total of 773 patients were enrolled in the study, out of which 357 patients were diagnosed with complicated appendicitis (46.2%). In the CAP group, age, body temperature, heart rate, eosinophil percentage, platelets, CRP, blood glucose, and urea nitrogen were significantly higher than in the UAP group, and systolic blood pressure, monocyte percentage, erythrocyte percentage, and total bilirubin were significantly lower in the CAP group than in the UAP group. Also compared to the UAP group, 89% of patients in the CAP group had WTOG24 h. Other key demographics, clinical presentation, and laboratory findings are summarized in Table 1.

Distribution of variable screening results and data sets

Appendicitis type was used as an outcome variable, and there were 31 predictor variables. The results of feature selection performed by Mean Decrease Accuracy are shown in Fig. 1. We included the top 15 predictor variables in the subsequent model construction, which were, in order, CRP、WTOG24 h、Muscular tone、HR、Platelet、Body temperature、UN、Eosinophils、SBP、Blood glucose、Monocytes、Age、Neutrophils、Lymphocytes and Erythrocyte. After removing redundant variables, patients were divided into a training set ($n = 580$) and a test set ($n = 193$) in a ratio of 3:1, and all features of the samples were comparable in the training and test data sets (Supplementary Table 1).

Comparison with other prediction models

First, we train the model on XGBoost, RF, SVM, and CART in the training dataset, respectively. In the training set, XGBoost, RF, and SVM perform very well in terms of discrimination with AUCs of 0.996, 1.0, and 0.951 and accuracies of 96.7%, 99.7%, and 91.2%, respectively, while possessing very good sensitivity, specificity, NPV,and precision, while performing the worst in the CART model with AUCs and accuracies of 0.82 and 80.3% (Table 2). Although XGBoost, RF, and SVM models perform well in the training set, it does not mean that they have better generalization ability.

Subsequently, we validated the model in the test dataset, and the results show that the XGBoost model exhibits fairly good discrimination, with AUC, sensitivity, specificity, precision, PPV, NPV, and accuracy of 0.914, 86.5%, 84.6%, 82.8%, 84.8%, 89.7%,and 85.5%, respectively, in comparison to RF, which, although it performs the best in the training set, shows a significant decrease

Table 1 Baseline characteristics of patients

Characteristic	Total ($n = 773$)	UAP ($n = 416$)	CAP($n = 357$)	*P* value
Gender(n,%)				0.109
Female	328(42)	188(45)	140(39)	
Male	445(58)	228(55)	217(61)	
Age,year(IQR)	41(30,55)	37(28,50)	46(33,61)	0.000
Residential area,(n,%)				0.088
Rural	461(60)	236(57)	225(63)	
Urban	312(40)	180(43)	132(37)	
ULDS, (*n*, %)				0.005
No	619(80)	349(84)	270(76)	
Yes	154(20)	67(16)	87(24)	
BT,℃(IQR)	37.3(36.8,37.9)	37.1(36.7,37.6)	37.5(36.9,38.2)	0.000
SBP,mmHg(IQR)	123(110,135)	128(113,139)	118(108,132)	0.000
DBP,mmHg(IQR)	75(68,83)	76(69,83)	75(67,83)	0.174
HR, Times/min(IQR)	85(75,101)	80(72,92)	94(79,108)	0.000
Appetite, (*n*, %)				0.858
Good	158(20)	88(21)	70(20)	
Moderate	223(29)	118(28)	105(29)	
Bad	392(51)	210(50)	182(51)	
Nausea or vomit, (*n*, %)				0.722
No	342(44)	187(45)	155(43)	
Yes	431(56)	229(55)	202(57)	
Diarrhea, (*n*, %)				0.492
No	727(94)	394(95)	333(93)	
Yes	46(6)	22(5)	24(7)	
MRLAP, (*n*, %)				0.226
No	336(43)	172(41)	164(46)	
Yes	437(57)	244(59)	193(54)	
Pain level, (*n*, %)				0.291
Slight	202(26)	118(28)	84(24)	
Moderate	410(53)	216(52)	194(54)	
Severe	161(21)	82(20)	79(22)	
Muscular tone, (*n*, %)				0.000
No	541(70)	269(65)	272(76)	
Slight	149(19)	125(30)	24(7)	
Obvious	83(11)	22(5)	61(17)	
TRLA, (*n*, %)				0.340
No	4(1)	1(0)	3(1)	
Yes	769(99)	415(100)	354(99)	
RPRLA, (*n*, %)				0.035
No	184(24)	112(27)	72(20)	
Yes	589(76)	304(73)	285(80)	
Leu,10^9/L(IQR)	9.3(6.5,13.8)	9.7(6.3,13.6)	9.1(6.6,14.1)	0.938
Neu,%(IQR)	77.2(66.8,86.4)	78.0(67.0,85.8)	76(66.5,86.6)	0.956
Lym,%(IQR)	14.2(7.8,22.4)	14(7.77,21.72)	14.6(7.9,23.1)	0.720
Mon,%(IQR)	6.5(4.6,8.4)	6.9(4.7,8.9)	6.2(4.5,7.8)	0.001
Eos,%(IQR)	0.8(0.1,2.1)	0.7(0.1,1.7)	1.2(0.1,2.6)	0.029
HB,g/L(IQR)	132(118,144)	131(118,144)	132(119,145)	0.821
Ery,10^{12}/L(IQR)	4.4(3.9,4.9)	4.5(4.0,5.0)	4.3(3.9,4.7)	0.004
Platelet,10^9/L(IQR)	210(170,269)	201(164,239)	232(179,303)	0.000

Table 1 (continued)

Characteristic	Total ($n=773$)	UAP ($n=416$)	CAP($n=357$)	*P* value
CRP,mg/L(IQR)	49.6(22.6,86.0)	43.8(12.6,64.5)	62.8(30.9,130.7)	0.000
BG,mmol/L(IQR)	5.3(4.6,6.6)	5.1(4.4,6.2)	5.9(4.9,7.3)	0.000
UN,mmol/L(IQR)	4.2(3.3,5.4)	3.8(3.0,4.9)	4.6(3.6,6.0)	0.000
Cr,μmol/L(IQR)	68(55,84)	69(56,81)	68(55,86)	0.562
TB,μmol/L(IQR)	13.6(9.5,19.4)	14.1(10.0,19.7)	13.0(8.8,19.1)	0.008
Appendix fecalith, (*n*, %)				0.006
No	535(69)	306(74)	229(64)	
Yes	238(31)	110(26)	128(36)	
WTOG 24 h, (*n*, %)				0.000
No	226(29)	186(45)	40(11)	
Yes	547(71)	230(55)	317(89)	

ULDS Underlying disease, *BT* Body temperature, *SBP* Systolic blood pressure, *DBP* Diastolic blood pressure, *HR* Heart rate, *MRLAP* Metastatic right lower abdominal pain, *TRLA* tenderness in the right lower abdomen, *RPRLA* Rebound pain in the right lower abdomen, *HB* Hemoglobin, *CRP* C-reactive protein, *UN* Urea nitrogen, *WTOG24 h* Whether the time from pain to operation is longer than 24 h, *Leu* Leukocyte, *Neu* Neutrophils, *Lym* Lymphocytes, *Eos* Eosinophils, *Mon* Monocytes, *Cr* Creatinine, *Ery* Erythrocyte, *TB* Total bilirubin, *BG* blood glucose

* $p < 0.05$ (significant)

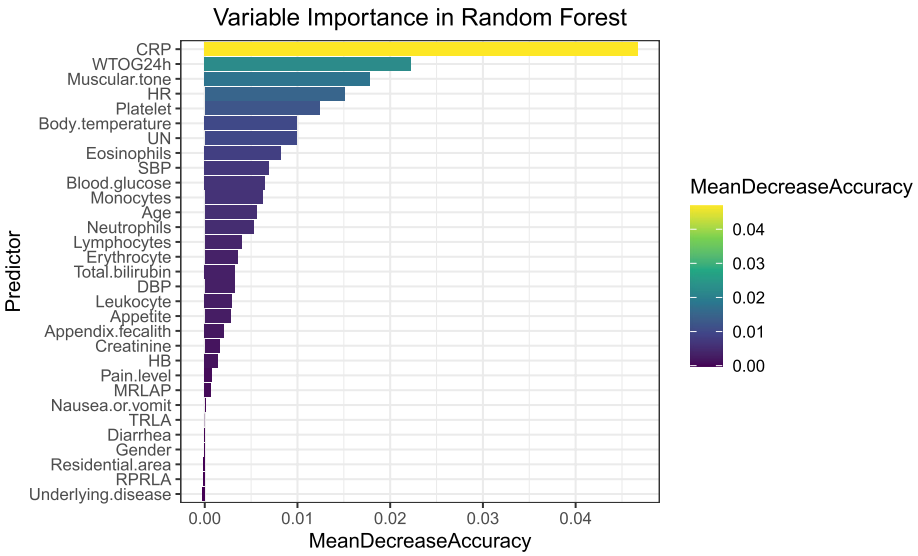

Fig.1 Variable importance graph. The larger the Mean Decrease Accuracy value corresponding to the predictor variable, the greater the impact on the accuracy of the model's predictions

in the test set in terms of specificity, PPV, and accuracy both decreased significantly, 69.2%, 71.2%, and 78.2%, respectively. The diagnostic performance of the SVM model also showed good discrimination in the test dataset, with AUC, sensitivity, NPV, and accuracy of 0.882, 86.5%, 87.1%, and 81.9%, respectively, and specificity and accuracy showing lower 77.9% and 77%, which is still a little less than XGBoost's performance in general. The corresponding ROC curves for the four models are shown in Fig. 2. To further assess the value of these predictive models for clinical applications, we plotted DCA curves (Fig. 3) using the test dataset for evaluating the expected net benefits of the models at the corresponding risk thresholds. The results showed that the XGBoost, RF, and SVM models all showed substantial net benefits, with the XGBoost model showing superior net clinical benefit.

To understand the accuracy of the model predictions, we used calibration curves in our test dataset to show the gap between the actual probabilities and the predicted probabilities as shown in Fig. 4, it is obvious that the XGBoost and SVM models show the best

Table 2 Diagnostic performance of different machine learning models

Outcome	Dataset	Xgboost	Random forest	CART	Support vector machine
AUC	Training	0.996(0.991,1.000)	1.000(1.000–1.000)	0.820(0.786–0.855)	0.951(0.932–0.970)
	Test	0.914(0.874,0.955)	0.870(0.819–0.921)	0.742(0.671–0.813)	0.882(0.833–0.931)
Sensitivity(%)	Training	95.9(94.3,97.5)	100.0(100.0,100.0)	72.8(69.2,76.4)	94.4(92.5,96.3)
	Test	86.5(81.7,91.3)	88.8(84.4,93.2)	71.9(65.6,78.2)	86.5(81.7,91.3)
Specificity(%)	Training	97.4(96.1,98.7)	99.4(98.8,100.0)	86.9(84.2,89.6)	88.5(85.9,91.1)
	Test	84.6(79.5,89.7)	69.2(62.7,75.7)	75.0(68.9,81.1)	77.9(72.0,83.8)
Precision(%)	Training	97.0(95.6,98.4)	99.3(98.6,100.0)	82.6(79.5,85.7)	87.5(84.8,90.2)
	Test	82.8(77.5,88.1)	71.2(64.8,77.6)	71.1(64.7,77.5)	77.0(71.1,82.9)
Accuracy(%)	Training	96.7(95.3,98.2)	99.7(98.8,100.0)	80.3(76.9, 83.5)	91.2(88.6, 93.4)
	Test	85.5(80.5,90.5)	78.2(71.7,83.8)	73.6(66.8, 79.7)	81.9(75.7, 87.0)
PPV(%)	Training	97.1(95.7,98.5)	99.3(98.6,100.0)	82.6(79.5,85.7)	87.5(84.8,90.2)
	Test	84.8(79.7,89.9)	71.2(64.8,77.6)	71.1(64.7,77.5)	77.0(71.1,82.9)
NPV(%)	Training	96.6(95.1,98.1)	1.00(100.0,100.0)	78.8(75.5,82.1)	94.9(93.1,96.7)
	Test	89.7(85.4,94.0)	87.8(83.2,92.4)	75.7(69.6,81.8)	87.1(82.4,91.8)

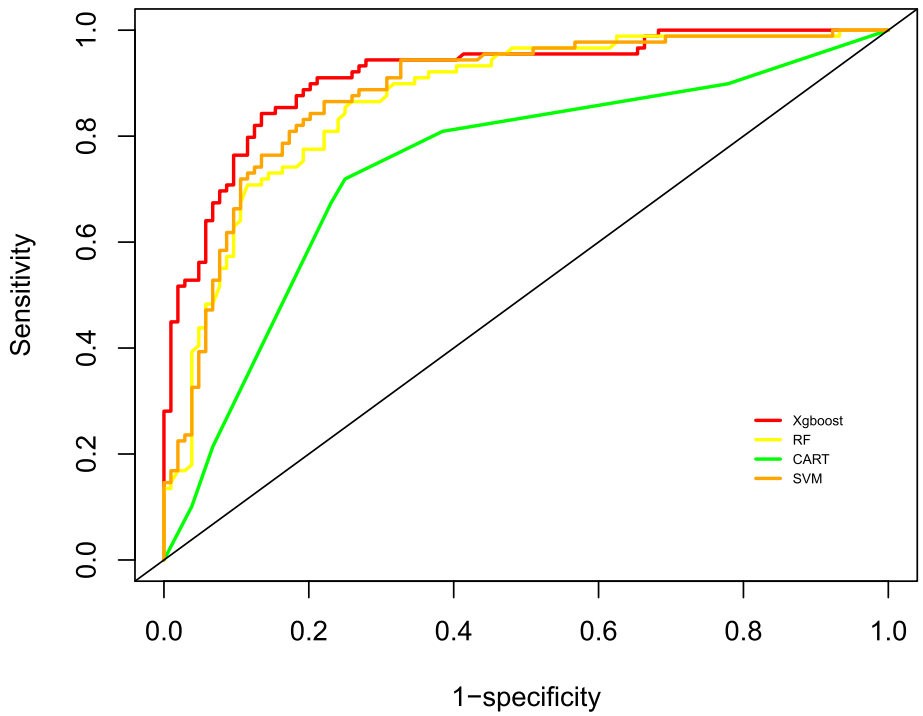

Fig. 2 ROC curves for XGBoost, RF, CART, and SVM models in the test dataset

calibration ability, which indicates that the predictions of the models are highly consistent with the actual values. Finally, we also calculated the Brier scores for the XGBoost, RF, SVM, and CART models, which were 0.118, 0.151, 0.202, and 0.139, respectively, indicating that all four prediction models have some predictive ability.

Feature-based model interpretation

We calculated and visualized the SHAP values for each feature in the Xgboost model. The swarm plot (Fig. 5A) shows an overview of the contribution of all patient features. In the swarm plot, features are sorted by the sum of the magnitude of the SHAP values of all samples, and the SHAP values are used to show the distribution of the

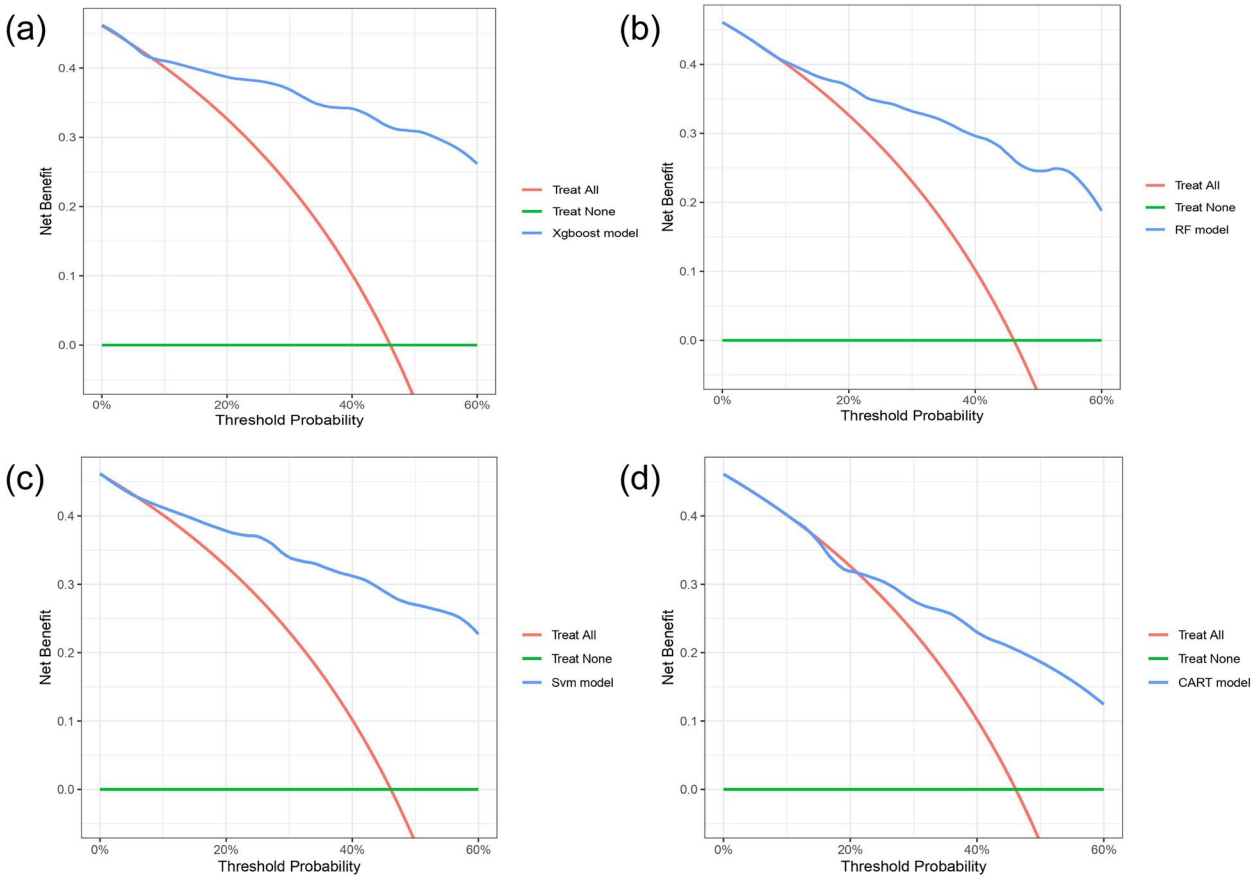

Fig. 3 DCA curves for (**a**) XGBoost, (**b**) RF, (**c**) SVM, and (**d**) CART models in the test dataset

impact of each feature. The bar graph shown in Fig. 5B illustrates the mean absolute value of the SHAP values for each feature. The graph shows that CRP,heart rate and WTOG24 h have higher mean values (Fig. 6B).

In addition Figs. 6A and B show the specific contribution of each feature in a single sample to the model prediction from two perspectives. Among them, CRP and WTOG24 h have the most significant impact on model prediction. low values of CRP and WTOG24 h significantly reduce the probability of model prediction in the negative category. Other characteristics such as age and systolic blood pressure also had an effect, but it was relatively small. Overall, model predictions were primarily driven by CRP and WTOG24 h, two features that are closely associated with disease onset and progression in clinical practice.

Discussion

In this study, XGBoost was introduced for the first time for the prediction of complicated appendicitis. The study focuses on building an XGBoost prediction model for the prediction of complicated appendicitis by combining clinical and laboratory findings and comparing the results with several other traditional machine learning models.Based on the order of importance of the variables, we finally included 9 laboratory indicators and 6 other clinical indicators, all of which were readily available before undergoing appendectomy, the top 5 predictors of which were CRP, WTOG24 h, abdominal muscle tension, heart rate, and platelet count, which is slightly different from other similar studies, which may be related to the methodology of the study and the various parameters included, but similarly, CRP levels were significant in several studies [11, 23–25]. In addition, in order to provide clinicians with a deeper understanding of the model, we introduced SHAP into the Xgboost prediction model to further reveal the decision-making mechanism of the model, and analyzed the total contribution of each feature in the model, and the results showed that CRP and WTOG24 h occupied the most important position in the model's contribution, which was in perfect agreement with our Mean Decrease Accuracy feature selection results. This is fully consistent with our Mean Decrease Accuracy feature selection results.

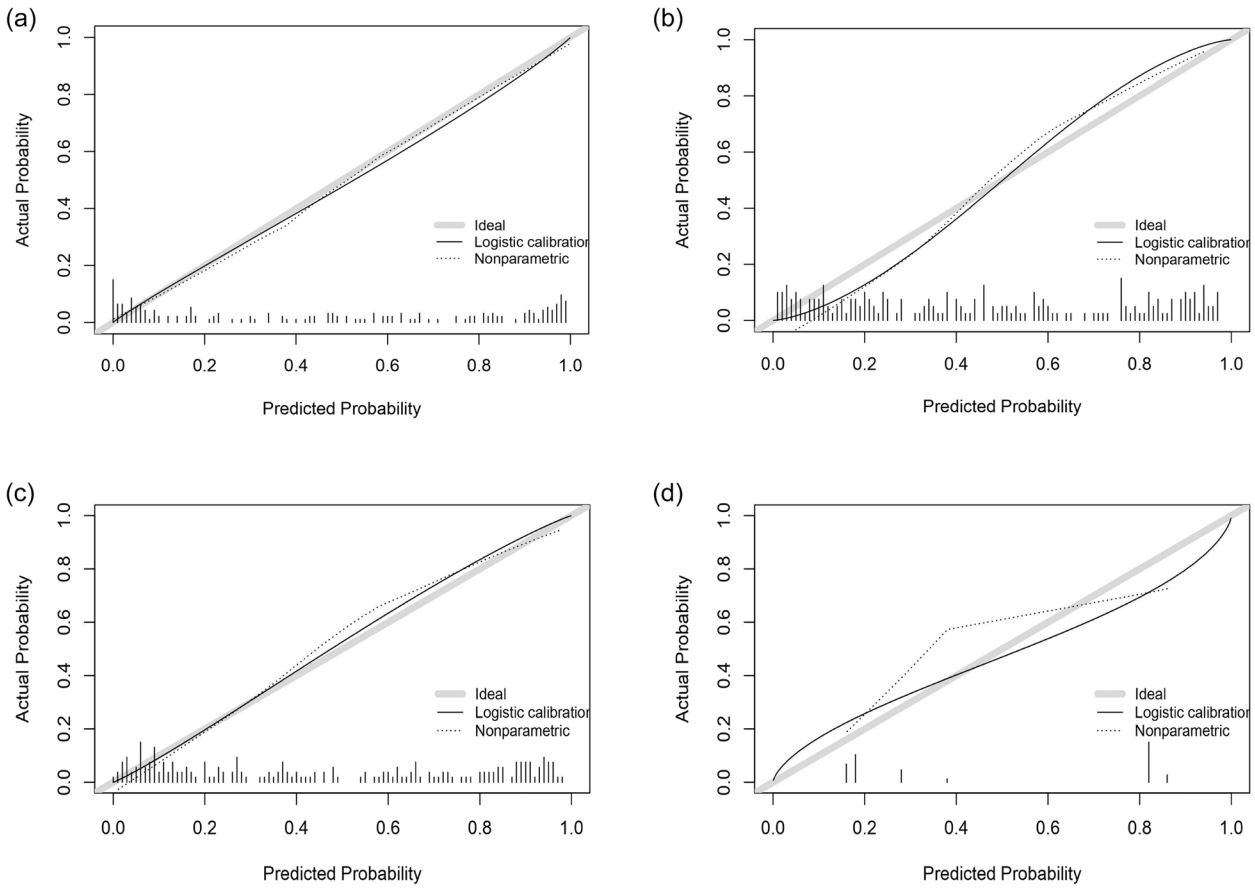

Fig. 4 Calibration curves for (**a**) XGBoost, (**b**) RF, (**c**) SVM, and (**d**) CART models in the test dataset

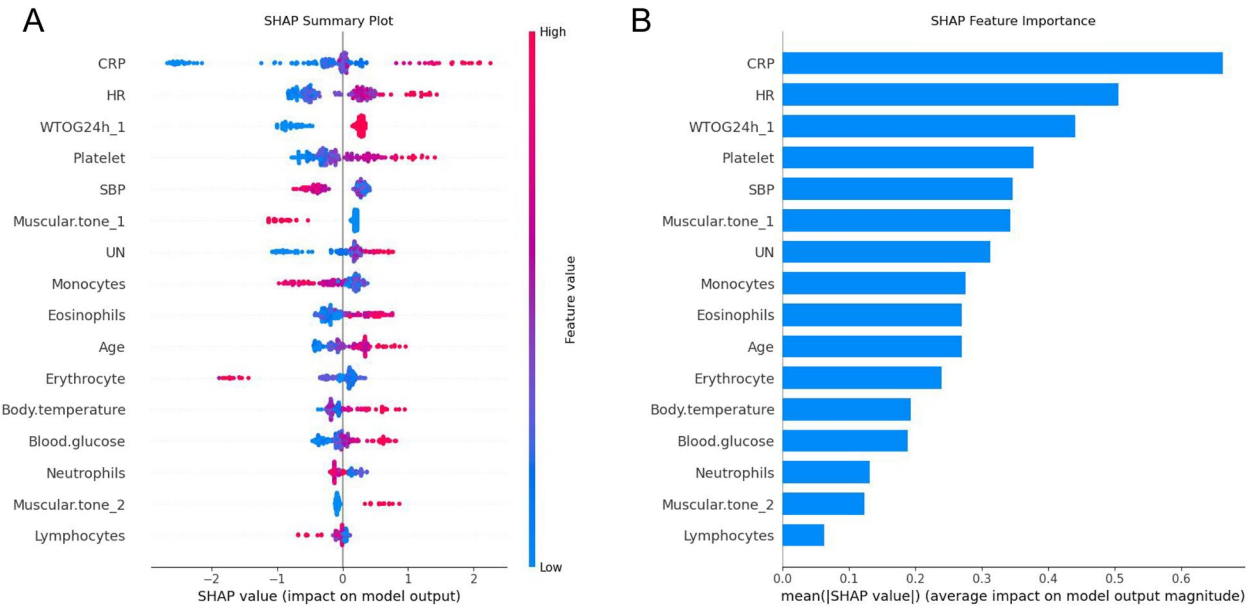

Fig. 5 The SHAP plots illustrated the feature-based model interpretation process. **A** The beeswarm plot used SHAP values to show the distribution of each feature's impacts. **B** The standard bar plot demonstrated the mean absolute value of the SHAP values for each feature

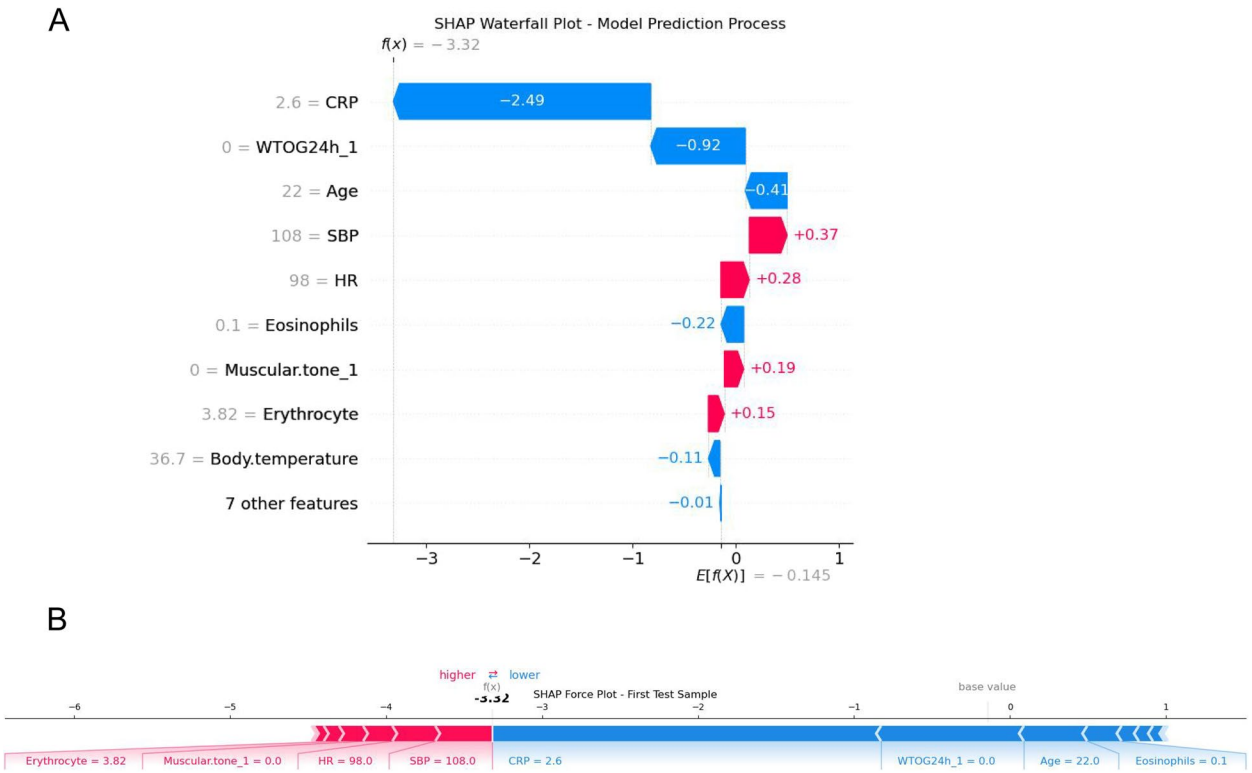

Fig. 6 SHAP plots demonstrated SHAP values from a case-based perspective. Sampled by model output, the overall SHAP plot (**A**) showed the decision process of all patients. The force plot (**B**) and the waterfall plot (**C**) demonstrated the proportion and absolute SHAP value of various features in the decision-making process for a single patient

CRP is a non-specific acute phase reactant that primarily stimulates cell-mediated immunity and chemotaxis in inflammation. It has been shown that CRP levels were significantly higher in perforated appendicitis cases than in non-perforated cases, which is similar to our findings, in addition serum CRP concentration and severe appendicitis were correlated, and elevated CRP was an independent risk factor for predicting perforation in acute appendicitis [26, 27]. Nina A Bickell et al. explored the risk of future appendiceal rupture in patients as the time since symptom onset without treatment increased, showing that the risk was negligible in the first 24 h of symptomatic untreated life, began to rise between 24 and 36 h, and then climbed to six percent in patients who had not been treated 36 h after symptom onset, and the guidelines recommend that for Treatment of patients with appendicitis should be limited to as few as 24 h after the onset of symptoms as possible [28, 29].

Several of the models we built were then internally validated in a separate cohort, and the XGBoost prediction model showed optimal predictive power, which cannot be separated from the superiority of the algorithm itself. Similarly, Qingqing Li et al. [30] constructed a variety of machine-learning models based on gene expression, including XGBoost, DT, SVM, K-Nearest Neighbors (KNN), LR, and RF for breast metastasis-assisted identification, with the XGBoost classifier achieving an overall higher average AUC (0.82); In addition, Tingting Fan et al [31] extracted patient data from the MIMIC-IV database and constructed an XGBoost model for predicting the risk of diabetic ketoacidosis-associated acute kidney injury in ICU patients and compared it with seven other machine learning models, and the XGBoost model performed the best among the eight machine learning models, with AUCs for the training and validation sets of 0.835 and 0.800 for the training and validation sets, respectively. The research on XGBoost will bring this excellent machine-learning method to the attention of more researchers and has many implications for future research. It is highly anticipated that in the future even better researchers will make more in-depth studies on XGBoost in the prediction of complex appendicitis, and even extend XGBoost to more medical fields.

Given the numerous adverse complications of complicated appendicitis, early and accurate diagnosis of the disease can greatly reduce complications. This study is dedicated to determining whether a patient has progressed to complicated appendicitis using only the

patient's preoperative vital signs, laboratory tests, and clinical presentation, The constructed predictive model improves clinical decision-making in patients with suspected complex appendicitis and can be targeted to change high-risk patients to a higher level of care with early preoperative preparation as well as surgical interventions, yet antibiotic therapy is also a safe and effective option for low-risk patients, which can save low-risk patients from acute surgical risks. Dan Liang et al. [32] incorporated clinical features, CT visual features, deep learning features, and radiomics features to develop a model using the CatBoost algorithm for the prediction of complicated appendicitis, and the model was validated at three other medical centers, and the results demonstrated moderate predictive performance with AUCs of 0.836, 0.793, and 0.72, respectively. Also for deep learning radiomics features, the model uses a manual depiction of regions of interest(ROI), which will be challenging to diagnose complicated appendicitis in emergency situations. Hui-An Lin et al. [33] utilized an ANN model to evaluate the diagnostic performance of nine different variable sets in a clinical model, and the results showed that Lin et al.'s model exhibited the most outstanding diagnostic performance, with a sensitivity and specificity of 85.7% and 91.7% in the test dataset, respectively, and that higher sensitivity and specificity could reduce the possibility of leakage and misdiagnosis, and that the high performance of the model was attributed to the incorporation of the periappendiceal fluid and fat stranding signs, which are two very important radiologic Characterization. In contrast, our study demonstrated a very high level of accuracy even without relying on radiologic imaging, and high accuracy means that the diagnostic test correctly identifies the patient's disease state in most cases, whether it is complex or uncomplicated appendicitis, which is critical for clinical decision making.This initiative will bring numerous benefits, first, by reducing the financial burden on patients while eliminating the time spent waiting for CT results; second, by reducing unnecessary radiation exposure to patients; and lastly, by allowing for adjunctive diagnostics in the absence of specialized equipment and personnel to make decisions for surgical decision making.

There are some limitations of our study. First, limited by the research platform and funding, our study was a single-center retrospective study with cases from the same medical institution and a limited number of cases, which may not be representative of a broad population, and we need to obtain a large sample in the future to further validate our findings; second, retrospective studies may have potential bias, which will result in a lack of extrapolation of our study, and we look forward to adopting a multicenter in our future studies, prospective design to improve the generalizability of the findings.

Conclusion

In this study, we developed and validated the XGBoost prediction model for risk assessment of complex appendicitis and compared it with the current mainstream algorithms (SVM, RF and CART). The results show that XGBoost has the strongest discrimination and calibration and also has good generalization ability. The predictive model developed in this study based on key clinical features and laboratory tests improves the ability to discriminate between complex and uncomplicated appendicitis, thus optimizing clinical decision-making. In current diagnostic practice, reliance on subjective clinical presentation and imaging may lead to missed or delayed diagnosis of complicated appendicitis, increasing the risk of complications such as perforation and abscess. This study provides a rapid and reliable diagnostic tool by integrating easily accessible clinical indicators, which fills the gaps in existing diagnostic methods and can reduce unnecessary imaging tests and decrease the risk of complications, thus improving patient prognosis and the efficiency of healthcare resource utilization.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12876-025-03847-6.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Acknowledgements
None.

Authors' contributions
SMC manuscript writing/editing, Data analysis. JFX data analysis. BBX data collection.YH data collection. MMT data collection. JYP participated in the design of the study and revised the manuscript for intellectual content. All authors read and approved the final manuscript.

Funding
This research did not receive any specific funding.

Data availability
The authors confirm that the data supporting the findings of this study are available within supplementary materials.

Declarations

Ethics approval and consent to participate
This research was approved by the ethics committee of Wenzhou Central Hospital. The ethics committee of Wenzhou Central Hospital waived the need for Informed Consent due to the retrospective nature of the study. All procedures were carried out in accordance with the Helsinki Declaration.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 December 2024 Accepted: 3 April 2025
Published online: 24 April 2025

References

1. Lamm R, et al. Diagnosis and treatment of appendicitis: systematic review and meta-analysis. Surg Endosc. 2023;37:8933–90. https://doi.org/10.1007/s00464-023-10456-5.
2. Bom WJ, Scheijmans JCG, Salminen P, Boermeester MA. Diagnosis of uncomplicated and complicated appendicitis in adults. Scandinavian journal of surgery. 2021;110:170–9. https://doi.org/10.1177/14574969211008330.
3. Di Saverio S, et al. Diagnosis and treatment of acute appendicitis: 2020 update of the WSES Jerusalem guidelines. World journal of emergency surgery : WJES. 2020;15:27. https://doi.org/10.1186/s13017-020-00306-3.
4. Sandstrom A, Grieve DA. Hyperbilirubinaemia: its utility in non-perforated appendicitis. ANZ J Surg. 2017;87:587–90. https://doi.org/10.1111/ans.13373.
5. Moris D, Paulson EK, Pappas TN. Diagnosis and management of acute appendicitis in adults: a review. JAMA. 2021;326:2299–311. https://doi.org/10.1001/jama.2021.20502.
6. Andersson M, Kolodziej B, Andersson RE. Validation of the appendicitis inflammatory response (AIR) score. World J Surg. 2021;45:2081–91. https://doi.org/10.1007/s00268-021-06042-2.
7. Coleman JJ, et al. The Alvarado score should be used to reduce emergency department length of stay and radiation exposure in select patients with abdominal pain. The journal of trauma and acute care surgery. 2018;84:946–50. https://doi.org/10.1097/ta.0000000000001885.
8. Gavriilidis P, Angelis N, Evans J, Di Saverio S, Kang P. Hyperbilirubinemia as a predictor of appendiceal perforation: a systematic review and diagnostic test meta-analysis. Journal of clinical medicine research. 2019;11:171–8. https://doi.org/10.14740/jocmr3724.
9. Akçiçek M, Ilgar M, Ünlü S. Is acute appendicitis complicated or uncomplicated? Approaching the question via computed tomography. Acta radiologica. 2023;64:1755–64. https://doi.org/10.1177/02841851221141221.
10. Mahankali SK, Ahamed SA, Gupta GSP, Razek A. CT based acute appendicitis severity index for acute appendicitis and validate its effectiveness in predicting complicated appendicitis. Emerg Radiol. 2021;28:921–7. https://doi.org/10.1007/s10140-021-01950-1.
11. Atema JJ, van Rossem CC, Leeuwenburgh MM, Stoker J, Boermeester MA. Scoring system to distinguish uncomplicated from complicated acute appendicitis. Br J Surg. 2015;102:979–90. https://doi.org/10.1002/bjs.9835.
12. Avanesov M, et al. Diagnostic prediction of complicated appendicitis by combined clinical and radiological appendicitis severity index (APSI). Eur Radiol. 2018;28:3601–10. https://doi.org/10.1007/s00330-018-5339-9.
13. Clifford S, et al. Validation and comparison of two new scoring systems for the prediction of complicated versus uncomplicated appendicitis. Ir J Med Sci. 2023. https://doi.org/10.1007/s11845-023-03594-1.
14. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. Cell. 2018;173:1581–92. https://doi.org/10.1016/j.cell.2018.05.015.
15. Phan-Mai TA, et al. Validity of machine learning in detecting complicated appendicitis in a resource-limited setting: findings from Vietnam. Biomed Res Int. 2023;2023: 5013812. https://doi.org/10.1155/2023/5013812.
16. Yazici H, et al. Predicting severity of acute appendicitis with machine learning methods: a simple and promising approach for clinicians. BMC Emerg Med. 2024;24:101. https://doi.org/10.1186/s12873-024-01023-9.
17. Mei Z, Xiang F, Zhen-hui L. Short-term traffic flow prediction based on combination model of Xgboost-LightGBM.2018 INTERNATIONAL CONFERENCE ON SENSOR NETWORKS AND SIGNAL PROCESSING (SNSP 2018). 2018:322–7. https://doi.org/10.1109/SNSP.2018.00069.
18. Wang J, Li B, Zeng Y. XGBoost-Based Android Malware Detection. 2017 13th International Conference on Computational Intelligence and Security (CIS).2017:268–72. https://doi.org/10.1109/CIS.2017.00065.
19. Chen T,Guestrin C.XGBoost: A Scalable Tree Boosting System.KDD'16: Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining. 2016:785–94. https://doi.org/10.1145/2939672.2939785.
20. Guan X, et al. Construction of the XGBoost model for early lung cancer prediction based on metabolic indices. BMC Med Inform Decis Mak. 2023;23:107. https://doi.org/10.1186/s12911-023-02171-x.
21. Moore A, Bell M. XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction: a UK Biobank cohort study. Clinical Medicine Insights Cardiology. 2022;16: 11795468221133611. https://doi.org/10.1177/11795468221133611.
22. Wang L, Wang X, Chen A, Jin X, Che H. Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. Healthcare (Basel). 2020;8. https://doi.org/10.3390/healthcare8030247.
23. Byun J, Park S, Hwang SM. Diagnostic algorithm based on machine learning to predict complicated appendicitis in children using CT, laboratory, and clinical features. Diagnostics (Basel). 2023;13. https://doi.org/10.3390/diagnostics13050923.
24. Xia J, et al. Performance optimization of support vector machine with oppositional grasshopper optimization for acute appendicitis diagnosis. Comput Biol Med. 2022;143: 105206. https://doi.org/10.1016/j.compbiomed.2021.105206.
25. Imaoka Y, et al. Validity of predictive factors of acute complicated appendicitis. World journal of emergency surgery : WJES. 2016;11:48. https://doi.org/10.1186/s13017-016-0107-0.
26. Akbulut S, et al. An investigation into the factors predicting acute appendicitis and perforated appendicitis. Ulusal travma ve acil cerrahi dergisi. 2021;27:434–42. https://doi.org/10.14744/tjtes.2020.60344.
27. Kubota A, et al. Treatment for appendicitis with appendicolith by the stone size and serum C-reactive protein level. J Surg Res. 2022;280:179–85. https://doi.org/10.1016/j.jss.2022.06.009.
28. Bickell NA, Aufses AH Jr, Rojas M, Bodian C. How time affects the risk of rupture in appendicitis. J Am Coll Surg. 2006;202:401–6. https://doi.org/10.1016/j.jamcollsurg.2005.11.016.
29. Mair A, et al. Safety of in-hospital delay of appendectomy - a propensity score-matched analysis of 4900 consecutive patients undergoing surgery for suspected appendicitis. Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract. 2025;29: 102003. https://doi.org/10.1016/j.gassur.2025.102003.
30. Li Q, et al. XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer. J Transl Med. 2022;20:177. https://doi.org/10.1186/s12967-022-03369-9.
31. Fan T, et al. Predicting the risk factors of diabetic ketoacidosis-associated acute kidney injury: a machine learning approach using XGBoost. Front Public Health. 2023;11: 1087297. https://doi.org/10.3389/fpubh.2023.1087297.
32. Liang D, et al. Development and validation of a deep learning and radiomics combined model for differentiating complicated from uncomplicated acute appendicitis. Acad Radiol. 2024;31:1344–54. https://doi.org/10.1016/j.acra.2023.08.018.
33. Lin HA, Lin LT, Lin SF. Application of artificial neural network models to differentiate between complicated and uncomplicated acute appendicitis. J Med Syst. 2023;47:38. https://doi.org/10.1007/s10916-023-01932-5.

Publisher's Note