ORIGINAL ARTICLE

Development of a diagnostic model for biliary atresia based on MMP7 and serological tests using machine learning

Yong Zhao[1] · An Wang[1] · Dingding Wang[1] · Dayan Sun[1] · Jiawei Zhao[1] · Yanan Zhang[1] · Kaiyun Hua[1] · Yichao Gu[1] · Shuangshuang Li[1] · Junmin Liao[1] · Peize Wang[1] · Jie Sun[1] · Jinshi Huang[1]

Abstract

Objective To develop a machine learning diagnostic model based on MMP7 and other serological testing indicators for early and efficient diagnosis of biliary atresia (BA).

Methods A retrospective analysis was conducted on patient information from those hospitalized for pathological jaundice at Beijing Children's Hospital between January 1, 2019, and December 31, 2023. Patients with serum MMP7, liver stiffness measurements, and other routine serological tests were included in the study. Six machine learning models were constructed, including logistic regression (LR), random forest (RF), decision tree (DET), support vector machine classifier (SVC), neural network (MLP), and extreme gradient boosting (XGBoost), to diagnose BA. The area under the receiver operating characteristic curve was used to evaluate the diagnostic efficacy of the various models.

Results A total of 98 patients were included in the study, comprising 64 BA patients and 34 patients with other cholestatic liver diseases. Among the six machine learning models, the XGBoost algorithm model and RF algorithm model achieved the best predictive performance, with an AUROC of nearly 100% in both the training and validation sets. In the training set, these two algorithm models achieved an accuracy, precision, recall, F1 score, and AUROC of 1. Through model interpretation analysis, serum MMP7 levels, serum GGT levels, and acholic stools were identified as the most important indicators for diagnosing BA. The nomogram constructed based on the XGBoost algorithm model also demonstrated convenient and efficient diagnostic efficacy.

Conclusion Machine learning models, especially the XGBoost algorithm and RF algorithm models, constructed based on preoperative serum MMP7 and serological tests can diagnose BA more efficiently and accurately. The most important influencing factors for diagnosis are serum MMP7, serum GGT, and acholic stools.

Keywords Biliary atresia · Diagnostic model · MMP7 · Machine learning

Yong Zhao and An Wang have contributed equally to this work as co-first authors.

✉ Dingding Wang
  wddlene@126.com

✉ Jinshi Huang
  jsdr2002@126.com

[1] Department of Neonatal Surgery, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China

Introduction

Biliary atresia (BA) is a rare and severe neonatal hepatobiliary disease, mainly characterized by neonatal biliary obstruction, bile stasis, and progressive liver fibrosis. The incidence of BA varies greatly worldwide, with rates of approximately 1:19,800–1:11,800 in Europe and the United States, and higher rates in East Asia [1, 2]. The etiology of BA remains unclear and may be related to various factors such as abnormal biliary tract development in early fetal life, viral infections, and immune disorders. Symptoms usually present as jaundice, acholic stools, and liver malfunction within 3 months after birth. Currently. Early diagnosis and timely Kasai portoenterostomy (KP) can effectively alleviate disease progression. Studies have reported that

KP performed within 60 days after birth can significantly improve patients' prognosis [3]. While the 20-year native liver survival rate of Patients who undergo surgery older than 90 days is less than 20% [4]. Therefore, early and accurate diagnosis of BA is crucial for improving patients' native liver survival.

Matrix metalloproteinase-7 (MMP7) may promote the formation and development of biliary atresia lesions by degrading collagen fibers and other matrix components in the bile duct wall during the pathogenesis of BA [5]. Studies from the United States, mainland China, and Taiwan have confirmed that MMP7 may be the best serum marker for diagnosing biliary atresia with good sensitivity and specificity [5–10].

Machine learning (ML), as a part of artificial intelligence, is being applied in various fields of medical research, such as disease diagnosis, severity assessment, and guiding treatment plans [11–14]. However, there is rare research on BA diagnosing based on serum MMP7 and other routine serological test results. The purpose of this study is to apply machine learning techniques to BA diagnosis based on patients' serum MMP7 and other routine serological test results, with the aim of constructing an efficient and accurate diagnostic model for BA, which will improving patients' native liver survival and prognosis by helping the early diagnosis and advancing the patient's age of surgery,

Methods

Study subjects

We retrospectively investigated the clinical data of children with pathological jaundice admitted to neonatal surgery department, Beijing Children's Hospital from January 1, 2019, to December 31, 2023. To explore the diagnostic efficacy of MMP7 for BA, we only included patients with complete preoperative testing, (MMP7, ultrasound and other laboratory data). This study was approved by the Medical Ethics Committee of Beijing Children's Hospital Affiliated to Capital Medical University, in line with the ethical guidelines of the Declaration of Helsinki (revised in 2013). Informed consent was exempted due to the retrospective nature of this study. The diagnosis of biliary atresia was based on intraoperative anatomical findings and cholangiography.

Serum MMP7 measurement

Serum MMP-7 levels were measured using an enzyme-linked immunosorbent assay (ELISA) kit (R&D Systems, DMP700, Minneapolis, MN, USA) [15]. All measurements were performed by Beijing Mygenostics Co., LTD. All samples were measured in triplicate, and the average serum MMP-7 levels were analyzed.

Liver stiffness measurement

In this study, liver stiffness measurement (LSM) was performed using an Aixplorer ultrasound system (SuperSonic Imagine SA, Aix-en-Provence, France) with a SuperLinear SL15-4 probe. LSM was obtained from an area of liver parenchyma approximately 1 cm in diameter, 0.5–1 cm below the liver capsule, avoiding large vessels. The detection and analysis of LSM were performed by two experienced ultrasound physicians.

Data collection and preparation

Baseline indicators included demographic factors such as gender, age, weight, mode of delivery, birth weight, gestational age at admission, maternal obstetric history, postnatal feeding status, etc. Symptoms such as acholic stool were also included. Preoperative examinations included ultrasound liver stiffness values, serum MMP7 levels, blood routine and biochemical laboratory test results. The final diagnosis of biliary atresia was based on intraoperative surgery findings and cholangiography.

Variables with missing values exceeding 20% were excluded from the analysis. The miceforest package was used to fill in missing values via multiple imputation by chained equations (MICE). To avoid high correlations between different features, we performed Spearman correlation analysis on features after handling missing values. The correlation coefficient of each feature relative to other feature was calculated. Features with high correlation ($r > 0.9$) were excluded from the analysis.

Prediction models and evaluation

We explored six machine learning algorithms, including logistic regression (LR), random forest (RF), decision tree (DET), support vector machine classifier (SVC), multilayer perceptron (MLP) and extreme gradient boosting (XGBoost) to predict BA. All models were developed using Python (version 3.10). The LR, RF, DET, MLP and SVC models were implemented using the Python Sklearn package. The XGBoost model was implemented using the XGbooster package. All patients were randomly divided into a training set and a validation set in a 7:3 ratio. Least absolute shrinkage and selection operator (Lasso) regression was used to perform feature selection on the training set data before inclusion in the model. The model was trained using tenfold cross-validation based on the training set data, and model performance was evaluated based on the average score. After

obtaining the optimal model parameters, validation was performed on the validation set.

Five metrics recording performance were obtained at each iteration, including area under the receiver operating characteristic (ROC) curve (AUROC), accuracy, precision, recall, F1 score. The models on each dataset were compared by the mean AUROC value over ten iterations. The best performing ML model was selected (as measured by its mean AUROC value). Calibration curves were used to assess the consistency between the predicted risk and observed risk of these ML models. The top ten features with the most significant influence on prediction were selected for model explanation using Shapley Additive Explanations (SHAP). Finally, based on the regression coefficients, the best-performing ML model was selected to construct a nomogram for predicting BA in the neonatal.

Data analysis

Patients were divided into BA and control groups, and variables between the two groups were displayed and compared. Continuous variables with normal distribution were analyzed by independent samples t test, and non-normally distributed variables were analyzed by Mann–Whitney U test. Categorical variables were analyzed by chi-square (χ^2) test or Fisher's exact test. All statistical analyses were performed using Python (version 3.10) and R software (4.2.1), with $P < 0.05$ considered significant.

Results

Queue characteristics

We retrospectively studied children who were hospitalized in our center due to pathological jaundice. Patients with incomplete preoperative serum MMP7, ultrasound and other laboratory data were excluded from the analysis (Fig. 1). This study included 98 patients, with 64 BA and 34 non-BA served as controls (Table 1). The non-BA group consist of 20 patients with neonatal hepatitis, 4 patients with digestive tract malformation, 4 patients with Alagille syndrome, 4 patients with progressive familial intrahepatic cholestasis (PFIC), and 2 patients with neonatal intrahepatic cholestasis caused by citrin deficiency (NICCD).

The baseline information, liver stiffness, and laboratory examination results of the two groups of patients are summarized. The incidence of serum MMP7, acholic stool, birth weight, weight at admission, total protein (TP), albumin (ALB), globulin (GLO), total cholesterol (TC), direct bilirubin (DBIL), low density lipoprotein (LDL), gamma-glutamyl transpeptidase (GGT), total bile acid (TBA), platelet (PLT) and Breastfeeding in the BA group was significantly
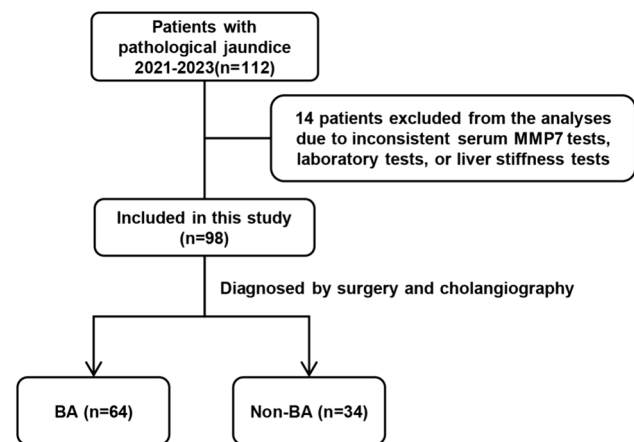

Fig. 1 Study flowchart

higher than that in the control group ($P < 0.05$). The premature birth and para 1 in the BA group were lower than those in the control group ($P < 0.05$). While other clinical features did not show significant statistical differences. To preliminarily explore the diagnostic value of each indicator, we independently constructed ROC curves for MMP7, LSM, TP, GGT, DBIL, TBA, PLT, and acholic stool (Fig. 2). The results showed that MMP7 (AUROC = 0.855, 95% CI 0.818–0.952), GGT (AUROC = 0.834, 95% CI 0.754–0.914), acholic stool (AUROC = 0.777, 95% CI 0.670–0.883), and LSM (AUROC = 0.750, 95% CI 0.645–0.855) had high independent diagnostic value.

Predictive performance of machine learning models

Spearman correlation analysis showed no strong correlation between any features ($r < 0.9$, Fig. 3). Therefore, all factors in the training set were included in the lasso regression for feature selection. Through lasso regression, a total of 10 features were selected for subsequent model construction, including: premature, first delivery, acholic stool, LSM, MMP7, GGT, TBIL, DBIL, VLDL, and PLT (Fig. 4A). Based on the training set data, the best parameters were selected through tenfold cross-validation to reduce model overfitting. The best diagnostic model parameters were obtained using Optuna for model hyperparameter optimization. The performance of all models was validated in the validation set. The results showed XGBoost and RF outperformed the other four models both in the training and validation sets (Fig. 4B and Table 2). It is particularly noteworthy that, due to the high diagnostic specificity of MMP7, both XGBoost and RF exhibited extremely high diagnostic efficacy in the training set under optimal optimization parameters. In the 30 training cases (19 BA and 11 non-BA), the accuracy, precision, recall, F1 score, and AUROC of these two models all reached 1 (Fig. 4B and Table 2). The

Table 1 Patient clinical data and baseline data

Variables	Total ($n=98$)	Non-BA ($n=34$)	BA ($n=64$)	Statistic	P
Age, Mean $\pm$ SD (day)	51.55 $\pm$ 29.76	45.06 $\pm$ 27.38	55.00 $\pm$ 30.60	$t=-1.59$	0.116
Weight, Mean $\pm$ SD (g)	4380.92 $\pm$ 1220.91	3908.53 $\pm$ 1213.92	4631.88 $\pm$ 1157.06	$t=-2.90$	0.005
Birthweight, Mean $\pm$ SD (g)	3064.64 $\pm$ 590.93	2801.18 $\pm$ 717.31	3204.61 $\pm$ 459.11	$t=-2.97$	0.005
LSM, Mean $\pm$ SD (kPa)	15.22 $\pm$ 18.16	8.68 $\pm$ 6.29	18.69 $\pm$ 21.26	$t=-3.49$	<0.001
MMP7, Mean $\pm$ SD (ng/ml)	53.44 $\pm$ 79.95	13.46 $\pm$ 6.90	74.68 $\pm$ 92.17	$t=-5.29$	<0.001
TP, Mean $\pm$ SD (g/l)	52.49 $\pm$ 5.67	49.54 $\pm$ 6.44	54.05 $\pm$ 4.54	$t=-3.63$	<0.001
ALB, Mean $\pm$ SD (g/l)	37.32 $\pm$ 4.30	35.22 $\pm$ 5.34	38.43 $\pm$ 3.15	$t=-3.22$	0.002
GLO, Mean $\pm$ SD (g/l)	15.16 $\pm$ 2.87	14.31 $\pm$ 2.35	15.62 $\pm$ 3.04	$t=-2.19$	0.031
TC, Mean $\pm$ SD (mmol/l)	4.23 $\pm$ 1.71	3.67 $\pm$ 1.09	4.53 $\pm$ 1.91	$t=-2.41$	0.018
ALP, Mean $\pm$ SD (U/l)	627.21 $\pm$ 409.26	540.82 $\pm$ 285.72	673.11 $\pm$ 457.12	$t=-1.53$	0.128
AST, Mean $\pm$ SD (U/l)	269.37 $\pm$ 268.45	256.58 $\pm$ 310.77	276.16 $\pm$ 245.43	$t=-0.34$	0.733
ALT, Mean $\pm$ SD (U/l)	162.30 $\pm$ 152.52	133.76 $\pm$ 140.31	177.46 $\pm$ 157.59	$t=-1.36$	0.178
GGT, Mean $\pm$ SD (U/l)	423.85 $\pm$ 360.77	186.88 $\pm$ 161.45	549.75 $\pm$ 374.61	$t=-6.67$	<0.001
TBIL, Mean $\pm$ SD (umol/l)	166.12 $\pm$ 71.07	154.21 $\pm$ 76.09	172.45 $\pm$ 68.02	$t=-1.21$	0.228
DBIL, Mean $\pm$ SD (umol/l)	99.22 $\pm$ 56.90	79.67 $\pm$ 66.57	109.60 $\pm$ 48.46	$t=-2.32$	0.025
IBIL, Mean $\pm$ SD (umol/l)	67.03 $\pm$ 55.79	74.91 $\pm$ 63.35	62.85 $\pm$ 51.37	$t=1.02$	0.311
TBA, Mean $\pm$ SD (umol/l)	119.58 $\pm$ 66.91	93.08 $\pm$ 63.27	133.65 $\pm$ 64.94	$t=-2.97$	0.004
TG, Mean $\pm$ SD (mmol/l)	1.27 $\pm$ 0.58	1.14 $\pm$ 0.53	1.33 $\pm$ 0.59	$t=-1.58$	0.118
HLD, Mean $\pm$ SD (mmol/l)	0.74 $\pm$ 0.39	0.78 $\pm$ 0.43	0.72 $\pm$ 0.36	$t=0.73$	0.467
LDL, Mean $\pm$ SD (mmol/l)	2.76 $\pm$ 1.44	2.29 $\pm$ 0.94	3.00 $\pm$ 1.60	$t=-2.37$	0.020
VLDL, Mean $\pm$ SD (mmol/l)	0.26 $\pm$ 0.12	0.25 $\pm$ 0.13	0.27 $\pm$ 0.12	$t=-0.88$	0.384
PALB, Mean $\pm$ SD (mg/l)	89.14 $\pm$ 39.02	88.15 $\pm$ 37.59	89.67 $\pm$ 40.04	$t=-0.18$	0.855
WBC, Mean $\pm$ SD (*10^9/l)	10.64 $\pm$ 3.53	10.73 $\pm$ 2.31	10.59 $\pm$ 4.05	$t=0.18$	0.856
RBC, Mean $\pm$ SD (*10^{12}/l)	3.53 $\pm$ 0.62	3.68 $\pm$ 0.60	3.46 $\pm$ 0.62	$t=1.69$	0.094
HB, Mean $\pm$ SD (g/l)	111.22 $\pm$ 23.46	116.79 $\pm$ 23.81	108.27 $\pm$ 22.91	$t=1.73$	0.087
PLT, Mean $\pm$ SD (*10^9/l)	371.04 $\pm$ 162.14	306.71 $\pm$ 151.22	405.22 $\pm$ 158.36	$t=-2.98$	0.004
APRI, Mean $\pm$ SD	2.03 $\pm$ 8.90	3.56 $\pm$ 14.89	1.22 $\pm$ 2.02	$t=1.24$	0.217
Gender, n (%)				$\chi^2=1.94$	0.164
No	44 (44.90)	12 (35.29)	32 (50.00)		
Yes	54 (55.10)	22 (64.71)	32 (50.00)		
Caesarean section, n (%)				$\chi^2=0.20$	0.658
No	46 (46.94)	17 (50.00)	29 (45.31)		
Yes	52 (53.06)	17 (50.00)	35 (54.69)		
Premature birth, n (%)				$\chi^2=13.71$	<0.001
No	82 (83.67)	22 (64.71)	60 (93.75)		
Yes	16 (16.33)	12 (35.29)	4 (6.25)		
G1, n (%)				$\chi^2=2.26$	0.133
No	59 (60.20)	17 (50.00)	42 (65.62)		
Yes	39 (39.80)	17 (50.00)	22 (34.38)		
P1, n (%)				$\chi^2=3.90$	0.048
No	48 (48.98)	12 (35.29)	36 (56.25)		
Yes	50 (51.02)	22 (64.71)	28 (43.75)		
Breastfeeding, n (%)				$\chi^2=4.05$	0.044
No	23 (23.47)	12 (35.29)	11 (17.19)		
Yes	75 (76.53)	22 (64.71)	53 (82.81)		
Acholic stools, n (%)				$\chi^2=33.31$	<0.001
No	28 (28.57)	22 (64.71)	6 (9.38)		
Yes	70 (71.43)	12 (35.29)	58 (90.62)		

LSM liver stiffness measurement, *MMP7* Matrix metalloproteinase-7, *TP* total protein, *ALB* albumin, *GLO* globular proteins, *TC* total cholesterol, *ALP* alkaline phosphatase, *AST* aspartate aminotransferase, *ALT* alanine transaminase, *GGT* gamma-glutamyl transpeptidase, *TBIL* total bilirubin, *DBIL* direct bilirubin, *IBIL* indirect bilirubin, *TBA* total bile acid, *TG* triglyceride, *HDL* high-density lipoprotein, *LDL* low-density lipoprotein, *VLDL* very low density lipoprotein, *PALB* prealbumin, *WBC* white blood cells, *RBC* red blood cells, *HB* hemoglobin, *PLT* platelet, *APRI* aspartate aminotransferase-to-platelet ratio index
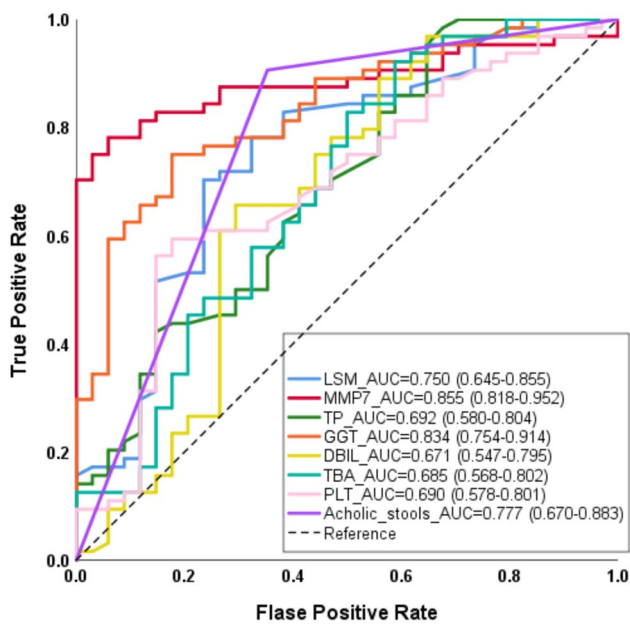
Fig. 2 ROC shows that LSM, MMP7, GGT, and acholic stool have relatively high independent diagnostic efficacy. *ROC curve* receiver operating characteristic curve, *LSM* liver stiffness measurement, *MMP7* matrix metalloproteinase-7, *GGT* gamma-glutamyl transferase

calibration curve shows that 5 models have better consistency between the predicted probability and the actual probability except for the LRM model (Fig. 4C).

Model interpretation and illustration

We used mean SHAP value to screen the ten features with the greatest impact on prediction BA from the six ML models. In multiple models, MMP7, GGT and acholic stool are the three most important features. We also analyzed the positive or negative contributions of features in diagnosing BA (Figs. 5 and 6). The AUROCs for diagnosing BA with only these three features are 0.855 (95% CI 0.818–0.952), 0.834 (95% CI 0.754–0.914), and 0.777 (95% CI 0.670–0.883), respectively.

To visualize the best-performing ML model (XGBoost model), a nomogram integrating 8 selected features was used for early diagnosis of BA (Fig. 7A). In the nomogram, the score corresponding to each feature value is obtained by drawing a line upward to the "Points" axis at a specific point. After obtaining the total score for all features, a line is drawn directly downward from the total score axis to the risk axis to diagnose BA. Based on the nomogram, we constructed an ROC curve, and the results showed that the AUROC reached 0.994 (95% CI 0.985–1), with a specificity of 1.0 and a sensitivity of 0.969 (Fig. 7B). This suggests that the nomogram can be conveniently and efficiently applied to the clinical diagnosis of BA.

Discussion

BA is a severe neonatal hepatobiliary disease, early diagnosis of BA can advance the age at surgery and improve the prognosis of the patient. We screened out the 10 parameters that had the greatest impact on the diagnosis and selected a total of six algorithms for model construction, to avoid bias caused by a single algorithm. Among them, the XGBoost and RF showed extremely high diagnostic accuracy and application value (AUROC = 1), the subsequent calibration curves also confirmed the good reliability of the two models. The XGBoost algorithm is an ensembled deep learning algorithm based on boosted decision tree models, which has the advantages of additional regularization schemes to prevent overfitting and efficient and flexible handling of missing data, and is therefore widely used in the diagnosis and prediction of clinical outcomes of various diseases [16–19]. The RF algorithm is a widely used ensembled deep learning algorithm, which works by constructing multiple decision trees and outputting the class patterns of these trees with high accuracy. It can handle a large number of features with relatively good interpretability, despite some drawbacks such as occupying a large amount of memory space and potentially long training times as the number of trees increases [20, 21]. SHAP model interpretation showed that serum MMP7, serum GGT, and acholic stool were the most important factors influencing the diagnosis of BA, and ROC analysis showed that all three factors had strong independent diagnostic efficacy. In addition, for the convenience of clinical application, we constructed a nomogram based on the XGBoost algorithm model. ROC analysis suggested that this nomogram could be conveniently and efficiently applied to the clinical diagnosis of BA (AUROC = 0.994, 95% CI 0.985–1). The application of ML diagnostic models provides an early and efficient method for BA diagnosis on the basis of traditional clinical diagnostic methods, which is expected for advancing the age at surgery and improve the prognosis of BA patients in the future.

As a member of the matrix metalloproteinase family, MMP7 is widely recognized as the best serum biomarker for the new generation of BA diagnosis [19]. Studies have reported that serum MMP7 levels in BA patients are significantly higher than in other cholestatic diseases [5, 8]. Currently, the AUROC of MMP7 for diagnosing BA is between 0.96 and 0.99, with sensitivity ranging from 0.86 to 1 and specificity ranging from 0.83 to 0.95 [5, 8–10, 22, 23]. However, the cut-off values for MMP7 in diagnosing BA vary widely between studies, ranging from 1.43 ng/mL to 52.85 ng/ml. The reasons for this may be due to different kit types, and ethnic differences [19, 24]. Serum MMP-7 levels are significantly positively correlated with
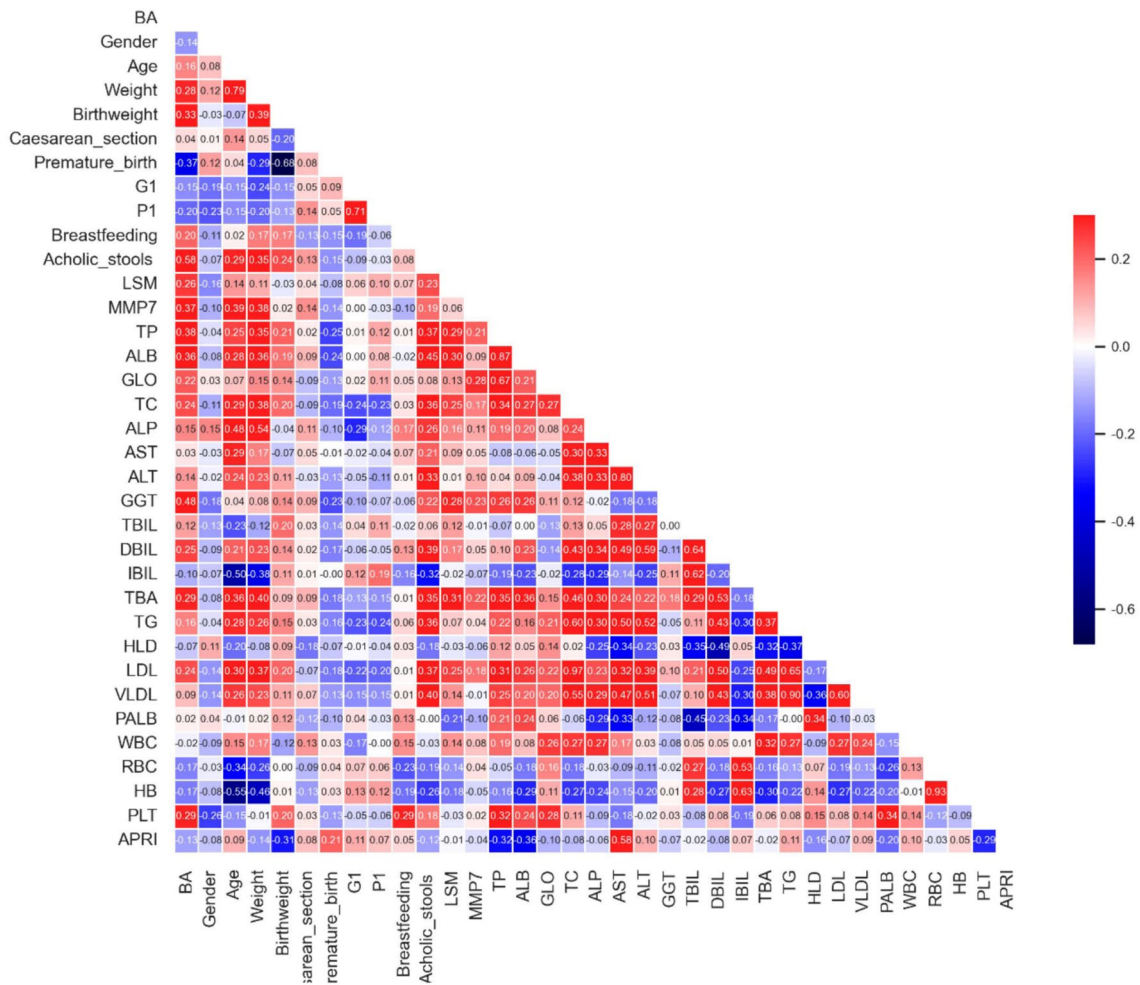
Fig. 3 Spearman correlation analysis of all features in the cohort

the age of the child, as MMP-7 is mainly produced by adenoepithelial cells, bile duct cells, and Kupffer cells in the liver, and older children usually have more severe liver fibrosis, more MMP7 is released into the blood. However, studies have reported that serum MMP-7 levels in neonates diagnosed with BA within 30–35 days are significantly lower than that in older infants, so the applicability of serum MMP-7 screening for BA in neonates (< 30 days) requires further study [20, 21]. Most scholars believe that serum MMP7 is positively correlated with the degree of liver fibrosis [9, 20, 21, 24]. However, Lertudomphonwanit et al. reported that serum MMP-7 was poorly correlated with liver fibrosis staging at diagnosis [5]. Despite some shortcomings, MMP7 is still undoubtedly an excellent serum diagnostic marker for BA. In this study, the cut-off value for MMP7 was 21.725 ng/ml, and the AUROC for MMP7 alone in diagnosing BA was 0.855 (95% CI 0.818–0.952). Its diagnostic efficacy alone was lower than in previous studies, which may be due to the higher proportion of neonates in this cohort. In this study, we

applied ML models to combine serum MMP7 with other clinical indicators, which can compensate for the potential deficiencies of MMP7 alone. Thanks to the extremely high diagnostic potential of MMP7, the RF and XGBoost models achieved nearly 100% diagnostic efficacy in both the training and validation sets.

Gamma-glutamyl transferase (GGT) is mainly derived from liver cells and, as an enzyme, catalyzes the metabolic process of glutathione [25]. Elevated serum GGT levels often suggest cholestasis, alcoholic liver disease, or fatty liver. It has been reported that elevated serum GGT levels help differentiate BA from other cholestatic conditions [26, 27]. Wu et al. reported that serum GGT levels had higher diagnostic efficacy for BA than MMP-7 [22]. However, more studies showed opposite conclusion [7, 9, 28]. GGT is significantly positively correlated with liver fibrosis, and studies have shown that GGT and glutathione act together on the extracellular membrane surface of bile duct epithelial cells to generate free radicals and exert pro-oxidative effects [29]. Elevated GGT levels after KP suggest a poor prognosis in
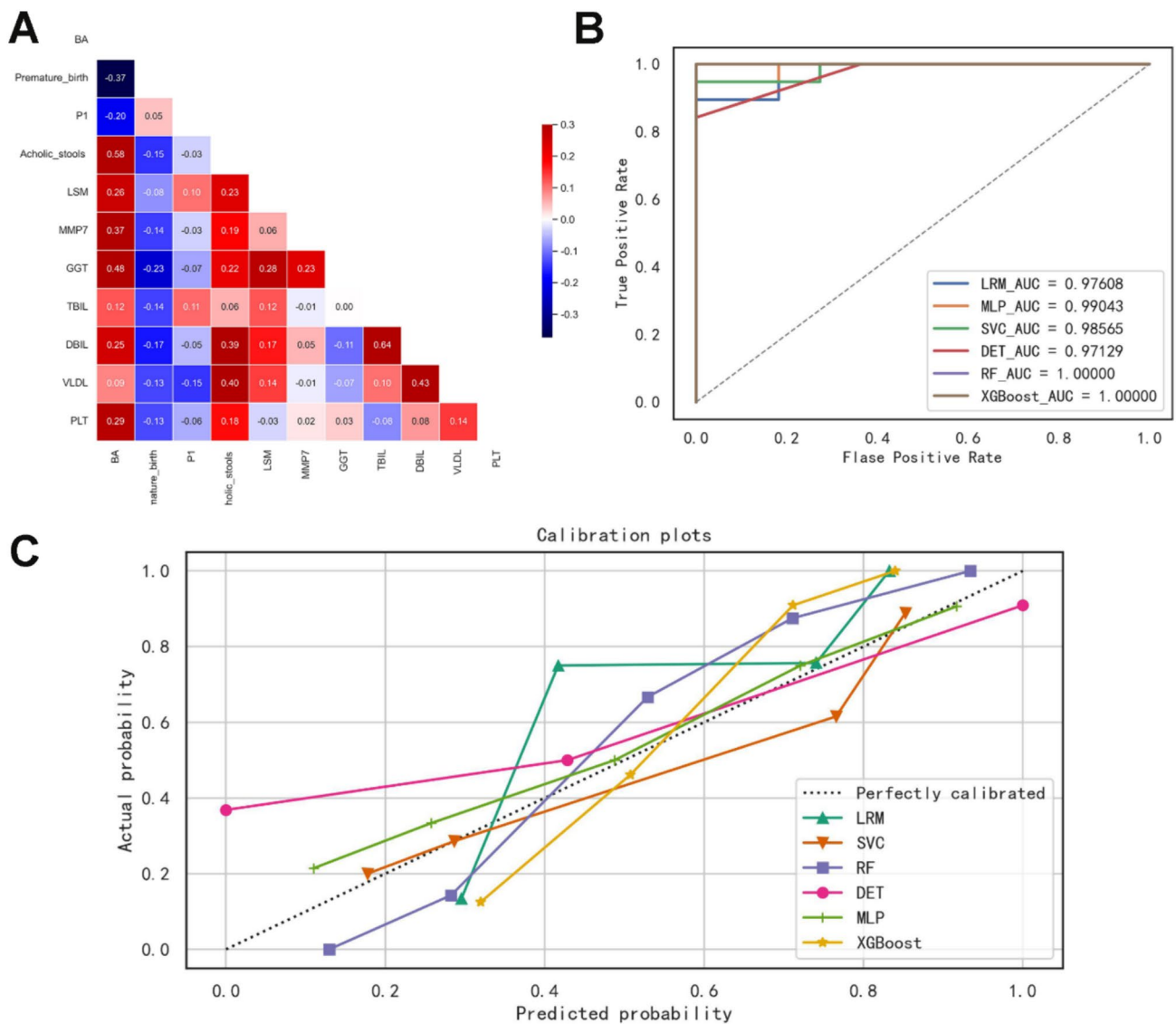
Fig. 4 **A** Correlation analysis of 10 features extracted by LASSO regression. **B** Area under the receiver operating characteristic curve (AUC) of all six machine learning models including logistic regression (LR), multilayer perceptron (MLP), support vector machine classifier (SVC), decision tree (DET), random forest (RF), and extreme gradient boosting (XGBoost). **C** Calibration curves for all six machine learning models. The calibration curve shows the consistency between predicted risk (x-axis) and actual risk (y-axis). The calibration curve shows that 5 models have better consistency between the predicted probability and the actual probability except for the LRM model

Table 2 Diagnostic efficacy of six machine learning diagnostic models on the training and validation sets

| Algorithms	Average performance of cross-validation on the training set					Performance on the validation set				
	Precision	Recall	F1 score	Accuracy	AUROC (95% CI)	Precision	Recall	F1 score	Accuracy	AUROC (95% CI)
LRM	0.884	0.980	0.923	0.886	0.963 (0.931–0.996)	0.857	0.947	0.900	0.867	0.976 (0.924–1.000)
MLP	0.890	0.940	0.909	0.871	0.943 (0.879–1.000)	0.900	0.947	0.923	0.900	0.990 (0.958–1.000)
SVC	0.890	0.940	0.909	0.871	0.955 (0.913–0.997)	0.857	0.947	0.900	0.867	0.986 (0.945–1.000)
DET	0.931	0.855	0.873	0.840	0.878 (0.801–0.956)	1.000	0.842	0.914	0.900	0.971 (0.914–1.000)
RF	0.935	0.955	0.938	0.914	0.990 (0.969–1.000)	1.000	1.000	1.000	1.000	1.000 (1.000–1.000)
XGBoost	0.947	0.890	0.904	0.886	0.990 (0.969–1.000)	1.000	1.000	1.000	1.000	1.000 (1.000–1.000)
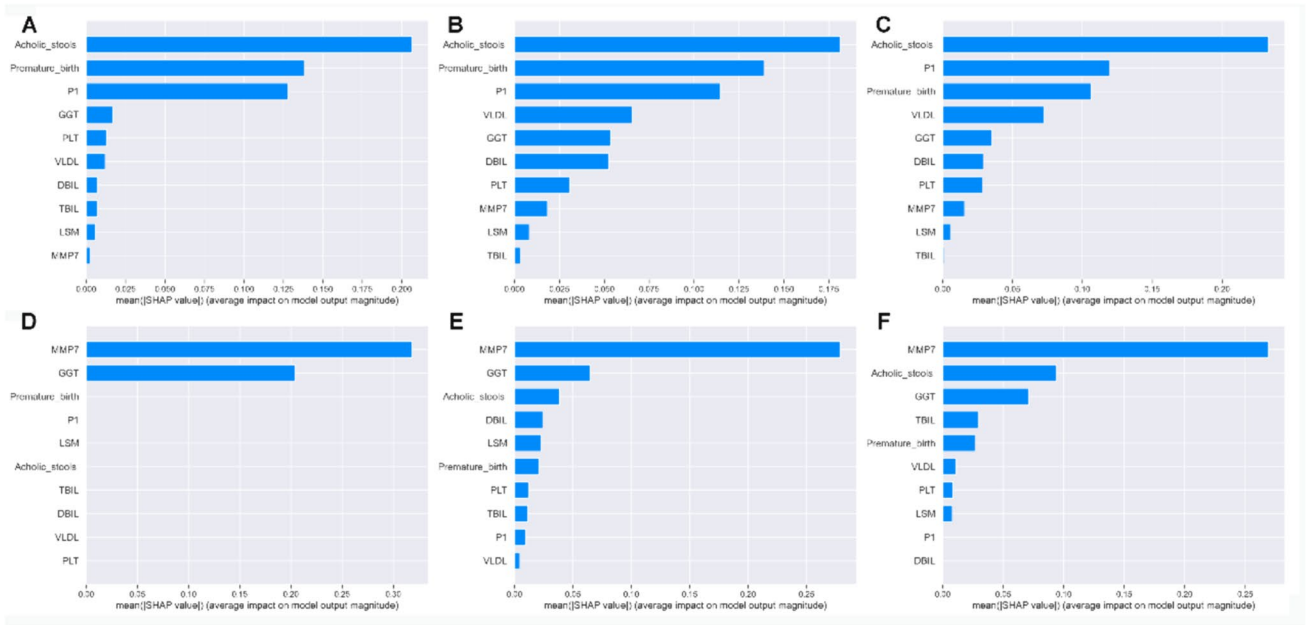
Fig. 5 Mean SHAP value was used to explain the importance of the top ten features in (**A**) logistic regression (LR), (**B**) multilayer perceptron (MLP), (**C**) support vector machine classifier (SVC), (**D**) decision tree (DET), (**E**) random forest (RF), and (**F**) extreme gradient boosting (XGBoost) models, with each model screening out the top ten features. *SHAP* Shapley additive explanations
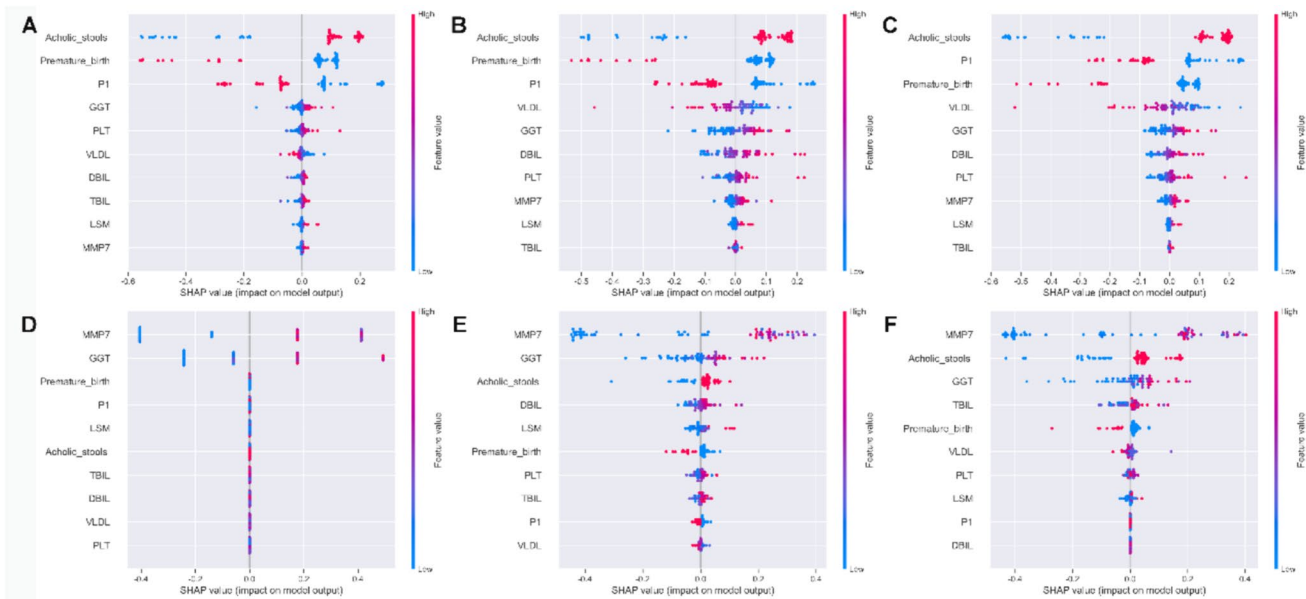

Fig. 6 SHAP value was used to explain the contribution of the top ten features to the prediction results in (**A**) logistic regression (LR), (**B**) multilayer perceptron (MLP), (**C**) support vector machine classifier (SVC), (**D**) decision tree (DET), (**E**) random forest (RF), and (**F**) extreme gradient boosting (XGBoost) models. Every point represents a data sample, with the color indicating whether the observed value of the feature is higher (red) or lower (blue) in diagnosing BA. *SHAP* Shapley additive explanations, *BA* biliary atresia

children, and Ihn et al. reported that children with serum GGT concentrations > 550 IU/L at 5 months after KP had lower 5-year native liver survival [30]. In this study, GGT alone had a high diagnostic efficacy, elevated GGT levels was ranked as one of the top three most important factors

for diagnosing BA by three algorithm models (DET, RF, and XGBoost).

Acholic stool is caused by obstruction of the extrahepatic bile ducts, which prevents normal drainage of bile into the intestinal lumen, resulting in progressively darker urine and

Fig. 7 **A** Nomogram based on the XGBoost model for the probability of diagnosing BA. **B** ROC curve shows that the nomogram has high diagnostic efficacy. *BA* biliary atresia
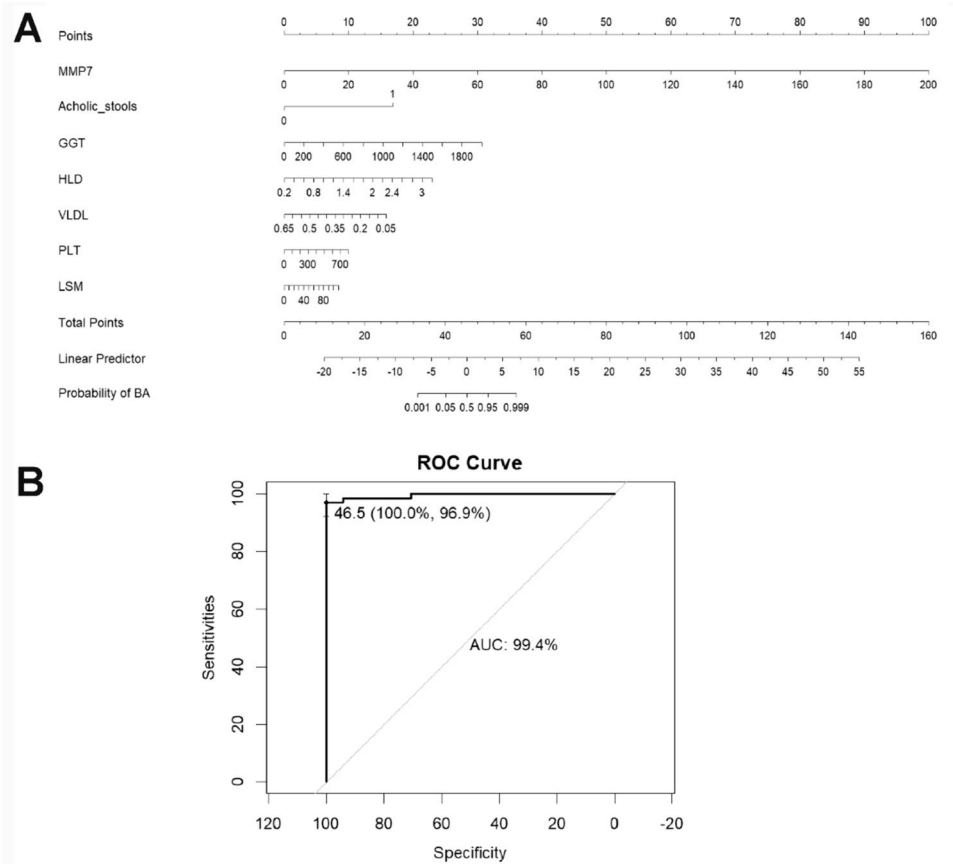
lighter stool color, even white like clay. A meta-analysis of 7 studies including 610 patients, reported an overall sensitivity of 87% and specificity of 78% for acholic stool in diagnosing BA [31]. In 2004, the infant stool color card (ISCC) was first proposed in Taiwan to early identify BA, which had a significant economic benefits and improved the prognosis of patients. Subsequently, there have been reports of its usage in Japan and the United States [32, 33]. In recent years, scholars in Japan and Italy have used machine learning technology to develop relevant mobile apps which could diagnose BA by analyzing photos of neonatal stools [34, 35]. In this study, five models out of six ranked acholic stool as one of the most three important predictive factors, and acholic stool alone also had high diagnostic efficacy, reflecting its value as a simple and non-invasive preoperative diagnostic indicator.

There are still some limitations in this study. First, this is a single-center retrospective study, which may lead to selection bias. Second, since there are certain differences in the baseline clinical data between the two groups of patients (such as birth weight, weight at admission, etc.), it may partially affect the effectiveness of the diagnostic model. In addition, the sample size of this study is limited, and a larger sample size of the training set is expected to further improve the effectiveness of the model. Therefore,

larger-scale prospective multicenter clinical cohort studies are needed in the future.

In conclusion, the XGBoost and RF models constructed by combining serum MMP7 and other serological characteristics can diagnose BA more efficiently. The most important factors affecting the model are serum MMP7 level, serum GGT level, and acholic stool. The nomogram based on the ML model can also be conveniently and efficiently applied to the clinical diagnosis of BA. However, it is still necessary to conduct prospective and multicenter large-sample studies to verify and optimize the model.

Author contributions Y.Z., A.W., D.D.W. and J.S.H. designed the researches; A.W., Y.Z., D.Y.S. and Y.N.Z. conducted data collection and follow-up; D.D.W., A.W., Y.Z. and T.Y. contributed to the analysis of the data; K.Y.H. and Y.C.G. conceived and directed the project; D.D.W., A.W., S.S.L., J.M.L. and D.Y.S. wrote the manuscript with the assistance and feedback of all the other co-authors.

Data availability The datasets during and/or analysis during the current study available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Hartley JL, Davenport M, Kelly DA (2009) Biliary atresia. Lancet (London, England) 374(9702):1704–1713

2. Lakshminarayanan B, Davenport M (2016) Biliary atresia: a comprehensive review. J Autoimmun 73:1–9

3. Lin JS, Chen SC, Lu CL et al (2015) Reduction of the ages at diagnosis and operation of biliary atresia in Taiwan: a 15-year population-based cohort study. World J Gastroenterol 21(46):13080–13086. https://doi.org/10.3748/wjg.v21.i46.13080

4. Nio M, Wada M, Sasaki H et al (2015) Effects of age at Kasai portoenterostomy on the surgical outcome: a review of the literature[J]. Surg Today 45(7):813–818. https://doi.org/10.1007/s00595-014-1024-z

5. Lertudomphonwanit C, Mourya R, Fei L, Zhang Y, Gutta S, Yang L, Bove KE, Shivakumar P, Bezerra JA (2017) Large-scale proteomics identifies MMP-7 as a sentinel of epithelial injury and of biliary atresia. Sci Transl Med. 9:eaan8462

6. Harpavat S (2019) MMP-7: the next best serum biomarker for biliary atresia? J Pediatr 208:8–9. https://doi.org/10.1016/j.jpeds.2019.01.026. (**Epub 2019 Mar 8 PMID: 30857775**)

7. Rohani P, Mirrahimi SB, Bashirirad H, Rahmani P, Kamran N, Alimadadi H, Hajipour M, Sohouli MH (2022) Serum matrix metalloproteinase-7 levels in infants with cholestasis and biliary atresia. BMC Pediatr 22(1):351. https://doi.org/10.1186/s12887-022-03409-9. (**PMID:35717157;PMCID:PMC9206322**)

8. Yang L, Zhou Y, Xu PP, Mourya R, Lei HY, Cao GQ et al (2018) Diagnostic accuracy of serum matrix metalloproteinase-7 for biliary atresia. Hepatology 68(6):2069–2077

9. Jiang J, Wang J, Shen Z, Lu X, Chen G, Huang Y et al (2019) Serum MMP-7in the diagnosis of biliary atresia. Pediatrics. https://doi.org/10.1542/peds.2019-0902

10. Wu JF, Jeng YM, Chen HL, Ni YH, Hsu HY, Chang MH (2019) Quantification of serum matrix metallopeptide 7 levels may assist in the diagnosis and predict the outcome for patients with biliary atresia. J Pediatr 208:30–7.e1

11. Mao HM, Huang SG, Yang Y, Cai TN, Guo WL (2023) Using machine learning models to predict the surgical risk of children with pancreaticobiliary maljunction and biliary dilatation. Surg Today 53(12):1352–1362. https://doi.org/10.1007/s00595-023-02696-8

12. Han X, Geng J, Zhang XX, Zhao L, Wang J, Guo WL (2023) Using machine learning models to predict acute pancreatitis in children with pancreaticobiliary maljunction. Surg Today 53(3):316–321. https://doi.org/10.1007/s00595-022-02571-y

13. Marcinkevics R, Reis Wolfertstetter P, Wellmann S, Knorr C, Vogt JE (2021) Using machine learning to predict the diagnosis, management and severity of pediatric appendicitis. Front Pediatr 9:662183. https://doi.org/10.3389/fped.2021.662183

14. Huang SG, Qian XS, Cheng Y, Guo WL, Zhou ZY, Dai YK (2021) Machine learning-based quantitative analysis of barium enema and clinical features for early diagnosis of short-segment Hirschsprung disease in neonate. J Pediatr Surg 56(10):1711–1717. https://doi.org/10.1016/j.jpedsurg.2021.05.006

15. Chi S, Peipei X (2022) Dynamic analysis of serum MMP - 7 and its relationship with disease progression in biliary atresia: a multicenter prospective study. Hepatol Int 16(4):954–963

16. Deng YH, Luo XQ, Yan P, Zhang NY, Liu Y, Duan SB (2022) Outcome prediction for acute kidney injury among hospitalized children via extreme gradient boosting algorithm. Sci Rep 12:8956

17. Rankin S, Han L, Scherzer R, Tenney S, Keating M, Genberg K et al (2022) A machine learning model for predicting mortality within 90 days of dialysis initiation. Kidney360 3:1556–1565

18. Yilmaz A, Hayiroglu MI, Salturk S, Pay L, Demircali AA, Coskun C et al (2022) Machine learning approach on high risk treadmill exercise test to predict obstructive coronary artery disease by using p, qrs, and t waves' features. Curr Probl Cardiol 48:101482

19. Harpavat S (2019) MMP-7: The Next Best Serum Biomarker for Biliary Atresia. J Pediatr 05(208):8–9. https://doi.org/10.1016/j.jpeds.2019.01.026

20. Mengna Z (2024) STRAS:a snakemake pipeline for genome-wide short tandem repeats annotation and score. Hum Genet. https://doi.org/10.1007/s00439-024-02662-5

21. Hunt ML, Blackburn GA, Siriwardena GM et al (2023) Using satellite data to assess spatial drivers of bird diversity. Remote Sens Ecol Conserv 9:483–500

22. Wu J-F, Lee C-S, Lin W-H, Jeng Y-M, Chen H-L, Ni Y-H et al (2018) Transient elastography is useful in diagnosing biliary atresia and predicting prognosis after hepatoportoenterostomy. Hepatology 68:616–624

23. Han YJ, Hu SQ, Zhu JH et al (2023) Accurate prediction of biliary atresia with an integrated model using MMP-7 levels and bile acids. World J Pediatr. https://doi.org/10.1007/s12519-023-00779-7

24. Sakaguchi H, Konishi K-I et al (2022) Serum matrix metalloproteinase-7 in biliary atresia: a Japanese multicenter study. Hepatology Res : Off J Jpn Soc Hepatol 52(5):479–487. https://doi.org/10.1111/hepr.13753

25. Lee DH, Buijsse B, Steffen L et al (2009) Association between serum gamma glutamyl transferase and cardiovascular mortality varies by age: the Minnesota heart survey. Eur J Cardiovasc Prev Rehabil 16(1):16–20

26. Robie DK, Overfelt SR, Xie L (2014) Differentiating biliary atresia from other causes of cholestatic jaundice. Am Surg 80(9):827–831

27. Chen X, Dong R, Shen Z et al (2016) Value of gamma-glutamyl transpeptidase for diagnosis of biliary atresia by correlation with age. J Pediatr Gastroenterol Nutr 63(3):370–373

28. Tang X, Lv Y, Pu L, Ma J, Jin S, Xiang B (2021) Matrix metalloproteinase-7 as a diagnostic marker for biliary atresia: a systematic review and meta-analysis. Indian J Surg 84:682–689

29. Everhart JE, Wright EC (2013) Association of gamma-glutamyl transferase (GGT) activity with treatment and clinical outcomes in chronic hepatitis C (HCV). Hepatology 57(5):1725–1733

30. Ihn K, Ho IG, Chang EY, Han SJ (2018) Correlation between gamma-glutamyl transpeptidase activity and outcomes after Kasai portoenterostomy for biliary atresia. J Pediatr Surg 53(3):461–467. https://doi.org/10.1016/j.jpedsurg.2017.10.001

31. Wang Li, Yang Y et al (2018) Early differential diagnosis methods of biliary atresia: a meta-analysis. Pediatr Surg Int 34(4):363–380. https://doi.org/10.1007/s00383-018-4229-1

32. Jimenez-Rivera C et al (2013) International incidence and outcomes of biliary atresia. J Pediatr Gastroenterol Nutr 56(4):344–354

33. Wang KS (2015) Newborn screening for biliary atresia. Pediatrics 136(6):e1663–e1669

34. Angelico R, Liccardo D, Paoletti M et al (2021) A novel mobile phone application for infant stool color recognition: an easy and effective tool to identify acholic stools in newborns. J Med Screen 28(3):230–237

35. Hoshino E, Hayashi K, Suzuki M et al (2017) An iPhone application using a novel stool color detection algorithm for biliary atresia screening. Pediatr Surg Int 33(10):1115–1121. https://doi.org/10.1007/s00383-017-4146-8