BMC Medicine

RESEARCH

Open Access

Development of an artificial intelligence-based multimodal diagnostic system for early detection of biliary atresia

Ya Ma[3†], Yuancheng Yang[1,2†], Yuxin Du[1,2], Luyang Jin[1,2], Baoyu Liang[1,2], Yuqi Zhang[1,2], Yedi Wang[3], Luyu Liu[3], Zijian Zhang[3], Zelong Jin[3], Zhimin Qiu[3], Mao Ye[4], Zhengrong Wang[3*] and Chao Tong[1,2*]

Abstract

Background Early diagnosis of biliary atresia (BA) is crucial for improving patient outcomes, yet remains a significant global challenge. This challenge may be ameliorated through the application of artificial intelligence (AI). Despite the promise of AI in medical diagnostics, its application to multimodal BA data has not yet achieved substantial breakthroughs. This study aims to leverage diverse data sources and formats to develop an intelligent diagnostic system for BA.

Methods We constructed the largest known multimodal BA dataset, comprising ultrasound images, clinical data, and laboratory results. Using this dataset, we developed a novel deep learning model and simplified it using easily obtainable data, eliminating the need for blood samples. The models were externally validated in a prospective study. We compared the performance of our model with human experts of varying expertise levels and evaluated the AI system's potential to enhance its diagnostic accuracy.

Results The retrospective study included 1579 participants. The multimodal model achieved an AUC of 0.9870 on the internal test set, outperforming human experts. The simplified model yielded an AUC of 0.9799. In the prospective study's external test set of 171 cases, the multimodal model achieved an AUC of 0.9740, comparable to that of a radiologist with over 10 years of experience (AUC = 0.9766). For less experienced radiologists, the AI-assisted diagnostic AUC improved from 0.6667 to 0.9006.

Conclusions This AI-based screening application effectively facilitates early diagnosis of BA and serves as a valuable reference for addressing common challenges in rare diseases. The model's high accuracy and its ability to enhance the diagnostic performance of human experts underscore its potential for significant clinical impact.

Keywords Biliary atresia, Multimodal, Artificial intelligence, Cholestasis, Children

†Ya Ma and Yuancheng Yang contributed equally to this work and are listed as co-first authors.

*Correspondence:
Zhengrong Wang
wenzcip@163.com
Chao Tong
tongchao@buaa.edu.cn
Full list of author information is available at the end of the article

Background

Biliary atresia (BA) is a rare idiopathic fibro-obliterative cholangiopathy that occurs during the perinatal period [1, 2]. The incidence of BA varies from 1:20,000 in North America and Europe to 1:5000–1.1:10,000 in Asia [3–7]. It is the leading cause of liver transplantation in children. Without treatment, BA inevitably progresses to end-stage liver disease and results in death within the first 2 years of life. The Kasai portoenterostomy procedure, followed by liver transplantation, has dramatically prolonged survival and improved quality of life since 1968 [8]. Early age at the time of the Kasai portoenterostomy is one of the most important factors in predicting surgical outcomes, with the best results achieved when performed before 30 to 45 days of age [9–11]; after that, the chance of success diminishes significantly with age. Unfortunately, the median age at the time of the Kasai procedure is approximately 60 days, with no improvement over time [9, 10]. The major cause for delay is the absence of effective and practical screening methods, thus presenting early diagnosis as a prominent and persisting clinical challenge.

Clinical profiles and noninvasive indicators could raise clinical suspicion and prompt further investigation. However, no single method in isolation can establish the diagnosis of BA. Most infants with BA have normal prenatal monitoring results and typically appear healthy at birth without feeding intolerance or growth retardation. Jaundice may be present early but is usually indistinguishable from physiological jaundice and other causes of neonatal cholestasis, including anatomic, infectious, genetic, metabolic, inflammatory, and endocrine conditions. Several laboratory findings, including elevated conjugated bilirubin and gamma-glutamyltransferase (GGT), have been proven to help facilitate BA recognition. However, the reliability of these serum markers alone is limited in terms of identifying BA. Ultrasound (US) is typically the initial and most commonly used imaging modality for neonatal cholestasis. Among various US features, gallbladder abnormalities and the "triangular cord" sign (TCS) are regarded as the most supportive and widely accepted BA indicators [12–15]. Nevertheless, the US is highly operator-dependent, and the sensitivity of the TCS varies dramatically in different studies, ranging from 17 to 100% [16]. Surgical exploration and intraoperative cholangiography, although invasive and associated with radiation risks, are still necessary to establish a definitive diagnosis because of the lack of pathognomonic symptoms or unequivocal biomarkers are [17]. However, a significant proportion of neonatal cholestasis cases do not have BA, and the issue of "over-testing" deserves special concern.

An accurate and noninvasive method for detecting BA largely depends on effectively integrating multimodal medical data, including the clinical manifestations, laboratory tests, and US features from the centers with significant experience. However, the combination and optimization of clinical data require multi-disciplinary collaboration, which can be limited due to the additional consumption of medical resources. This is especially true when dealing with rare clinical conditions, where institutions and clinicians with substantial diagnostic expertise are hard to acquire. Given the low incidence of the disease, the limited accuracy of noninvasive tests, and the imperative of early diagnosis and intervention, it is time to shift our attention to the field of artificial intelligence-aided (AI-aided) diagnostic technology [18, 19]. Among AI techniques, Deep learning offers significant advantages in data analysis [20–25].

Deep learning has shown promising results in the diagnosis of BA based on gallbladder US images [26–28], achieving an area under the curve (AUC) of 0.956 [26]. However, diagnosis methods that rely solely on gallbladder images may overlook other relevant information and may not be adequate for complex cases in which the gallbladder cannot be recognized or exhibits minimal morphological changes. Up to now, efforts to effectively utilize multimodal BA data with AI technology have not yielded significant breakthroughs. Current AI methods encounter challenges when handling multimodal medical data, including significant morphological variations and dispersed information among the data modalities [29–32].

In this study, we aimed to enhance the diagnosis of BA by fusing multimodal medical data using deep learning. To emulate clinical diagnostic strategies, we collected clinical data related to BA diagnosis and built a multimodal dataset, which is currently the largest dataset available in the literature. Our proposed multimodal deep learning method allowed for precise and noninvasive BA diagnosis. To address the challenges of noticeable differences and scattered information across data types, we introduced an attention mechanism for intra- and inter-modality fusion to aggregate critical and informative data from each modality. In addressing the common problem of missing modalities in multimodal data, our approach diverged from conventional interpolation methods by incorporating prior knowledge into our method. To further assess the generalization ability of the model, we conducted prospective validation using an external test cohort.

Methods

Data collection and processing

This study was divided into two parts. The first part was a retrospective study. In this part, we reviewed the medical records of patients retrospectively from November

2016 to August 2022 at a single tertiary center, which served as the national referral center for BA. The study received approval from the Institutional Review Board of our center. Infants aged < 6 months with conjugated hyperbilirubinemia (the ratio of the direct to total bilirubin levels > 20% when the total direct bilirubin serum was ≥ 85 μmol/L and ≥ 17.1 μmol/L when the total bilirubin was < 85 μmol/L) and those suspected of having BA were enrolled in this study. The exclusion criteria included (1) patients with unclear final diagnosis and (2) patients with unavailable data. Reference standard: The diagnosis of BA was ruled out if cholangiography showed a patent biliary tree or if there was recovery from cholestasis during the clinical follow-up period. Confirmation of the diagnosis was achieved through surgical exploration and intraoperative cholangiogram. We also randomly selected some infants who had no significant liver disease. The second part of the study was a prospective study conducted from September 2022 to November 2023 at the same center. Patients who met the aforementioned criteria were enrolled in the study. This subset of patients was designated as the external test dataset. The sample allocation flowchart is shown in Fig. 1A.

Demographic characteristics, medical histories, and laboratory test results were collected from the patient's medical records. In addition to US images of the gallbladder and TCS, we also incorporated images of the liver capsule and parenchyma as input data for BA classification. Small nodules and unevenly thickened liver capsules are frequently observed on the outer surface of the liver during surgery in BA patients. While these manifestations are not commonly utilized in clinical settings, we included the US features of the liver parenchyma and liver capsule to test if AI exhibits heightened sensitivity towards these subtle changes. The region of interest (ROI) box was marked with a rectangle covering the minimum area of the corresponding position. After initial annotations were completed by the primary experts, these annotations underwent a peer review by a second independent annotator. Seven clinical information variables were included: gender, age, body weight, preterm status, jaundice, clay stool, and dark urine. Fourteen laboratory test data variables were included: total bilirubin (TB), direct bilirubin (DB), DB-to-TB ratio, total bile acid (TBA), total protein (TP), albumin (ALB), alanine aminotransferase (ALT), aspartate aminotransferase (AST), gamma-glutamyl transpeptidase (GGT), alkaline phosphatase (ALP), platelets (PLTs), white blood cells (WBCs), prothrombin time (PT), and activated partial thromboplastin time (APTT). The input data for the multimodal model comprised 4 types of two-dimensional US images and 21 variables (7 clinical information and 14 laboratory test parameters).

Multimodal model

To enhance the integration of medical data in various forms, our model comprised four essential components that enabled effective fusion and adaptation to different data modalities: data mapping and enhancement, feature extraction, modality fusion, and multi-loss joint training modules. These modular parts were plug-and-play modules that could be freely switched to satisfy different task requirements.

As shown in Fig. 1C, the data were processed and mapped to the input format. Then, the data were passed through the feature extraction module of each modality to obtain a high-dimensional semantic feature representation. For the selection of the feature extraction modules, we used Swin Transformer V2 [33] as the visual encoder and MLP as the numeric data encoder. Following this, the feature fusion module undertook a comprehensive analysis of the semantic feature integration. We proposed a novel feature fusion mechanism called Self-Masked Attention for visual intra-modality fusion and MLP for inter-modality fusion (Additional file 1: Supplementary Information S2).

In the training phase, we combined and arranged each patient's four types of images, and then connected them

(See figure on next page.)

Fig. 1 Flow diagram of the study, multimodal model design, and retrospective results. **A** Flow diagram of the inclusion and exclusion of patients in the study. **B** ROC curves of the multimodal model and four experts. **C** Overview of the proposed multimodal deep model method. This model mainly comprises four parts: data mapping and enhancement, feature extraction, modality fusion, and multi-loss joint training modules. Each part can be switched freely to satisfy different task requirements. In the feature extraction stage, we use the Swin transformer as a visual encoder to extract features from the input images. During the modality fusion stage, we employ a self-masked attention mechanism to fuse visual features, followed by the fusion of different modalities using a multi-layer perceptron (MLP). The fused features are subsequently used for classification and prediction. **D** Changes in diagnostic outcomes of radiologists when assisted by the multimodal model on the internal test dataset. Illustration of changes in AUC, sensitivity, specificity, accuracy, PPV, and NPV of the four experts before and after the assistance of the model. Circles represent the diagnostic outcomes of the radiologists when the diagnosis was established independently, squares represent the diagnostic outcomes when aided by the model, and stars denote the performance of the model. Expert 1, Expert 2, and Expert 3 refer to radiologists with more than 10 years, more than 5 years, and more than 1 year of experience in pediatric US, respectively. Expert 4 refers to a radiologist without experience in pediatric US

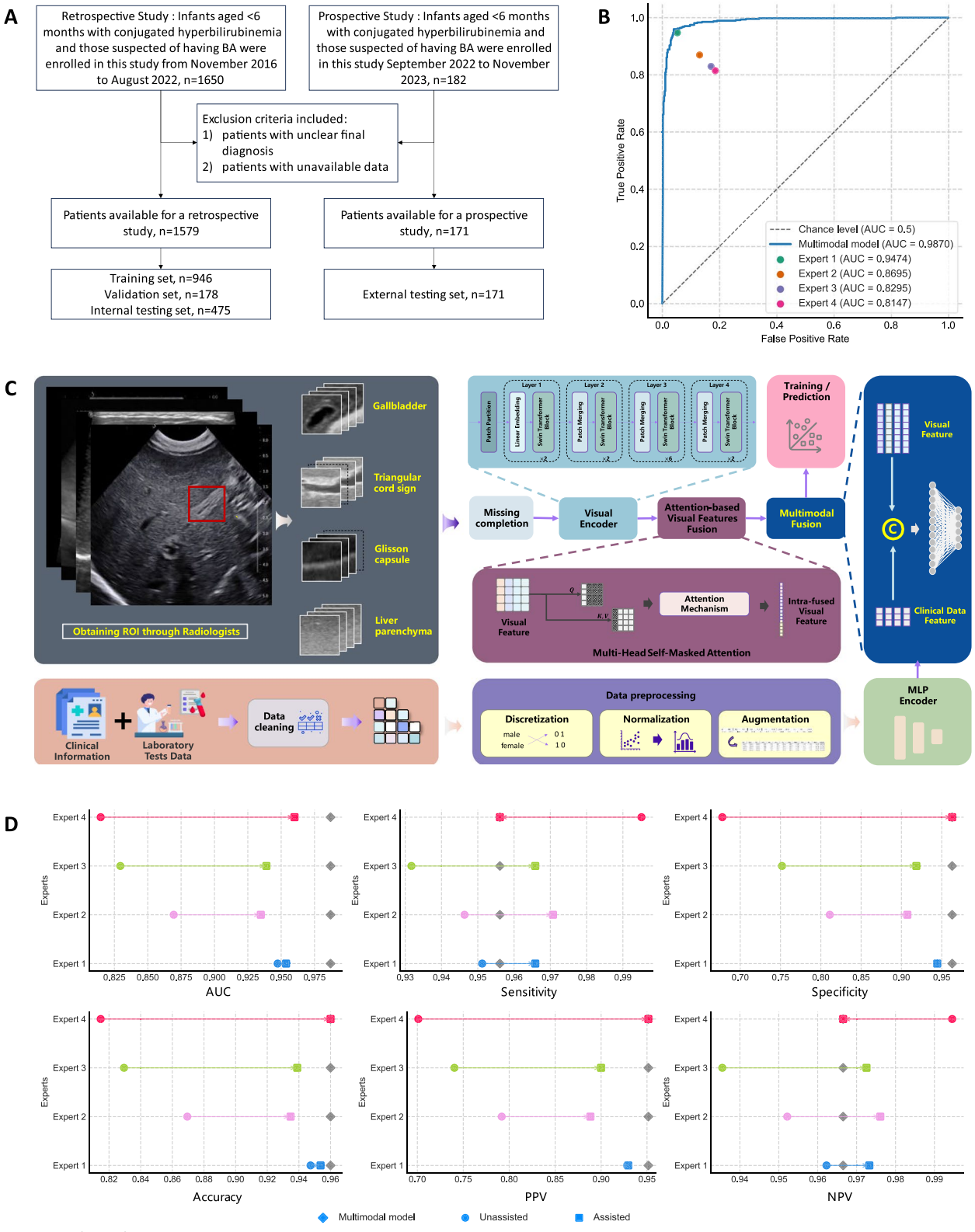

Fig. 1 (See legend on previous page.)

with the corresponding clinical and laboratory test data to form the input data. Each patient contributed multiple input samples, and their diagnostic annotations served as the labels for each sample. This method enabled the model to comprehensively learn the associations between each position and modality within the combination data.

Simplified multimodal model

Our model was built on a modular framework that enabled the flexible interchangeability of components to satisfy different task requirements. To identify the meaningful input data in the BA diagnosis process and optimize the applicability of our model in clinical settings, we selected and combined different modalities of data and trained a more compact model for prediction (Additional file 1: Fig. S8). In this simplified model, we reduced the number of encoders and heads while retaining the MLP-based fusion mechanism for modality fusion. Notably, the training and inference procedures of the simplified model remained consistent with those of the main model to ensure the robustness and accuracy of the simplified model.

Diagnosis by radiologists

We conducted a comparative study to evaluate the effectiveness of our model versus clinical experts in terms of BA diagnosis. In this part, four radiologists (three radiologists from our center with more than 10 years, more than 5 years, and more than 1 year of experience in BA diagnosis, and one radiologist from another center without experience in BA diagnosis) independently reviewed all the US images. All these radiologists were provided with associated information typically available in the clinical setting, including demographic data, clinical signs and symptoms, laboratory features, and US images, which were also provided as input data to the AI system. The radiologists were blinded to the patients' identities and final diagnoses. To assess the ability of our model to enhance the performance of the radiologists, we further presented the predictive results of the model to the radiologists, who then made the diagnosis by referring to the predictions of the model.

Visualization

One way to explain a black-box diagnostic model is to visualize its decision-making process by showing its attention distribution for images. However, it is challenging to generate separate attention maps for multiple images using the gradient derivation method. Therefore, we used the single-modal model to display the attention areas of the images. We used the Grad-CAM method [34] to calculate and visualize the attention maps from the output results to the input space.

Statistical analysis

We evaluated the performance of the models and radiologists by calculating AUC, sensitivity, specificity, accuracy, positive predictive value, and negative predictive value. The 95% confidence intervals of sensitivity and specificity were calculated using the "exact" Clopper-Pearson confidence interval. The 95% confidence intervals of AUCs were obtained using Delong's method. When comparing the AUCs, the *P* value was also calculated by the Delong test [35]. A P value < 0.05 indicated a statistically significant difference.

Results

Data and dataset

In the retrospective part, we enrolled 681 BA patients and 898 non-BA patients (458 with other cholestasis diseases and 440 infants without liver disease). To ensure balanced distributions among different categories, we randomly split the dataset into training, validation, and test sets at a ratio of 6:1:3. The training set comprised 408 cases of BA, 274 cases of other cholestasis, and 264 cases without liver disease. The validation set consisted of 68 cases of BA, 46 cases of other cholestasis, and 44 cases without liver disease. The test set contained 205 cases of BA, 138 cases of other cholestasis, and 132 cases without liver disease. The clinical characteristics of the included patients are displayed in Additional file 1: Table S1. Following the retrospective study, the prospective study included 171 cases, comprising 70 cases of BA, 55 cases with other cholestasis diseases, and 46 cases without liver disease.

Predictive outcomes of the multimodal model

On the internal test set, the multimodal model achieved an AUC of 0.9870, a sensitivity of 0.9561, a specificity of 0.9630, an accuracy of 0.9600, a positive predictive value of 0.9515, and a negative predictive value of 0.9665. A positive correlation between diagnostic experience and diagnostic accuracy was observed (Table 1). The multimodal model outperformed all four radiologists (Table 1 rows 2–5, Fig. 1B), including one radiologist with more than 10 years of pediatric US experience, who achieved an AUC of 0.9474, a sensitivity of 0.9512, a specificity of 0.9444, an accuracy of 0.9474, a positive predictive value of 0.9286, and a negative predictive value of 0.9623. The Delong test indicated significant differences in AUCs between the four experts and the model.

Following conventional multimodal processing approaches, data are typically directly projected and fed into a transformer-based encoder. We also explored early data fusion methods (Additional file 1: Supplementary Information S13.1), but the results were unsatisfactory, with an AUC of 0.7431, a sensitivity

Table 1 Performance of the models and experts under various data settings on the internal test dataset

	Data	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy	PPV	NPV	*P* value[1]
Multimodal model	US modality, CI and LTD	0.9870 (0.9777, 0.9958)	0.9561 (0.9183, 0.9797)	0.9630 (0.9329, 0.9821)	0.9600	0.9515	0.9665	-
Expert 1		0.9474 (0.9277, 0.9680)	0.9512 (0.9121, 0.9764)	0.9444 (0.9100, 0.9686)	0.9474	0.9286	0.9623	< 0.001
Expert 2		0.8695 (0.8507, 0.9068)	0.9463 (0.9060, 0.9729)	0.8111 (0.7592, 0.8560)	0.8695	0.7918	0.9522	< 0.001
Expert 3		0.8295 (0.8107, 0.8729)	0.9317 (0.8881, 0.9622)	0.7519 (0.6959, 0.8022)	0.8295	0.7403	0.9355	< 0.001
Expert 4		0.8147 (0.8081, 0.8648)	0.9951 (0.9731, 0.9999)	0.6778 (0.6185, 0.7331)	0.8147	0.7010	0.9946	< 0.001
Unimodal model	Gallbladder	0.9666 (0.9512, 0.9835)	0.8361 (0.7743, 0.8866)	0.9572 (0.9247, 0.9784)	0.9091	0.8945	0.9333	-
	TCS	0.9418 (0.9136, 0.9689)	0.8421 (0.7786, 0.8933)	0.9602 (0.9258, 0.9816)	0.9093	0.8893	0.9412	-
	Liver capsule	0.7977 (0.7500, 0.8452)	0.7468 (0.6705, 0.8133)	0.7421 (0.6738, 0.8027)	0.7442	0.7833	0.7012	-
	Liver parenchyma	0.8389 (0.8019, 0.8730)	0.7363 (0.6697, 0.7958)	0.7566 (0.7005, 0.8068)	0.7479	0.7922	0.6948	-
	CI and LTD	0.9595 (0.9433, 0.9766)	0.9317 (0.8881, 0.9622)	0.8852 (0.8410, 0.9206)	0.9053	0.8604	0.9447	-
Simplified multimodal model	Gallbladder, TCS	0.9590 (0.9457, 0.9778)	0.8439 (0.7868, 0.8907)	0.9630 (0.9329, 0.9821)	0.9116	0.9454	0.8904	< 0.001
	Gallbladder, TCS, and CI	0.9731 (0.9596, 0.9857)	0.9024 (0.8533, 0.9394)	0.9444 (0.9100, 0.9686)	0.9263	0.9250	0.9273	0.0110
	Gallbladder, TCS, CI, and LTD	0.9791 (0.9662, 0.9898)	0.9463 (0.9060, 0.9729)	0.9481 (0.9145, 0.9714)	0.9474	0.9327	0.9588	0.0420
	US modality	0.9535 (0.9398, 0.9760)	0.8390 (0.7814, 0.8865)	0.9407 (0.9055, 0.9658)	0.8968	0.9149	0.8850	< 0.001
	US modality and CI	0.9799 (0.9689, 0.9918)	0.8878 (0.8364, 0.9275)	0.9444 (0.9100, 0.9686)	0.9200	0.9239	0.9173	0.2070
	US modality and LTD	0.9795 (0.9671, 0.9902)	0.9268 (0.8822, 0.9585)	0.9444 (0.9100, 0.9686)	0.9368	0.9268	0.9444	0.0805

Expert 1, Expert 2, and Expert 3 refer to radiologists with more than 10 years, more than 5 years, and more than 1 year of experience in BA diagnosis, respectively. Expert 4 refers to a radiologist without experience in BA diagnosis

TCS triangular cord sign, *CI* clinical information, *LTD* laboratory test data, AUC area under receiver operating characteristic curve, *PPV* positive predictive value, *NPV* negative predictive value

[1] The values were calculated using the Delong test to compare the AUCs

of 0.8018, a specificity of 0.6465, and an accuracy of 0.7444. Moreover, with consideration of the differences between modalities, we used cross-attention mechanisms to fuse the image and numeric data modalities (Additional file 1: Supplementary Information S13.2), this approach achieved an AUC of 0.8323, a sensitivity of 0.9171, a specificity of 0.8963, and an accuracy of 0.9053. The results indicated that both intra- and intermodal differences warranted refined processing, particularly in medical data where instances of the same data type may vary with respect to objects and information density. Consequently, additional work beyond simply partitioning the data based on their modalities may be necessary.

On the external test set, the multimodal model achieved an AUC of 0.9740, a sensitivity of 0.9130, a specificity of 0.9510, an accuracy of 0.9357, a positive predictive value of 0.9515, and a negative predictive value of 0.9665. These results were comparable to the radiologist with over 10 years of experience (AUC = 0.9766), and no statistically significant differences were observed (Table 2, Fig. 2B). We compared the diagnostic results between radiologists and the model and the model exhibited fewer errors in its predictions compared to the radiologists (Table 3).

In total, there were 6 false negative cases and 5 false positive cases. Of these 11 cases, all but 2 were male, with 1 false negative case and 1 false positive case

Table 2 Performance of the models and experts on the external test dataset

	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy	PPV	NPV	*P* value[1]
Multimodal model	0.9740 (0.9490, 0.9959)	0.9420 (0.8582, 0.9840)	0.9510 (0.8893, 0.9839)	0.9600	0.9515	0.9665	-
Expert 1	0.9766 (0.9473, 0.9994)	0.9565 (0.8782, 0.9909)	0.9902 (0.9466, 0.9998)	0.9474	0.9286	0.9623	0.9489
Expert 2	0.9064 (0.8778, 0.9559)	0.9710 (0.8992, 0.9965)	0.8627 (0.7804, 0.9229)	0.8695	0.7918	0.9522	0.0044
Expert 3	0.8070 (0.7926, 0.8839)	0.9317 (0.8881, 0.9622)	0.7519 (0.6959, 0.8022)	0.8295	0.7403	0.9355	< 0.001
Expert 4	0.6667 (0.5612, 0.7018)	0.4493 (0.3292, 0.5738)	0.8137 (0.7245, 0.8840)	0.8147	0.7010	0.9946	< 0.001
Expert 1 with machine aided	0.9766 (0.9516, 0.9998)	0.9710 (0.8992, 0.9965)	0.9804 (0.9310, 0.9976)	0.9474	0.9286	0.9623	0.7839
Expert 2 with machine aided	0.9123 (0.8690, 0.9558)	0.9130 (0.8203, 0.9674)	0.9118 (0.8391, 0.9589)	0.8695	0.7918	0.9522	0.0015
Expert 3 with machine aided	0.8947 (0.8704, 0.9484)	0.9855 (0.9219, 0.9996)	0.8333 (0.7466, 0.8998)	0.8295	0.7403	0.9355	< 0.001
Expert 4 with machine aided	0.9006 (0.8417, 0.9401)	0.8406 (0.7326, 0.9176)	0.9412 (0.8764, 0.9781)	0.8147	0.7010	0.9946	0.0016

Expert 1, Expert 2, and Expert 3 refer to radiologists with more than 10 years, more than 5 years, and more than 1 year of experience in BA diagnosis, respectively. Expert 4 refers to a radiologist without experience in BA diagnosis

AUC area under receiver operating characteristic curve, *PPV* positive predictive value, *NPV* negative predictive value

[1] The values were calculated using the Delong test to compare the AUCs

being female. Among the 6 false negative cases, where the model failed to correctly diagnose BA, 3 cases presented with evidently small or difficult-to-identify gallbladders, while 3 cases exhibited slightly irregular gallbladders—a known challenge in radiological assessments. This suggests that the model's performance may be influenced by the quality of the imaging data or by atypical presentations of the condition. For the 5 false positive cases, where the model incorrectly diagnosed BA in patients without the condition, 3 patients were under 1 month of age. Additionally, 2 cases of choledochal cysts were misclassified as cystic BA. In 4 of the 5 false positive cases (including the two choledochal cysts), the GGT levels were significantly elevated, ranging from 528 to 990 U/L. This indicates that the model may be more prone to false positives in very young infants and patients with choledochal cysts, as well as those with elevated GGT levels.

Diagnostic performance with the assistance of a multimodal model

On the internal test set, all radiologists demonstrated an improvement in diagnostic AUC when assisted by the model, with greater enhancements observed among those with less experience (Fig. 1D). In terms of sensitivity, the performance of all radiologists with over 1 year of experience improved and surpassed that of the model, while the performance of the radiologist without diagnostic experience declined to the same level as that of the model. A high sensitivity is essential for effective screening of this relatively rare but highly lethal disease because a highly sensitive test reduces the risk of false negative results. All radiologists exhibited improved specificity, except for the radiologist with more than 10 years of experience. Similar trends were observed for accuracy and PPV, where radiologists with less experience showed greater improvements. Similar to sensitivity, the NPV of

(See figure on next page.)

Fig. 2 Prospective external cohort characteristics and results. **A** Heat map of clinical and laboratory test characteristics for the prospective external cohort of 171 cases. The color intensity of each element correlates to its values. **B** ROC curve analysis of the multimodal models and experts on the external test dataset. **C** Distribution of each group within the external test cohort. **D** Changes in diagnostic outcomes of radiologists when assisted by the multimodal model on the external test dataset. Illustration of changes in AUC, sensitivity, specificity, accuracy, PPV, and NPV of the four experts before and after the assistance of the model. Circles represent the diagnostic outcomes of the radiologists when the diagnosis was established independently, squares represent the diagnostic outcomes when aided by the model, and stars denote the performance of the model. Expert 1, Expert 2, and Expert 3 refer to radiologists with more than 10 years, more than 5 years, and more than 1 year of experience in pediatric US, respectively. Expert 4 refers to a radiologist without experience in pediatric US

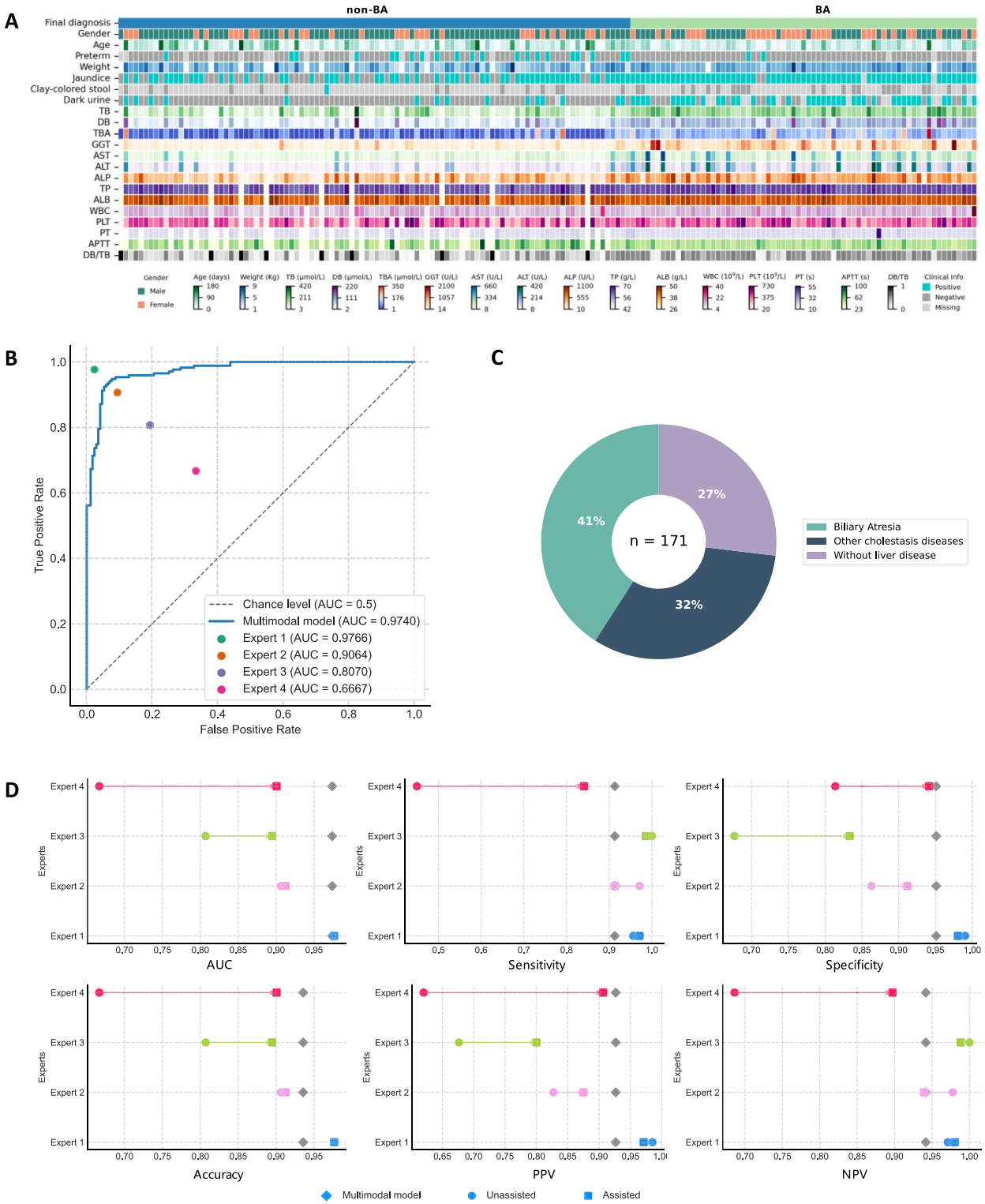

Fig. 2 (See legend on previous page.)

the least experienced radiologist declined to the level of the model, while other radiologists demonstrated higher NPV and outperformed the model.

On the external test set, the same trend was observed, with less experienced radiologists showing more notable improvements in AUC with the assistance of the model

Table 3 Comparison of diagnostic results of the model and experts

Test cohort	Experts	Doctor incorrect, model correct	Doctor correct, model incorrect
Internal test	Expert 1	18	12
	Expert 2	50	7
	Expert 3	72	10
	Expert 4	78	9
External test	Expert 1	1	8
	Expert 2	13	8
	Expert 3	31	9
	Expert 4	52	6

Expert 1, Expert 2, and Expert 3 refer to radiologists with more than 10 years, more than 5 years, and more than 1 year of experience in BA diagnosis, respectively. Expert 4 refers to a radiologist without experience in BA diagnosis

(Fig. 2D). Specifically, the AUC of the radiologist without experience in BA diagnosis increased from 0.6667 to 0.9006, sensitivity improved from 0.4493 to 0.8406, and specificity increased from 0.8137 to 0.9412. However, the radiologist with over 10 years of experience did not show any improvement in AUC after AI assistance. Overall, the results suggest that experience plays a crucial role in the performance of radiologists, while our model can significantly enhance the diagnostic accuracy of less experienced radiologists.

Predictive outcomes of the simplified multimodal model

Using easily obtained data, even avoiding pain from blood sampling, our simplified model achieved an AUC of 0.9799 (Table 1 row 15), which was comparable to the performance of the multimodal model mentioned above. Moreover, the combination of the gallbladder and the TCS yielded an AUC of 0.9590 (Table 1 row 11). By incorporating clinical information and laboratory test features, the model further improved its AUC to 0.9791 (Table 1 row 13). All the ROC curves are shown in Fig. 3A.

Predictive outcomes of the unimodal model

The results were consistent with clinical experience. The gallbladder achieved the highest performance, with

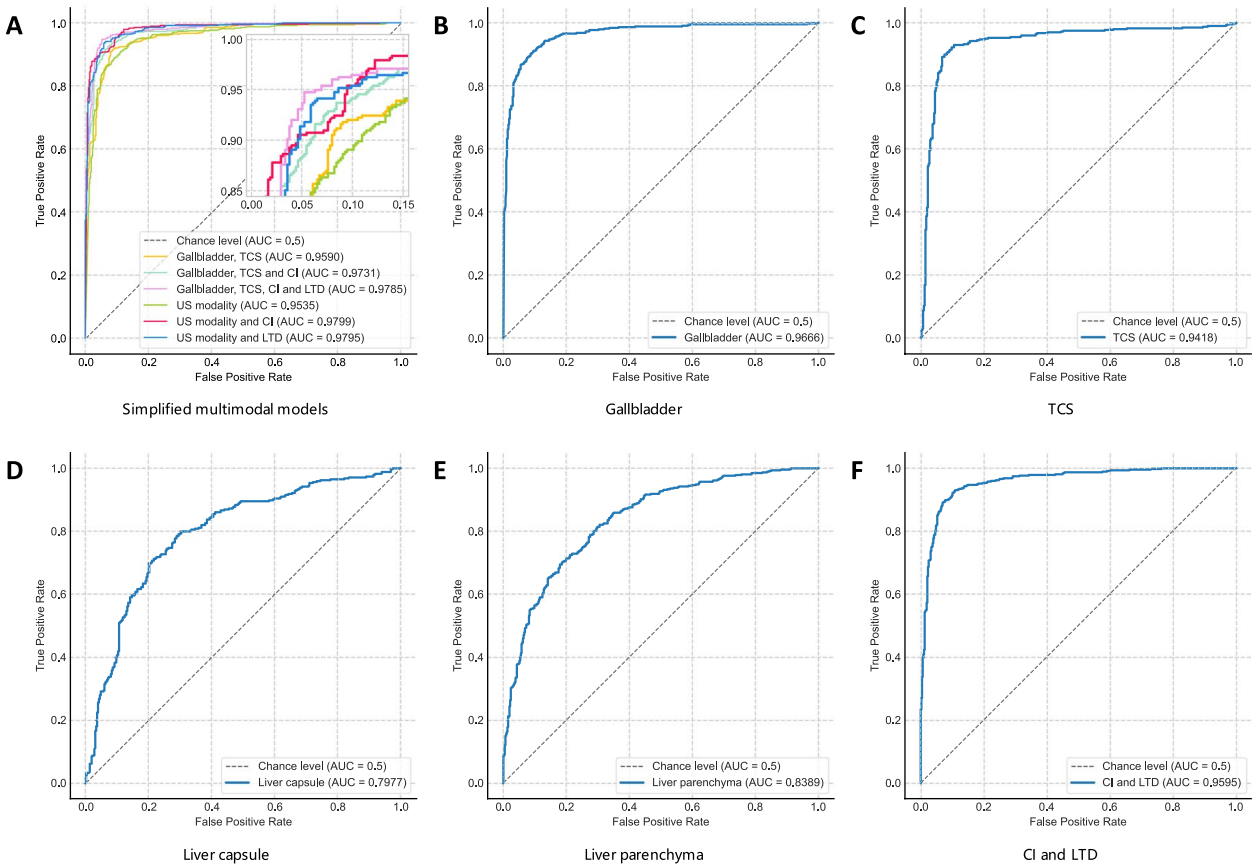

Fig. 3 Prospective external cohort characteristics and results. **A** ROC curves of simplified multimodal models under different data settings. **B**–**E** ROC curves of unimodal models based on the gallbladder, triangular cord sign (TCS), liver capsule, and liver parenchyma US images. **F** ROC curve of unimodal based on the combination of clinical information (CI) and laboratory test data (LTD)

an AUC of 0.9666, a sensitivity of 0.9572, a specificity of 0.8415, an accuracy of 0.9091, a positive predictive value of 0.8945, and a negative predictive value of 0.9333 (Table 1 row 6; Fig. 3B). The TCS achieved an AUC of 0.9418, a sensitivity of 0.9602, a specificity of 0.8421, an accuracy of 0.9093, a positive predictive value of 0.8893, and a negative predictive value of 0.9412 (Table 1 row 7; Fig. 3C). The AUCs of the liver capsule and liver parenchyma were 0.7977 and 0.8389, respectively (Table 1 rows 8–9; Fig. 3D, E), indicating moderate classification performance. The numeric data yielded unexpectedly high diagnostic accuracy with an AUC of 0.9595, a sensitivity of 0.9317, a specificity of 0.8852, an accuracy of 0.9053, a positive predictive value of 0.8604, and a negative predictive value of 0.9447 (Table 1 row 10; Fig. 3F), showcasing the novel capabilities of machine learning in deciphering the complex relationships among numeric variables.

Visualization results

As depicted in Fig. 4, the red areas represent the most important parts that significantly contributed to the prediction, while the blue areas represent less influential regions. These findings generally align with the expectations of clinicians. For gallbladder images, the model primarily focused on the contour of the gallbladder and its surrounding areas. Similarly, for TCS images, the model directed its attention towards TCS regions. When analyzing liver capsule images, the model concentrated on specific regions within the location of the liver capsule. However, the model did not exhibit any specific regions of emphasis on liver parenchyma images. Instead, it appeared to extract global information from the images, possibly due to the dispersion of crucial information throughout the liver parenchyma.

Discussion

Although progress in diagnostic strategies for rare diseases incorporating AI and its applications lags behind that of other medical disciplines, accelerating automated diagnosis for rare diseases holds great significance in providing high-quality, safe, and efficient healthcare, especially in primary care settings [36]. This study represents the pioneering effort to implement a multimodal intelligent screening tool that integrates demographic, clinical, laboratory, and US features using the largest available dataset of patients with BA. Clinical experience has shown that US images should be interpreted in conjunction with a patient's clinical condition, rather than in isolation. Therefore, we employed a self-masked attention mechanism to integrate both intra- and intermodal information. This approach addresses the issue of missing data and enhances the high-precision diagnosis of BA through the comprehensive analysis of multimodal data. As expected, the proposed multimodal deep learning model outperformed models that relied solely on US parameters or various combinations of these parameters, indicating that each piece of scattered information contributes to improving the prediction.

Constructing a high-precision and scalable multimodal model posed several challenges. The first challenge revolved around designing a scalable deep learning model capable of effectively processing multimodal inputs. Unlike previous multimodal approaches, which relied on large amounts of data to achieve excellent generalization performance [37, 38], we prioritized the effective utilization of multimodal data and the modularization of key steps, enabling proper adjustments based on the specific data format and task type. The second challenge centered on developing efficient multimodal fusion techniques and addressing the issue of missing modalities. Traditional multimodal models relied heavily on global features and intermodal information for diagnostic predictions, while disregarding the specific information contained within individual modalities [23]. Our proposed approach incorporated a self-masked attention mechanism for image data fusion to extract intra-modal information and an MLP for inter-modal fusion. Unlike the self-attention mechanism, the self-masked attention mechanism reduces the attention paid towards itself during the computation process, allowing for a greater focus on information derived from other parts. This characteristic is particularly advantageous in scenarios involving limited data, as it allocates more attention to external elements. Drawing on previous methods that used fusion feature space interpolation for completion [39], we introduced prior knowledge and proposed a novel methodology for fusion and supplementation of information across different modalities to address missing values.

In addition to highly precise predictions, our model has the advantage of providing clinical decision support. With the assistance of AI, all radiologists showed improvements in AUC, and these improvements were negatively correlated with their diagnostic experience. On the external test set, with the assistance of the model, the AUC of the radiologist without diagnostic experience increased from 0.6667 to 0.9006, sensitivity improved from 0.4493 to 0.8406, and specificity increased from 0.8137 to 0.9412. It seems that experience is an important factor in the performance of radiologists, which is not surprising given the complexity of medical imaging interpretation. However, our model has demonstrated its ability to greatly enhance the diagnostic proficiency of less experienced radiologists. This is a significant advancement with important implications for BA diagnosis, especially in resource-limited settings. The model may provide an accurate and reliable reference for

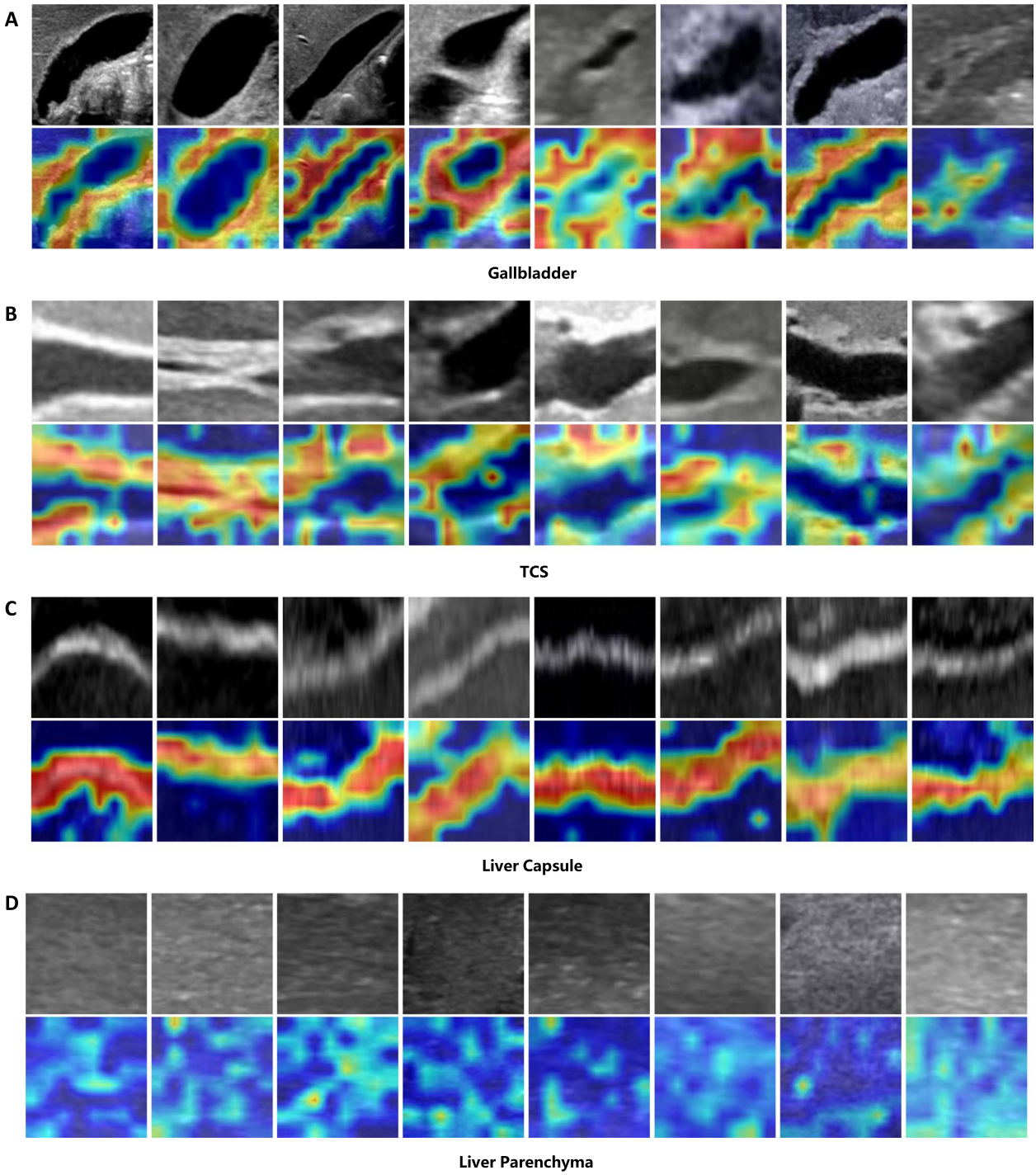

Fig. 4 Visualization of the class activation maps generated by the unimodal model across four types of US images. The red regions in the heatmaps highlight the areas that attract the most attention from the model. Each subfigure consists of two rows: the upper row presents the original images fed into the model, while the lower row showcases the corresponding heatmap results

pediatricians, most of whom only encounter BA cases a handful of times throughout their career. Moreover, it offers an effective balance between cost and benefit by preventing the cycle of misdiagnosis and correction that often occurs when an inexperienced doctor makes an error and requires additional consultations. Patients could receive accurate diagnoses and treatment recommendations without the need to travel long distances or

bear the high costs associated with multiple consultations and specialist referrals. Misdiagnoses can sometimes trigger a cascade of unnecessary tests and treatments as doctors attempt to rule out potential conditions. An AI system that delivers a reliable diagnosis from the outset could help minimize these unnecessary investigations, ensuring a more focused and efficient use of healthcare resources.

In this study, we explored innovative approaches for BA prediction based on certain US features. While there have been successful applications of US imaging of liver capsule and parenchyma for intelligent liver fibrosis prediction in adult studies [40–42], it remains unclear whether this relationship in BA can be recognized by deep learning. Therefore, we extracted features of the liver capsule and parenchyma from US images for intelligent classification of BA. The results indicated that the liver capsule and parenchyma achieved moderate classification performance, with AUC values of 0.7977 and 0.8389, respectively. However, it should be noted that visual inspections of changes in the liver capsule and echotexture of the parenchyma are considered to be inaccurate and unreliable in identifying BA. Another unanticipated finding was the remarkable performance achieved by clinical information and laboratory test results (AUC = 0.9595). These findings imply that artificial intelligence demonstrates exceptional proficiency in detecting extremely subtle structural changes and managing intricate numerical tasks, which pose challenges for human interpretation.

The present study has several limitations that need to be acknowledged. Firstly, advanced techniques such as shear-wave elastography and new biomarkers like MMP-7, which have been reported to be effective in diagnosing BA, were not included in the model due to limited availability for reliable assessment. Secondly, deep learning is a data-driven analytical technique that lacks transparent inference explanations. We employed visualization techniques and comprehensive ablation experiments to address this limitation. However, these methods infer explanations based on the results rather than the internal reasoning process of the model. In the future, we can incorporate case-based methods to leverage the self-explained capabilities of machine learning [43, 44]. Thirdly, the validation of this model was conducted at a single institution, a major referral center for BA treatment in northern China, and as such, the patient population reflects the characteristics of BA patients in this region. However, to ensure the broader applicability of our findings, a multicenter study involving diverse healthcare settings and patient populations is warranted.

With heightened awareness of the challenges faced by existing BA screening systems and the need for limited invasiveness in neonates, it is now an opportune time to shift attention from traditional diagnostic patterns based on experience to intelligent diagnostic technology. The latter is expected to ensure that patients with rare diseases in different institutions have access to high-quality care equally. Our study is the first attempt to apply deep learning-based algorithms to highly complex scenarios for BA diagnosis using multimodal data, without requiring sophisticated equipment or specialized skills. It has the potential to assist in selecting neonates with suspected BA for intraoperative cholangiography in a timely manner. This would help streamline diagnostic workflows and minimize the need for costly interventions, ensuring that only the most necessary tests and treatments are administered, thereby optimizing resource utilization. Moreover, by enhancing diagnostic precision, our model could reduce readmissions and complications, which represent significant financial burdens on healthcare systems. Additionally, the integration of our model could increase labor efficiency by automating routine diagnostic tasks, allowing clinicians to focus on more complex cases.

Conclusions

In this study, we developed and validated an AI-based diagnostic system for the early detection of BA using a comprehensive multimodal dataset. By integrating ultrasound images, clinical data, and laboratory results, our multimodal deep learning models achieved high accuracy, outperforming human experts in retrospective evaluations and demonstrating robust performance in prospective validation. Additionally, the significant improvement in diagnostic accuracy for less experienced radiologists when assisted by our AI system further underscores the value of this technology as a supportive tool in medical practice. Moreover, our approach addresses common challenges in rare disease diagnostics, providing a framework that could be adapted to other conditions with similar diagnostic complexities.

Abbreviations

BA	Biliary atresia
US	Ultrasound
AI	Artificial intelligence
AUC	Area under the curve
PPV	Positive predictive value
NPV	Negative predictive value
GGT	Gamma-glutamyltransferase
TCS	Triangular cord sign
TB	Total bilirubin
DB	Direct bilirubin
TBA	Total bile acid
TP	Total protein
ALB	Albumin
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
GGT	Gamma-glutamyl transpeptidase
ALP	Alkaline phosphatase
PLTs	Platelets

WBCs	White blood cells
PT	Prothrombin time
APTT	Activated partial thromboplastin time
MLP	Multi-layer perceptron
ROI	Region of interest

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12916-025-03962-x.

Additional file 1: Supplementary Information S1-S14, Tables S1-S9, Figures S1-S11. Supplementary Information S1. Data statistics. Supplementary Information S2. Statistical analysis. Supplementary Information S3. Multimodal model design. Supplementary Information S4. Swin Transformer Encoder. Supplementary Information S5. Cross-Validation. Supplementary Information S6. Unimodal model with different parameters. Supplementary Information S7. Ablation experiment of missing value completion. Supplementary Information S8. Ablation experiment of self-masked attention. Supplementary Information S9. Ablation experiment of multi-loss. Supplementary Information S10. Ablation experiment of pretraining. Supplementary Information S11. Ablation experiment of alternative vision backbones. Supplementary Information S12. Simplified model training process. Supplementary Information S13. Comparative experiment with the EDLM model. Supplementary Information S14. Alternative multimodal model design. Table S1. Statistics of the clinical information and laboratory test data in our collected dataset. Table S2. Results of 5-fold and 3-fold cross-validations performed on our collected dataset. Table S3. Performance comparison between the parameters of the multimodal model and those of the retrained unimodal models on single-modality data. Table S4. Ablation experiment for evaluating the missing value completion strategy of our proposed model. Table S5. Ablation experiment for comparing the impacts of different attention mechanisms on the model. Table S6. Ablation experiment for evaluating the multi-loss training method in our proposed model. Table S7. Ablation experiment for evaluating the performance of pretrained parameters on the model. Table S8. Ablation experiment for evaluating the performance of models using different backbone visual encoders. Table S9. Performance comparison between our visual encoding model and the EDLM model on the EDLM dataset. Fig. S1. The architectures of the Swin transformer V2 and Swin transformer V2 block. Fig. S2. Illustration of the multi-loss computation. Fig. S3. The ROC curves produced in the stratified k-fold experiments for evaluating the influence of the dataset splitting strategy. Fig. S4. ROC curves illustrating the ablation experiment settings for evaluating the missing value completion strategy with the prior knowledge copying method. Fig. S5. ROC curves depicting the ablation experiment settings for evaluating the self-masked attention method. Fig. S6. ROC curves of the ablation experiment settings for evaluating the multi-loss method. Fig. S7. ROC curves of the ablation experiment settings for evaluating the performance of the pretrained parameters. Fig. S8. ROC curves of the ablation experiment settings for evaluating the performance of models using different backbone visual encoders. Fig. S9. Overview of the simplified model training process. Fig. S10. Overview of the proposed alternative multimodal deep model with the early fusion strategy. Fig. S11. Overview of the proposed alternative multimodal deep model with latent feature fusion.

Additional file 2. CLAIM Guideline Checklist.

Acknowledgements
The authors thank Jie Zhou and Jianqiu Huang for their valuable discussions regarding the imaging analysis.

Authors' contributions
Designing the study: YCY, YM. Methodology: YCY, YXD, LYJ, YM. Algorithms: YCY, YXD, LYJ. Validation: YCY, YXD, LYJ. Result analysis: YCY, YM. Data curation: YM, YDW, LYL, ZJZ, ZLJ, ZMQ, MY. Writing—Original Draft & Revise: YCY, YM. Writing—Figure & Table: YCY, YM, LYJ. Revise: YCY, YM, YXD, LYJ, BYL, YQZ. Visualization: YXD, YCY. Supervision: CT, ZW. Funding acquisition: CT, YM. All authors read and approved the final manuscript.

Funding
This study is partially supported by the National Natural Science Foundation of China (62176016, 72274127, 82202197), National Key R&D Program of China (No. 2021YFB2104800), Guizhou Province Science and Technology Project: Research and Demonstration of Sci. & Tech Big Data Mining Technology Based on Knowledge Graph (supported by Qiankehe (2021) General 382), and Capital Health Development Research Project (2022–2-2013).

Data availability
Due to privacy considerations and in compliance with applicable data protection regulations, the data used in this study cannot be openly shared or made publicly available. To mitigate these concerns, access to the data is restricted, and interested researchers from reputable institutions may request access by contacting the corresponding author via email. The data will be made available to qualified researchers upon completing a data sharing agreement and adhering to relevant privacy and ethical guidelines. The source code of this study is available online on https://github.com/hiyyc/Multimodal-AI-for-BiliaryAtresia-Diagnosis.

Declarations

Ethics approval and consent to participate
The study received approval from the Institutional Review Board of Capital Institute of Pediatrics. For the prospective part, written informed consent was obtained from the guardians of each participant. For the retrospective part, the requirement for informed consent was waived.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Computer Science and Engineering, Beihang University, Beijing, China. [2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China. [3]Department of Ultrasound, Capital Institute of Pediatrics, Beijing, China. [4]Department of General Surgery, Capital Institute of Pediatrics, Beijing, China.

Received: 28 September 2024 Accepted: 20 February 2025
Published online: 27 February 2025

References

1. Hartley JL, Davenport M, Kelly DA. Biliary atresia Lancet Lond Engl. 2009;374:1704–13.
2. Govindarajan KK. Biliary atresia: Where do we stand now? World J Hepatol. 2016;8:1593–601.
3. Chung PHY, Zheng S, Tam PKH. Biliary atresia: East versus west. Semin Pediatr Surg. 2020;29: 150950.
4. McKiernan PJ, Baker AJ, Kelly DA. The frequency and outcome of biliary atresia in the UK and Ireland. Lancet Lond Engl. 2000;355:25–9.
5. Cavallo L, Kovar EM, Aqul A, McLoughlin L, Mittal NK, Rodriguez-Baez N, et al. The Epidemiology of Biliary Atresia: Exploring the Role of Developmental Factors on Birth Prevalence. J Pediatr. 2022;246:89-94.e2.
6. Tiao M-M, Tsai S-S, Kuo H-W, Chen C-L, Yang C-Y. Epidemiological features of biliary atresia in Taiwan, a national study 1996–2003. J Gastroenterol Hepatol. 2008;23:62–6.
7. Wada H, Muraji T, Yokoi A, Okamoto T, Sato S, Takamizawa S, et al. Insignificant seasonal and geographical variation in incidence of biliary atresia in Japan: a regional survey of over 20 years. J Pediatr Surg. 2007;42:2090–2.
8. Kasai M, Kimura S, Asakura Y, Suzuki H, Taira Y, Ohashi E. Surgical treatment of biliary atresia. J Pediatr Surg. 1968;3:665–75.
9. Hopkins PC, Yazigi N, Nylund CM. Incidence of Biliary Atresia and Timing of Hepatoportoenterostomy in the United States. J Pediatr. 2017;187:253–7.
10. Parolini F, Boroni G, Milianti S, Tonegatti L, Armellini A, Garcia Magne M, et al. Biliary atresia: 20–40-year follow-up with native liver in an Italian centre. J Pediatr Surg. 2019;54:1440–4.

11. Schreiber RA, Harpavat S, Hulscher JBF, Wildhaber BE. Biliary Atresia in 2021: Epidemiology, Screening and Public Policy. J Clin Med. 2022;11:999.
12. Zhou L-Y, Wang W, Shan Q, Liu B, Zheng Y, Xu Z, et al. Optimizing the US Diagnosis of Biliary Atresia with a Modified Triangular Cord Thickness and Gallbladder Classification. Radiology. 2015;277:181–91.
13. Sandberg JK, Sun Y, Ju Z, Liu S, Jiang J, Koci M, et al. Ultrasound shear wave elastography: does it add value to gray-scale ultrasound imaging in differentiating biliary atresia from other causes of neonatal jaundice? Pediatr Radiol. 2021;51:1654–66.
14. Humphrey TM, Stringer MD. Biliary atresia: US diagnosis. Radiology. 2007;244:845–51.
15. El-Guindi MA-S, Sira MM, Konsowa HA-S, El-Abd OL, Salem TA-H. Value of hepatic subcapsular flow by color Doppler ultrasonography in the diagnosis of biliary atresia. J Gastroenterol Hepatol. 2013;28:867–72.
16. Napolitano M, Franchi-Abella S, Damasio MB, Augdal TA, Avni FE, Bruno C, et al. Practical approach to imaging diagnosis of biliary atresia, Part 1: prenatal ultrasound and magnetic resonance imaging, and postnatal ultrasound. Pediatr Radiol. 2021;51:314–31.
17. Wang KS. Section on Surgery, Committee on Fetus and Newborn, Childhood Liver Disease Research Network. Newborn Screening for Biliary Atresia Pediatrics. 2015;136:e1663-1669.
18. Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. Lancet Digit Health. 2020;2:e486–8.
19. Alnaggar OAMF, Jagadale BN, Saif MAN, et al. Efficient artificial intelligence approaches for medical image processing in healthcare: comprehensive review, taxonomy, and analysis. Artif Intell Rev. 2024;57:221. https://doi.org/10.1007/s10462-024-10814-2.
20. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
21. Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw. 2015;61:85–117.
22. Placido D, Yuan B, Hjaltelin JX, Zheng C, Haue AD, Chmura PJ, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. Nat Med. 2023;29:1113–22.
23. Steyaert S, Pizurica M, Nagaraj D, Khandelwal P, Hernandez-Boussard T, Gentles AJ, et al. Multimodal data fusion for cancer biomarker discovery with deep learning. Nat Mach Intell. 2023;5:351–62.
24. Witowski J, Heacock L, Reig B, Kang SK, Lewin A, Pysarenko K, et al. Improving breast cancer diagnostics with deep learning for MRI. Sci Transl Med. 2022;14:eabo4802.
25. Soenksen LR, Kassis T, Conover ST, Marti-Fuster B, Birkenfeld JS, Tucker-Schwartz J, et al. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. Sci Transl Med. 2021;13:eabb3652.
26. Zhou W, Yang Y, Yu C, Liu J, Duan X, Weng Z, et al. Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images. Nat Commun. 2021;12:1259.
27. Zhou W, Ye Z, Huang G, Zhang X, Xu M, Liu B, et al. Interpretable artificial intelligence-based app assists inexperienced radiologists in diagnosing biliary atresia from sonographic gallbladder images. BMC Med. 2024;22:29.
28. Hsu F-R, Dai S-T, Chou C-M, Huang S-Y. The application of artificial intelligence to support biliary atresia screening by ultrasound images: A study based on deep learning models. PLoS ONE. 2022;17: e0276278.
29. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2:719–31.
30. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18:500–10.
31. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med. 2019;25:433–8.
32. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28:31–8.
33. Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, et al. Swin Transformer V2: Scaling Up Capacity and Resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. p. 12009–19.
34. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE; 2017. p. 618–26.
35. Sun X, Xu W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. IEEE Signal Process Lett. 2014;21:1389–93.
36. Banerjee J, Taroni JN, Allaway RJ, Prasad DV, Guinney J, Greene C. Machine learning in rare disease. Nat Methods. 2023;20:803–14.
37. Kim W, Son B, Kim I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In: Proceedings of the 38th International Conference on Machine Learning. PMLR; 2021. p. 5583–94.
38. Nagrani A, Yang S, Arnab A, Jansen A, Schmid C, Sun C. Attention Bottlenecks for Multimodal Fusion. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2021. p. 14200–13.
39. Lee C, Schaar M van der. A Variational Information Bottleneck Approach to Multi-Omics Data Integration. In: Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. PMLR; 2021. p. 1513–21.
40. Xue L-Y, Jiang Z-Y, Fu T-T, Wang Q-M, Zhu Y-L, Dai M, et al. Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis. Eur Radiol. 2020;30:2973–83.
41. Lee JH, Joo I, Kang TW, Paik YH, Sinn DH, Ha SY, et al. Deep learning with ultrasonography: automated classification of liver fibrosis using a deep convolutional neural network. Eur Radiol. 2020;30:1264–73.
42. Liu X, Song JL, Zhao J, Chen YQ, Zhang JQ. Extracting and describing liver capsule contour in high-frequency ultrasound image for early HBV cirrhosis diagnosis. In: IEEE international conference on multimedia and expo, ICME 2016, seattle, WA, USA, july 11–15, 2016. IEEE Computer Society; 2016. p. 1–6.
43. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell. 2019;1:206–15.
44. Chen H, Gomez C, Huang C-M, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med. 2022;5:156.

Publisher's Note