Multimodal Model for the Diagnosis of Biliary Atresia Based on Sonographic Images and Clinical Parameters

Supplementary Note 1: Ultrasound image and video acquisition criteria

Both gallbladder and triangular cord ultrasound (US) images were required to meet the following acquisition criteria: Imaging Technique: (1) Acquired using high-frequency ultrasound without artifacts, obstructions, or significant noise. (2) Target Visualization: Clear and complete depiction of the target structure (gallbladder or triangular cord). (3) Image Cleanliness: No obstructive interference (e.g., measurement calipers, annotations, arrows, etc.) within the target region. (4) Gallbladder-Specific: Maximum long-axis view of the gallbladder must be captured. (5) Triangular Cord-Specific: Clear visibility of the right portal vein lumen. For video acquisitions: (1) Gallbladder Video: Need include the longest complete sectional view of the gallbladder. (2) Triangular Cord Video: Need capture the thickest sectional view. All videos were recorded for 3-10 seconds to ensure adequate dynamic assessment.

The shear wave elastography (SWE) images were acquired according to the following standardized criteria: (1) Region of interest (ROI) Placement: A rectangular electronic ROI (minimum 2.0cm×2.5cm) was positioned 1.0-3.0cm from the liver capsular surface for SWE measurements. (2) Anatomical Avoidance: Special attention was paid to avoid any vessels, biliary tracts, bones, or artifacts from lung gas. (3) Color Fill: >90% of the ROI area displayed valid color-coded elasticity data. (4) Clean Acquisition: No interference objects were positioned in the ROI box.

Supplementary Note 2: Preprocessing of triangular cord US images

In this study, two pre-processing strategies were adopted for the triangular cord US images. Firstly, a simple edge detection method was used to identify and extract the US imaging area, so that to remove the irrelevant and redundant information outside the US imaging area (such as black background, text information, etc.). The processed image was reviewed by a junior radiologist, and the segmentation error image would be manually corrected. In the second strategy, all images were manually drawn from the original image by a junior radiologist using the software Image J (version 1.52a) to draw a rectangular bounding box that contains the entire triangular cord in the smallest range. The cropped image of the area where the rectangular box

is located was used as the training image. The sizes of the above two processed images were uniformly adjusted to 224×224 pixels. The preliminary experimental results showed that the model trained by the image containing the entire US imaging area obtained by the first strategy was better than that of the rectangular area manually outlined by the second strategy (AUC 0.896 vs. 0.821). Therefore, in the subsequent multimodal fusion modeling, the images obtained by the first strategy were used for model training and testing.

Supplementary Note 3: Preprocessing of clinical information

To address missing attribute data for a minority of infants, we imputed the missing values by calculating the respective means for numerical data and the mode for categorical data. And then the indicators initially enrolled in this study were used to build a diagnostic model, and the SHapley Additive exPlanations (SHAP) value was calculated to estimate the contribution of each indicator in the inference stage. The SHAP values quantified each feature's marginal contribution to the model's output, offering both global and local insights into its behavior. The process involved adjusting the model's predictions by isolating the effect of each feature, then calculating contributions based on these marginal effects and the frequency of each feature in the dataset. These individual contributions were then summed to produce the SHAP value for each sample, providing a comprehensive explanation of the model's decision-making process.

Supplementary Note 4: Model architecture details

The modeling process was structured to comprehensively integrate multimodal data through a systematic architecture, which was mainly divided into three parts: (1) image modality modeling, (2) image and clinical information alignment, and (3) ensemble training and inference strategy. The proposed framework is optimized using standard categorical cross-entropy loss. In the first part, two single-modality imaging models (Gallbladder model and Triangular cord model) were developed using ResNet-101 pretrained on ImageNet-1K as image encoder. A single bottleneck module consisted of three sequentially connected convolutional layers: a 1×1 convolutional layer for channel dimension reduction, followed by a 3×3 convolutional layer with stride=1 and padding=1 for spatial feature extraction, and a final 1×1 convolutional layer for channel dimension restoration. Each convolutional layer was immediately followed by batch normalization and ReLU activation. The skip connection

pathway used identity mapping to directly add input features to bottleneck outputs, thereby maintaining gradient flow and stabilize training.

In the second part, the study adopted different strategies for image-image and image-clinical data fusion. These methods aim to effectively integrate information from different modalities to improve the overall performance of the model. For image-image fusion, the process focuses on combining gallbladder image features and triangulated image features. Each image feature is processed through a ResNet-101 encoder, which generates a 2,048-dimensional feature vector. To create a joint representation, these vectors are concatenated along the channel dimension to produce a unified 4,096-dimensional vector that captures the combined features of the two image types. This approach ensures that the rich spatial and semantic information contained in each image can be preserved and seamlessly integrated. On the other hand, in order to address the difficulty of fusing clinical data with image features due to the large modality gap and dimensionality difference, the study introduced a novel multimodal fusion module to align clinical information with image features and realize the interaction between clinical data and multi-stage image features. Specifically, the input clinical information is first processed using a multi-layer perceptron (MLP). Each MLP consists of two linear layers with ReLU activation applied in the middle. The purpose of this architecture is to map clinical data into a feature space that matches the dimension of the image features extracted by the ResNet encoder. After MLP processing, the mapped clinical features are further transformed into channel attention weights through a sigmoid gate. These channel weights are element-wise multiplied with the intermediate features and the final image features after each residual block. This fusion module will help learn the complementary information between different modalities while mitigating modality noise. Finally, the joint features are processed through a softmax classifier to obtain the final diagnosis of the image (Supplementary Fig. 1).

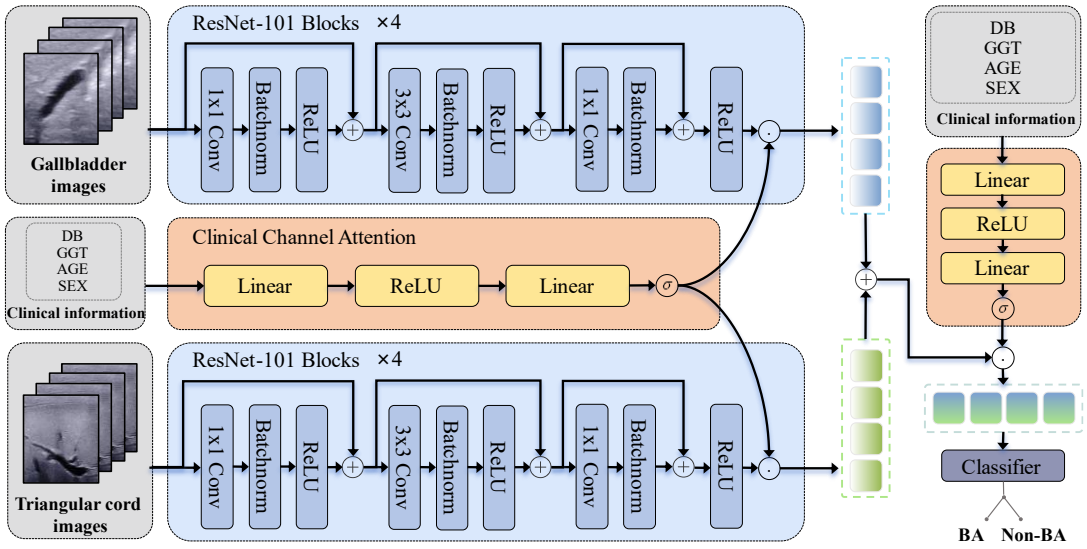
Supplementary Note 5: Segmentation Model Development for US videos

This study evaluated the impact of frame selection ratios, determined by a segmentation model, on the diagnostic performance of an AI model for gallbladder and triangular cord ultrasound video analysis. The diagnostic model's performance was compared across different selection ratios (1%, 5%, and 10%). The model performed optimally when 5% of the video

frames were selected by the segmentation model for diagnosis, balancing both sensitivity (86.3%) and specificity (81.8%) effectively (Supplementary Table 2). This indicated that an intermediate frame selection strategy improves diagnostic reliability compared to very strict (1%) or lenient (10%) selection criteria. This study provided a critical frame optimization strategy for AI-assisted ultrasound diagnosis, improving the reliability of automated diagnostic systems.

Supplementary Figure

Supplementary Figure 1 The detailed architecture of the multi-modal model.



Each image modality would go through ResNet-101 encoder and interact with clinical channel attention features. In the backend of model, two image modality features are concatenated and interact with clinical channel attention features. Finally, the fused multi-modal features would go through the softmax classifier and obtain diagnostic results.

Supplementary Tables

Supplementary Table 1. The number of infants with both gallbladder and triangular cord US images in each cohort and the corresponding number of images and videos.

Type of data	Ultrasound features	Training cohort		Internal test cohort		External test cohort	
		BA (n=194)	Non-BA (n=190)	BA (n=48)	Non-BA (n=40)	BA (n=83)	Non-BA (n=73)
Images	Gallbladder	994	955	245	257	332	252
	Triangular cord	1303	1006	114	106	445	312
Videos	Gallbladder	-	-	-	-	80	66
	Triangular cord	-	-	-	-	80	66

Note: BA, biliary atresia. Non-BA, non-biliary atresia.

Supplementary Table 2. The comparison of model performance on different ratio of selection from gallbladder and triangular cord videos.

Ratio*	AUROC	AUPR	Accuracy (%)	Sensitivity (%)	Specificity (%)
Top 1%	0.921 (0.868, 0.930)	0.912 (0.879, 0.942)	81.5 (74.2, 87.4)	85.0 (75.3, 92.0)	77.2 (65.3, 86.7)
Top 5%	0.930 (0.876, 0.966)	0.945 (0.923, 0.968)	84.2 (77.3, 89.7)	86.3 (76.7, 92.9)	81.8 (70.4, 90.2)
Top 10%	0.905 (0.864, 0.956)	0.912 (0.896, 0.964)	80.1 (72.7, 86.3)	84.2 (75.8, 90.4)	73.6 (63.2, 82.3)

Note: AUROC, area under receiver operating characteristic curve; AUPR, area under Precision-Recall curve.

* The proportion of video frames selected by the segmentation model that are used for diagnosis.