# Detection of Ventricular Septal Defect in Pediatric Cardiac Ultrasound Videos Using Parasternal View and Faster R-CNN

Muhammad Nasrudin[1], Shindi Shella May Wara[1], Amri Muhaimin[1], Nur Indah Nirmalasari[2] & Mega Rizkya Arfiana[3]

*[1]Department of Data Science, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur*
*[2]Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember*
*[3]Department of Children's Health Sciences, Faculty of Medicine, Airlangga University*
**nasrudin.fasilkom@upnjatim.ac.id*

## ABSTRACT

Congenital heart disease (CHD), particularly ventricular septal defect (VSD), remains a major contributor to pediatric morbidity, while echocardiographic diagnosis is highly dependent on operator expertise and image quality. This study examines the feasibility of an object-detection-based intelligent imaging framework for localizing VSD in pediatric cardiac ultrasound videos acquired from the parasternal long-axis view. Rather than proposing a novel detection algorithm, this work adopts a system-oriented approach by evaluating the Faster R-CNN framework under practical clinical constraints, including limited annotated data and heterogeneous ultrasound characteristics. Three convolutional neural network backbones such as ResNet50, ResNet101, and Inception-ResNet V2 are comparatively analyzed within a unified detection pipeline. Experimental results indicate that the ResNet101-based model achieves the highest localization performance at an intersection-over-union threshold of 0.5, while ResNet50 provides more consistent precision across stricter localization thresholds. Although false-positive detections are observed in acoustically challenging frames, the proposed framework maintains real-time feasibility at approximately 7–8 frames per second. The findings offer practical insights into accuracy–efficiency trade-offs and backbone selection for the development of clinically aware intelligent echocardiography systems, supporting the application of information and communication technology in pediatric cardiac imaging.

**Keywords**: Congenital Heart Disease, Echocardiography, Faster R-CNN, Object Detection, Ventricular Septal Defect

## 1. INTRODUCTION

Congenital heart disease (CHD) is one of the most prevalent congenital abnormalities and remains a leading cause of pediatric morbidity and mortality worldwide. Epidemiological studies report that CHD affects approximately 8–9 per 1,000 live births, with a substantial proportion of cases occurring in low- and middle-income countries where access to specialized cardiac care is limited [1], [2]. Among various CHD subtypes, ventricular septal defect (VSD) is the most frequently diagnosed condition in children and may result in severe complications, including pulmonary arterial hypertension, ventricular dysfunction, and arrhythmias, if not detected and managed at an early stage [3].

Transthoracic echocardiography is the primary imaging modality for diagnosing VSD due to its non-invasive nature and ability to provide real-time cardiac visualization. However, echocardiographic interpretation is highly operator-dependent and requires considerable expertise, particularly in pediatric patients where image quality is often affected by patient movement, small anatomical structures, and acoustic artifacts [4]. These challenges contribute to inter-observer variability and increase the risk of delayed or inaccurate diagnosis, especially in healthcare settings with limited availability of pediatric cardiology specialists.

Recent advances in artificial intelligence and deep learning have demonstrated significant potential in automating medical image analysis and supporting clinical decision-making. In the domain of echocardiography, convolutional neural networks (CNNs) have been applied to tasks such as cardiac view classification, chamber segmentation, and congenital defect detection, showing promising performance under controlled experimental conditions [5]–[9]. Object detection frameworks, in particular, enable not only the identification of pathological findings but also their spatial localization, which is clinically relevant for understanding defect morphology and severity. Despite these advances, many existing studies emphasize algorithmic accuracy and rely on large curated datasets, while paying limited attention to deployment-related constraints encountered in real clinical environments.

From an information and communication technology (ICT) perspective, the development of intelligent echocardiography systems for pediatric care introduces several practical challenges that remain underexplored. These include limited availability of annotated data, heterogeneous ultrasound acquisition protocols, strong dependence on acoustic conditions, and the need to balance detection accuracy with computational efficiency for near real-time use. Addressing these constraints requires a system-oriented evaluation that goes beyond proposing novel algorithms and instead focuses on robustness, feasibility, and design trade-offs within realistic clinical settings.

In this context, the present study investigates the application of a Faster R-CNN-based object detection framework for localizing VSD regions in pediatric cardiac ultrasound videos acquired from the parasternal long-axis view. Rather than introducing a new detection architecture, this work aims to analyze how different widely adopted CNN backbones ResNet50, ResNet101, and Inception-ResNet V2 affect detection performance, localization consistency, and processing speed within a unified pipeline. By framing the problem as the design and evaluation of a clinically constrained intelligent imaging system, this study seeks to provide practical insights for ICT-driven diagnostic support in pediatric cardiology and to inform future development of deployable echocardiography-based decision support tools.

## 2. MATERIAL AND METHODS

This section outlines the research methodology applied to the proposed models. The workflow of the methodology used in this study is illustrated in Figure 1. Figure 1 illustrates the research methodology flowchart for detecting ventricular septal defect (VSD) in pediatric cardiac ultrasound data using the Faster R-CNN model with three different feature extractors: ResNet50, ResNet101, and Inception-ResNet V2. The process begins with the collection and preprocessing of ultrasound video data, which includes frame extraction and annotation to ensure the data is suitable for model training. The prepared data is then used to train and detect VSD using the Faster R-CNN model with each of the feature extractors. Finally, the performance of each model is evaluated using specific metrics, such as average precision (AP), to determine the most effective feature extractor for VSD detection. This methodology highlights a comparative approach to optimizing deep learning techniques for medical image analysis.

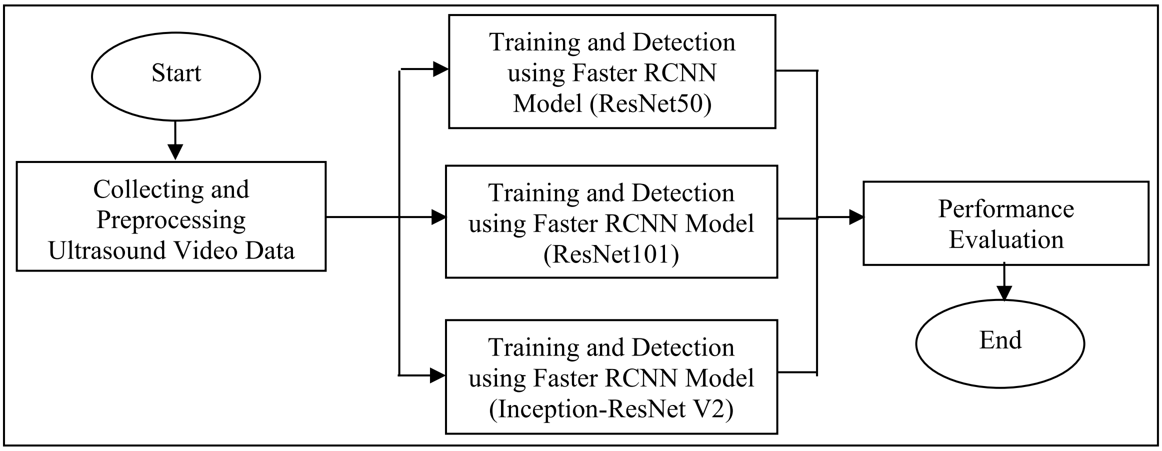

Start
Collecting and Preprocessing Ultrasound Video Data
Training and Detection using Faster RCNN Model (ResNet50)
Training and Detection using Faster RCNN Model (ResNet101)
Training and Detection using Faster RCNN Model (Inception-ResNet V2)
Performance Evaluation
End

FIGURE 1. Research Methodology Flowchart

## 2.1 DATASET

The data used in this study were secondary data in the form of pediatric cardiac ultrasound videos obtained from the medical records of patients in the Cardiology Division of the Department of Pediatric Health, RSUD Dr. Soetomo. The ultrasound videos in the medical records comprised several specific views; however, this study focused exclusively on the parasternal long-axis view, utilizing 11 videos with VSD and 1 normal pediatric cardiac ultrasound video. The ultrasound videos used had a frame width of 640 pixels and a frame height of 480 pixels. Each video had a different frame rate, resulting in a varying number of frames across the videos.

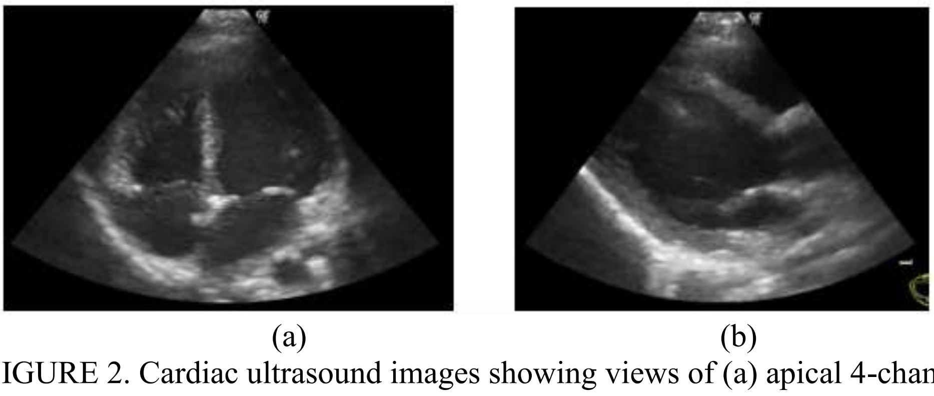
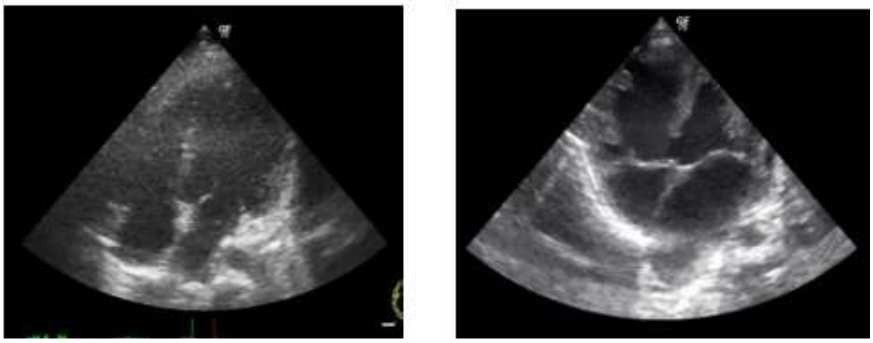

(a) (b)

FIGURE 2. Cardiac ultrasound images showing views of (a) apical 4-chamber and (b) parasternal long axis

A ventricular septal defect (VSD) is a congenital heart defect characterized by an abnormal connection between the left and right ventricles of the heart, leading to hemodynamic disturbances. VSD is the most common congenital heart defect found in children and is the second most frequently encountered defect in adults after bicuspid aortic valve. Most VSDs close spontaneously; however, those that fail to close can lead to complications such as pulmonary arterial hypertension, ventricular dysfunction, and an increased risk of arrhythmias [10].

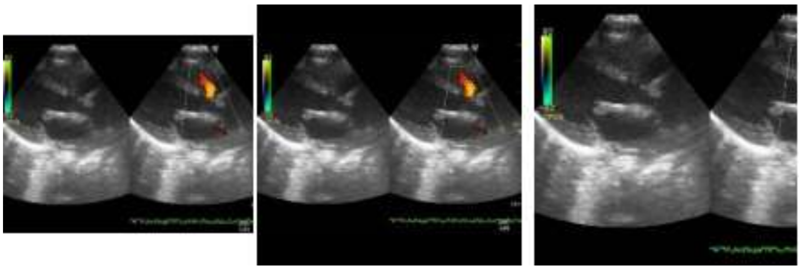


(a) (b)

FIGURE 3. Pediatric Cardiac Ultrasound Images (a) Normal heart and (b) Heart with Ventricular Septal Defect (VSD)

## 2.2 DATA PREPROCESSING AND ANNOTATION

Data preprocessing involved selecting pediatric cardiac ultrasound videos that clearly displayed VSD and one normal cardiac ultrasound video. A total of 10 VSD videos and 1 normal cardiac ultrasound video were utilized. One normal cardiac ultrasound video and one VSD video were designated for testing purposes. The ultrasound videos for training and validation were extracted into individual image frames based on the frame rate and duration of each video. The frames, originally with a resolution of 480 pixels in height and 640 pixels in width, were resized to 360 pixels in height and 480 pixels in width. Padding was then added to the height of the frames to achieve a uniform size of 480×480 pixels.



(a) (b) (c)

FIGURE 4. Pediatric Cardiac Ultrasound Images (a) Original Image (b) Resize and Cropped Image, and (c) Cropped image

Frames displaying VSD were annotated with ground truth boxes in collaboration with clinical supervisors. A total of 88 frames with VSD were obtained, with 80% allocated for training data and 20% for validation data. Data augmentation was performed by cropping the images and adjusting their brightness (both increasing and decreasing) and contrast to enhance the dataset diversity.

## 2.3 FASTER REGION-BASED CONVOLUTIONAL NEURAL NETWORK

Faster R-CNN is part of the development series of region-based convolutional neural networks (R-CNNs). R-CNNs are an object detection approach that applies deep learning models. Faster R-CNN is an advancement of its predecessor architectures, namely R-CNN and Fast R-CNN. Unlike R-CNN and Fast R-CNN, which use selective search to generate proposal regions, Faster R-CNN utilizes a Region Proposal Network (RPN). Faster R-CNN consists of two stages: the RPN for proposing region proposals and the Fast R-CNN detector as the detection network. The general architecture of Faster R-CNN is shown in Figure 5 [11].

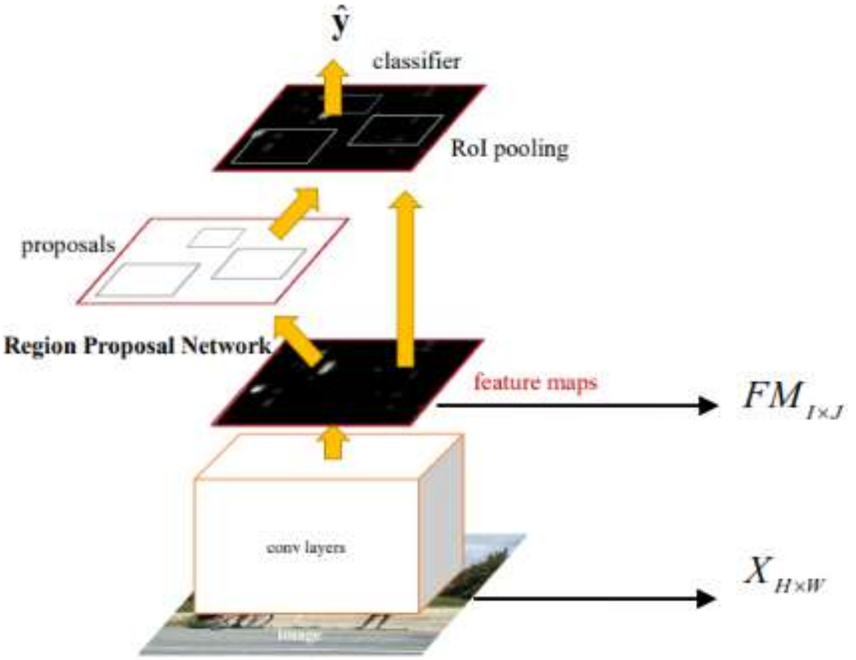


FIGURE 5. Faster R-CNN

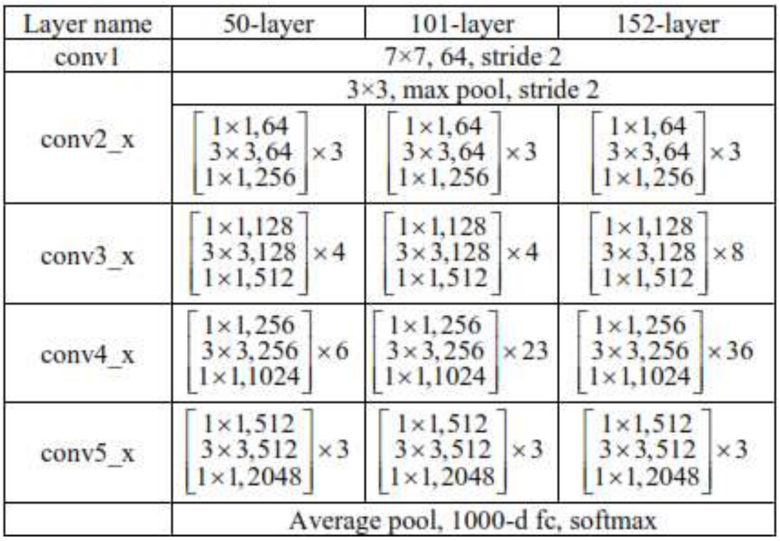
The first stage in Faster R-CNN involves using a feature extractor or a ConvNet consisting of several convolutional layers and pooling layers to extract features from the input image, resulting in a convolutional feature map. The feature extractor is based on a CNN model initialized with a pre-trained model. In this study, three feature extractors were used ResNet50, ResNet101, and Inception-ResNet V2.

ResNet, or Residual Network, is a convolutional neural network (CNN) model that incorporates shortcut connections and is widely used for image recognition tasks. ResNet adopts residual learning across consecutive layers. Let $z$ denote the input to the first layer of a residual block, and let $f(z,\{w_i\})$ represent the residual function learned by the stacked layers, where $w_i$ denotes the weights of the $i$-th layer. The shortcut connection performs identity mapping, and the output of the residual block is given by

$$y = f(z,\{w_i\}) + z$$

assuming that $f(z,\{w_i\})$ and $z$ have the same spatial dimensions. These shortcut connections do not introduce additional parameters or increase computational complexity significantly [11]. Several ResNet architectures have been developed based on this principle, including ResNet50 and ResNet101. The architectural details are shown in Table 1. In layers from conv2 to conv5, building blocks are represented in parentheses, where ×3 indicates the presence of 3 stacked blocks.

TABLE 1.
Detailed Architecture of ResNet



| Layer name | 50-layer | 101-layer | 152-layer |
|---|---|---|---|
| conv1 | 7×7, 64, stride 2 | | |
| conv2_x | 3×3, max pool, stride 2 | | |
| | $\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$ | $\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$ | $\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$ |
| conv3_x | $\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$ | $\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$ | $\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$ |
| conv4_x | $\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$ | $\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$ | $\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$ |
| conv5_x | $\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$ | $\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$ | $\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$ |
| | Average pool, 1000-d fc, softmax | | |

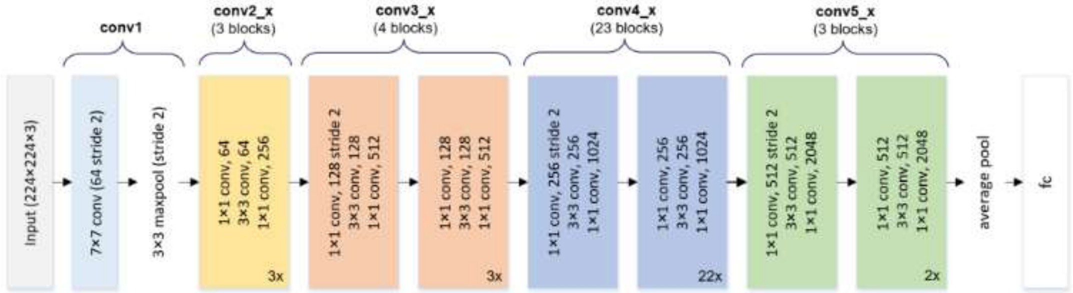


FIGURE 6. The Architecture of ResNet-101

Inception-ResNet V2 is a combination of residual connections and the Inception architecture. It comprises a stem, Inception-ResNet, and reduction blocks. The overall architecture of Inception-ResNet V2 with an input image of size 299×299 is shown in Figure 7 [12].

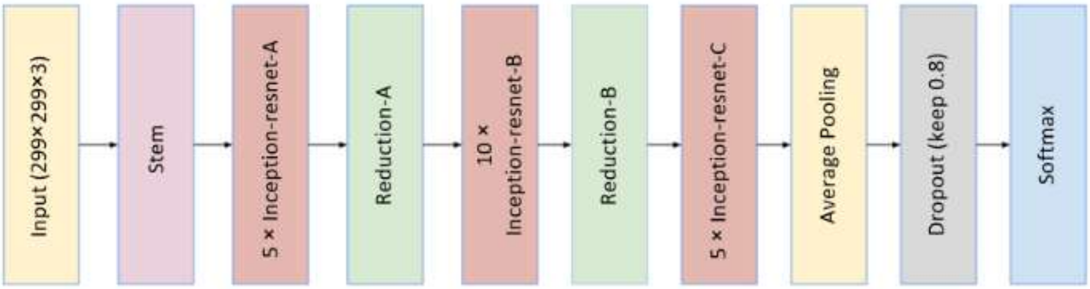


FIGURE 7. The Architecture of Inception-ResNet V2

Szegedy et al. [12] introduced Inception-ResNet V2 as a hybrid architecture that combines the benefits of the Inception module and residual connections to improve deep learning performance. The model builds upon the success of Inception modules by integrating filters of varying sizes in parallel to extract diverse features, while residual connections facilitate efficient gradient flow and faster convergence in deep networks. The architecture includes a stem for feature extraction, Inception-ResNet blocks for deeper hierarchical learning, and reduction blocks for dimensionality reduction. Designed for large-scale image classification tasks, Inception-ResNet V2 achieves state-of-the-art accuracy on benchmarks like ImageNet while maintaining computational efficiency. By incorporating residual connections, the model alleviates issues like vanishing gradients, making it scalable to greater depths. This innovation has paved the way for advancements in both classification and object detection tasks, showcasing its versatility and robustness in handling complex datasets.

## 2.4 STOCHASTIC GRADIENT DESCENT (SGD)

Stochastic Gradient Descent (SGD) is an optimization algorithm. SGD optimizes parameters by minimizing the loss function or cost function. Parameter optimization using SGD involves selecting a mini-batch of $m$ training samples and calculating the average gradient in each iteration. The calculated gradient is then used to update the parameters. The steps for parameter updating using stochastic gradient descent in RPN are as follows:

1. Determine the learning rate schedule $\in_1, \in_2, \ldots$
2. Determine the initial parameters $\theta_k = \theta_0$
3. Select one sample from the training data corresponding to the target $y_l$, where $l = 1, 2, \ldots, n$, with $n$ being the total number of training samples. Perform operations using $\theta_k$ on the selected $x_l$.
4. Take a sample in the form of a mini-batch consisting of $m$ anchors.
5. Calculate the gradient estimate of the loss function with respect to all parameters $\theta_k$ by finding the partial derivatives using the chain rule method.
$$\hat{g}_k \leftarrow \frac{1}{m} \nabla_{\theta_k} \sum_i^m L\left(\{p_i\},\{t_i\}\right) = \frac{1}{m} \nabla_{\theta_k} \sum_i^m L\left(L_{cls} + L_{reg}\right)$$
6. Update the parameter $\theta_k \leftarrow \theta_{k-1} - \in_k \hat{g}_k$

Repeat step 3 using the updated $\theta_k$ parameters or weights until the stopping criterion is met.

## 2.5 EVALUATION METRIC

The evaluation metric used in this study is Average Precision (AP). Mathematically, AP is expressed in Equation (1).

$$AP = \sum_{q=0} \left(r_{q+1} - r_q\right) p_{interp}\left(r_{q+1}\right), \tag{1}$$

$$p_{interp}\left(r_{q+1}\right) = \max_{\overline{r}:\overline{r} \geq r_q + 1} p\left(\tilde{r}\right), \tag{2}$$

$p\left(\tilde{r}\right)$ in Equation (2) represents the precision at a specific recall. Precision and recall can be calculated using Equations (3) and (4).

$$Precision = \frac{TP}{TP + FP}, \tag{3}$$

$$Recall = \frac{TP}{TP + FN}, \tag{4}$$

These values are calculated based on the Intersection over Union (IoU) between the predicted bounding box and the ground truth box, which is compared against a predefined IoU threshold.

## 3. RESULTS AND DISCUSSION

There are nine heart ultrasound videos with Ventricular Septal Defect (VSD) taken from the parasternal long axis view. These ultrasound videos were obtained from the medical records of pediatric patients aged between 3 months and 13 years, with VSD sizes ranging from 0.26 cm to 1.34 cm.

TABLE 2.
Histogram of a Single Data Frame

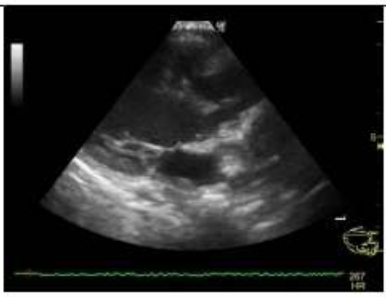
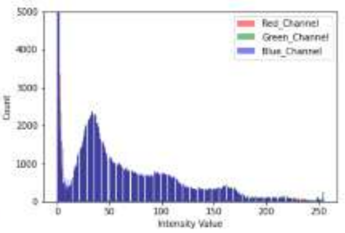
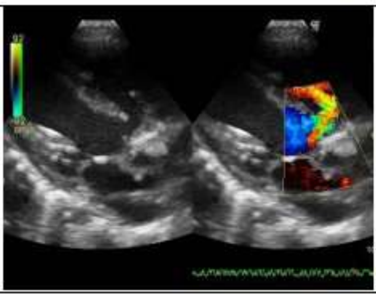
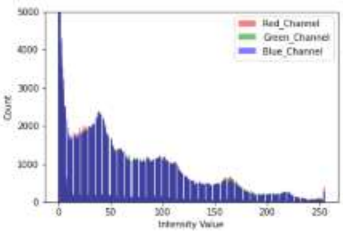
| ID | Frame | Image Histogram |
|---|---|---|
| P1 |  |  |
| P2 |  |  |

Table 2 displays one of the frames used in the study along with its corresponding image histogram. The histogram of each frame shows a high frequency of intensity values at 0, as the images are predominantly black. However, in Table 2, the histogram's y-axis is limited to a maximum value of 5000 for better visualization of the intensity distribution. A common artifact in ultrasound imaging is acoustic shadows, which appear as dark areas caused by tissue absorption or reflection. This artifact contributes to the overall darker appearance of the ultrasound images. Frames extracted from the nine cardiac ultrasound videos in the parasternal long axis view were selected to ensure the presence of visible VSD. Table 3 summarizes the number of frames used in each stage: training, validation, and testing. The testing data consists of both VSD and normal heart ultrasound videos, including frames with visible VSD as well as frames without it. The AP (Average Precision) value is calculated exclusively based on the frames where VSD is visible.

TABLE 3.
Division of the Data

| Training | Validation | Testing |
|---|---|---|
| 520 | 27 | 208 |

### 3.1 FASTER R-CNN MODEL FOR VSD DETECTION IN ULTRASOUND VIDEOS

This study employs the Faster R-CNN model with different feature extractors to detect VSD in ultrasound videos. The Faster R-CNN framework utilized in this study is illustrated in Figure 6, as proposed by Girshick et al. (2016). Faster R-CNN operates in two stages: the first stage is the Region Proposal Network (RPN), and the second stage is the detector network, which uses Fast R-CNN as the backbone. The input image, with a height $H = 480$ pixels, a width $W = 480$ pixels, and three RGB channels, is initially processed through several convolutional layers in the feature extractor to generate a convolutional feature map. The feature extractor in Faster R-CNN is based on a CNN backbone network. Each CNN model incorporates varying types of layers and parameters, which subsequently affect the model's performance and speed in detecting VSD. The number of parameters in the Faster R-CNN model varies depending on the feature extractor used. The Faster R-CNN model with the Inception-ResnNetV2 feature extractor has the highest number of parameters compared to other feature extractors. The large number of parameters significantly impacts the training process. The training is conducted with a batch size of 1, resulting in 100,000 iterations.

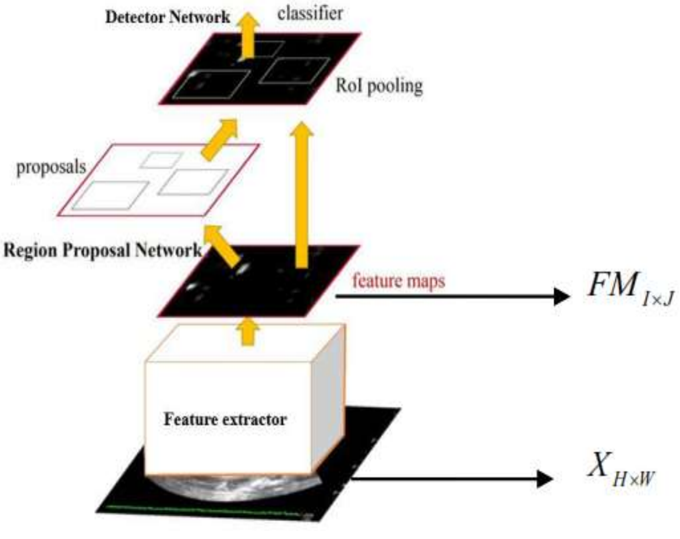


FIGURE 8. Illustration of Faster R-CNN on Ultrasound Images

TABLE 4.
Number of Parameters in the Model

| Feature extractor | Number of Paremeters |
|---|---|
| ResNet50 | 56,605.858 |
| ResNet101 | 94.642.338 |
| Inception-ResNetV2 | 118.733.282 |

Each model was trained for 100,000 iterations and initialized using a pretrained CNN classification model on ImageNet. The evaluation results for each model are shown in Table 5.

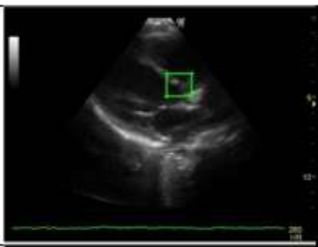
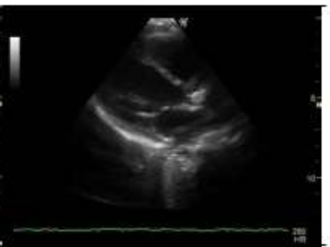
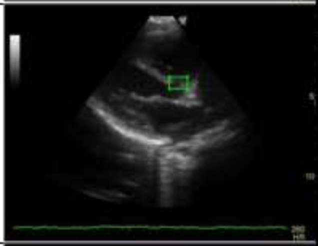
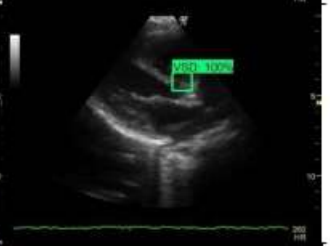
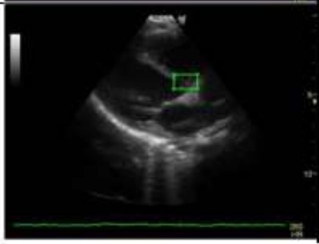
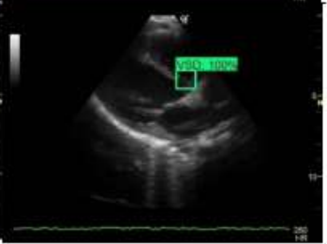
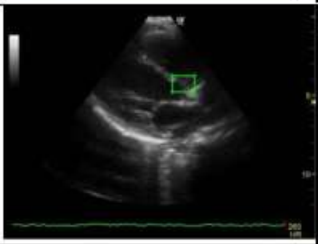
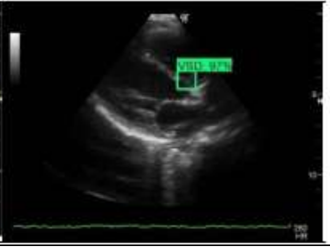
TABLE 5.
Average Precision of Models

| Feature extractor | TP | FP | $AP^{0.5}$ | $AP^{[0.5:0.95]}$ |
|---|---|---|---|---|
| ResNet50 | 23 | 4 | 0.744 | 0.285 |
| ResNet101 | 23 | 4 | 0.758 | 0.277 |
| Inception-ResNetV2 | 20 | 7 | 0.717 | 0.246 |

The Faster R-CNN model based on ResNet101 achieved the highest $AP^{0.5}$ among all models, with a value of 75.8% on the validation data using an IoU threshold of 0.5. A total of 23 VSD bounding boxes were accurately predicted on the validation data. Meanwhile, the model based on ResNet50 recorded the highest mean AP, with a value of 27.7%, when using IoU thresholds ranging from 0.5 to 0.95 in 0.05 increments. The Faster R-CNN model with ResNet50 also accurately predicted 23 VSD bounding boxes, but it generated four incorrect bounding box predictions based on an IoU threshold of 0.5.

The $AP^{0.5}$ achieved during testing with this model is 71.3%. Some frames that did not display VSD were falsely detected as containing VSD by this model. On average, the Faster R-CNN model with ResNet101 required 0.13 seconds to detect a single frame, equivalent to processing approximately 7–8 frames per second.

TABLE 6.
Detection results with the Faster R-CNN model based on ResNet101

| Frame | Ground-Truth | Prediction |
|---|---|---|
| 15 |  |  |
| 50 |  |  |
| 73 |  |  |
| 124 |  |  |

## 4. CONCLUSION

This study demonstrated the effectiveness of Faster R-CNN in detecting ventricular septal defect (VSD) in pediatric cardiac ultrasound videos. By employing CNN-based feature extractors, including ResNet50, ResNet101, and Inception-ResNet V2, the proposed method achieved notable performance in both validation and testing phases. ResNet101 yielded the highest $AP^{0.5}$ of 75.8%, while ResNet50 achieved the best average precision across multiple IoU thresholds, indicating its robustness for diverse detection tasks. The results highlight the potential of Faster R-CNN as a diagnostic aid for early detection of congenital heart defects, reducing reliance on manual evaluations and enhancing diagnostic accuracy. However, limitations such as false positive detections and processing time suggest avenues for further optimization. Future work may focus on integrating advanced data augmentation techniques, multi-view echocardiography, and lightweight architectures to improve both accuracy and scalability for real-world clinical applications.

## REFERENCES

[1] M. J. van der Bom *et al.*, "The global burden of congenital heart disease in adults," *Circulation*, vol. 134, no. 8, pp. 568–578, 2016.

[2] W. Dakkak *et al.*, "Ventricular septal defects: Pathophysiology, clinical presentation, and management," *Journal of Pediatric Cardiology*, vol. 38, no. 2, pp. 112–124, 2024.

[3] S. Ghosh, K. Raghunathan, and S. Ishikawa, "The impact of ventricular septal defects on cardiac function and surgical outcomes," *International Journal of Cardiology*, vol. 257, pp. 56–62, 2018.

[4] L. G. Rudski *et al.*, "Guidelines for the echocardiographic assessment of cardiac structure and function," *Journal of the American Society of Echocardiography*, vol. 33, no. 5, pp. 256–275, 2020.

[5] R. Arnaout *et al.*, "Deep-learning models improve automated classification of congenital heart disease on echocardiographic screening," *Nature Medicine*, vol. 26, no. 8, pp. 1224–1228, 2020.

[6] D. Ouyang *et al.*, "Video-based artificial intelligence for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, pp. 252–256, 2020.

[7] X. Jiang *et al.*, "A deep learning-based method for pediatric congenital heart disease detection with standard echocardiographic views," *World Journal of Pediatric Surgery*, vol. 6, no. 3, e000580, 2023.

[8] J. Wang, X. Jiang, and H. Liu, "AI-driven echocardiography: A new era in congenital heart disease diagnosis," *Circulation: Cardiovascular Imaging*, vol. 14, no. 2, pp. 115–127, 2021.

[9] M. Nasrudin *et al.*, "On the YOLOv4 architecture for fast and real-time congenital heart disease detection via ultrasound videos," *MATEMATIKA*, vol. 38, no. 2, pp. 103–114, 2022.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[12] C. Szegedy *et al.*, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[13] L. Chen *et al.*, "Automated ventricular septal defect screening in echocardiography using deep learning," *IEEE Access*, vol. 9, pp. 126937–126948, 2021.

[14] J. G. S. M. de Carvalho *et al.*, "Deep learning-based view classification and segmentation in echocardiography," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 837–847, 2021.