MATEMATIKA, MJIM, Volume 38, Number 2, 103–114

On the YOLOv4 Architecture for Fast and Real Time Congenital Heart Disease Detection Via Ultrasound Videos

[1]Muhammad Nasrudin*, [2]Nur Iriawan, [3]Kartika Fithriasari, [4]Anindya Apriliyanti Pravitasari and [5]Taufiq Hidayat

[1,2,3]Department of Statistics, Faculty of Science and Analytics Data
Institut Teknologi Sepuluh Nopember, 60111 Surabaya, Indonesia

[4] Department of Statistics, Faculty of Mathematics and Natural Sciences
Padjajaran University, 45363 Sumedang, Indonesia

[5]Department of Child Health, Faculty of Medicine
Airlangga University, 60132 Surabaya, Indonesia

*Corresponding author: nasrudin.sta068.its@gmail.com

Article history
Received: 30 September 2021
Received in revised form: 23 May 2022
Accepted: 6 July 2022
Published online: 31 August 2022

Abstract Congenital Heart Disease (CHD) is one of the most frequent cardiac defects in infants, and it is becoming more common. Various research studies have been conducted for CHD identification based on clinical and non-clinical data. This study conducts an artificial intelligence system for real-time congenital heart disease (CHD) detection using ultrasound videos. The YOLOv4 (You Only Look Once) is employed for localizing the congenital defect through ultrasound videos. The performance is evaluated by mean Average Precision (mAP) which compares the classification results with the medical ground truth. The model has great performance in CHD detection with mAP values of training data of 98.36% for YOLOv4 and 87.24% for YOLOv4 tiny. This is useful for doctors and radiologists who require a simple, fast, yet accurate model for the detection of CHD.

Keywords Artificial Intelligence; Congenital Heart Disease; Mean Average Precision; YOLOv4

Mathematics Subject Classification 20E28, 68U10, 68T07.

1 Introduction

Congenital Heart Disease (CHD) is a structural abnormality of the heart and blood vessels that appears at birth and is the main cause of child death from all congenital disorders. The main cause is the failure of the formation of cardiac structures in the early stages of fetal formation in the womb. The incidence of CHD worldwide is estimated at 1.2 million cases out of 135 million live births each year. Of these, around 300,000 cases are categorized as severe CHD that require complex surgery to survive [1]. Based on a statement issued by the British Congenital Cardiac Association, several conditions were found in patients with CHD that were potentially at risk for severe infection with COVID-19 infection.

The most common type of CHD found is called Ventricular Septal Defect (VSD), which is a defect in the heart where there is a hole in the interventricular septum that separates the right and left ventricles. This defect in the interventricular septum results from incomplete septal formation during complex embryological morphogenesis of the heart [2]. One technique to detect CHD is to perform echocardiography. Echocardiography, also known as cardiac ultrasound, is an examination of the heart and large blood vessels using ultrasound waves [3]. This examination is intended to assist doctors in determining the diagnosis, predicting the prognosis of cases of heart and blood vessel disease, and selecting the appropriate medical procedures. This makes echocardiography an important role in the development of modern heart disease. Examination with echocardiography produces morphology regarding the chambers and valves of the heart, as well as function and hemodynamic conditions (without inserting the device into the body), so it has a low risk and almost no side effects. The inspection works by sending sound waves and recording the echoing waves. The wave is then measured and displayed by a device, which then produces visualization through the monitor directly. Captured visualizations can be saved in the form of video or can be cropped from video frames as images or pictures. Unfortunately, not all pediatricians can read the results of the visualization. Only doctors with a pediatric heart specialty can read and determine the diagnosis and manifestations of CHD. If not detected early and not treated properly, 50% of CHD deaths will occur in the first month of life. Seeing these problems, this study wants to analyze the results of the ultrasound video to help pediatricians in detecting CHD early so that they can provide appropriate referrals to cardiologists. This will help the pediatric cardiologist work.

This research utilizes technology and a deep learning approach. In the health and medical industry, the accuracy of predicting disease is very important and requires effective decisions in taking analysis and predicting the accuracy of a patient's disease. Early detection and good early treatment are needed to minimize the level of morbidity and mortality. Medical data processing is often done with a deep learning approach because it can process very complex data structures [4][5][6]. The model of object detection includes You Only Look Once (YOLO) [7], SSD [8], Fast-R-CNN [9], Faster-R-CNN [10], and SPP-Net [11]. A deep learning approach is used to perform early detection of CHD. The detection is the result of an analysis of the ultrasound video of the child's heart. The architecture incorporated into the Convolutional Neural Network (CNN) called YOLO, which allows converting video into a sequential image as input, is employed in this research. Many object detection studies for health cases have been carried out, such as malaria detection research using a single-stage detector (SSD) and a two-stage detector Faster R-CNN [12]. The Faster R-CNN has achieved 71.0% of mAP, while SSD can reach 71.4% of mAP. The SSD model has the highest speed for detection speed but has the worst detection. The Faster R-CNN model is the slowest in detection speed, but the accuracy is comparable to the SSD model. Detection of septal defects has been researched by [13] using deep learning-based multiclass instance segmentation. Another approach is a random forest predicting each frame but can't automatically annotate video data [14]. [15] has succeeded in using the COCO dataset to perform object detection by comparing several architectures, including CenterMask, ASFF, ATSS, EfficientDet, YOLOv3, and YOLOv4. As a result, YOLOv4 can detect objects twice faster than EfficientDet and experience an increase in Average Precision (AP) and frames per second (FPS) by 10% and 12%, respectively, compared to YOLOv3. YOLOv4 has also been used to detect malaria, obtaining an accuracy of 96% with a time resolution of 30fps [12]. It is very fast compared to SSD and has very good accuracy. An evaluation of deep learning methods for small object detection was conducted by [16] and the result was that YOLO has better performance than Fast R-CNN, Faster R-CNN, and RetinaNet.

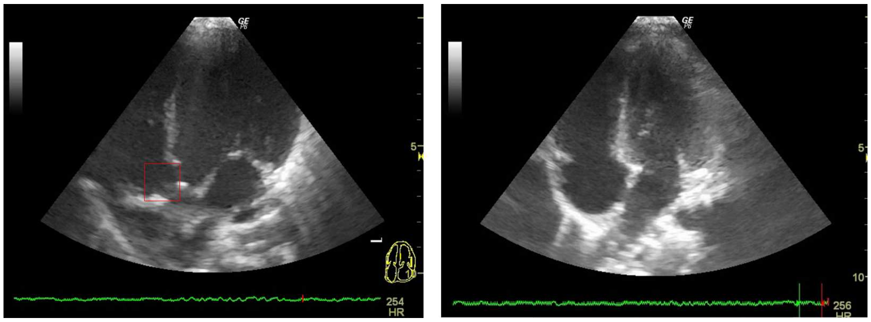
In contrast to previous research, which tries to detect chamber heart defects, this study focuses on finding the defect in the VSD. It is challenging due to the size matter and the abstract visualization of VSD. This study uses a deep learning approach with the YOLOv4 architecture to detect the presence or absence of CHD in children through ultrasound video. It would assist doctors and medical personnel in early and rapid detection of heart defects in children. So that doctors and medical personnel in the remote area can participate in detecting and treating cases of CHD in children early. That in turn will help the handling of CHD patients correctly, precisely, and quickly.

2 Material and Methods

2.1 Dataset

This study focused on the Ventricular Septal Defect (VSD). The dataset used in this study contains 20 ultrasound videos with VSD recorded from Dr. Soetomo hospital with 640×480 pixel. The video is converted into a number of frames, which is calculated by the length of the video duration multiplied by the number of frames per second (fps). Therefore, it consists of the 744 frames used in this study. Cropping frames are used to make a square in 480×480 pixel. All VSD detection models in this study were trained using 80% of the dataset and 20% for testing data.

Image annotation aims to provide information in the form of class names and positions in the form of bounding boxes of objects to be detected. A Bounding box is the most common type of data labeling annotation used in computer vision in the form of a rectangular box that is used to determine the location of the target object. The bounding box is usually represented by two coordinates and/or by one coordinate and the width and height of the bounding box. The labeling process only uses one class, namely VSD. The labeling process uses GUI-based software, namely Labellmg by Tzutalin [17]. Figure 1 depicts an example of VSD with corresponding defect location in red bounding box (Figure 1(a)) and normal cardiac (Figure 1(b)).



(a) (b)

Figure 1: Ultrasound Image of (a) VSD (red bounding box is the defect location) and (b) Normal Cardiac

2.2 YOLO for Fast and Real Time Object Detection

YOLO is an approach in deep learning that performs object detection [7,18]. It is targeted at real-time processing and framing objects as a single regression problem from direct image pixels to separate spatial bounding boxes and associated probability classes. YOLO performs object detection and recognition like the human brain. When human look at something, the brain instantly recognizes and concludes what is being seen. This study used YOLO architecture since it is very fast and accurate.

2.2.1 YOLOv4 Architecture

YOLOv4 is the improved version of YOLOv3 which has great performance in speed and accuracy. The features added in YOLOv4 are two methods called Bag of Freebies (BoF) and Bag of Special (BoS). Both of them are applied to the detector module's backbone. BoF affects the training process, such as data augmentation, soft labelling, and cost function, to imbalance class which can produce accuracy in the system without increasing inference costs. The word Bag refers to a set of methods or strategies, and Freebies means that the inference accuracy will increase without having to affect the load on the hardware.

BoF for the backbone in YOLOv4 uses data augmentation which is intended to increase the variability in the input images, so that the object detection model designed has a higher resistance to images obtained from different environments. YOLOv4 uses BoF for the backbone, namely CutMix and Mosaic Data Augmentation, DropBlock Regularization, and Class Label Smoothing. In the detector, the BoF used are CIoU-loss, CmBN, DropBlock regularization, Mosaic data augmentation, Self-Adversarial Training, Eliminate grid sensitivity, Multiple anchors for single ground truth, Cosine annealing scheduler, Optimal hyperparameters, and Random training shapes. While BoS, the word specials refers to getting something of value at a discount or low price. By analogy, it can be said as a series of modules that only increase the inference cost by a small amount but significantly increase the accuracy of object detection. That is the Bag of Specials meant. The types of BoS used in the backbone are Mish Activation, Cross-stage Partial Connections (CSP), Multi-input Weighted Residual Connections (MiWRC). The detectors used, on the other hand, are Mish Activation, SPP-block, SAM-block, PAN Path-Aggregation Block, and DIOU NMS [15]. The difference between YOLOv3 and YOLOv4 architecture for detecting the VSD is presented in Figure 2 and Figure 3, respectively.

2.2.2 Object Detection with YOLOv4

Localization is carried out in the YOLOv4 classification, which means that there is an additional object location assignment in the form of bounding boxes (bx, by, bh, bw). The steps for using YOLOv4 to detect objects are as follows.

(a) Reads image of any size.

(b) Resize the image. The larger the size of the given image, the more accurate the prediction. However, it impacts the slower computational process and vice versa. Another step is to create a grid on the image with the size of S×S grids. For example, for 416×416 pixel image size with S=13, we get 32×32 pixel grid cells. The visualization of the grid example shown in Figure 4.

(c) Perform object detection mapping on each grid cell with fully connected layers and activation function. In this study, we use Leaky ReLU (*Rectified Linear Unit*) and *mish* activation function. In each grid cell, fully connected and activation functions are performed to obtain object detection scores, so that a map for the class probabilities is obtained for each grid cell. Equation (2) is the

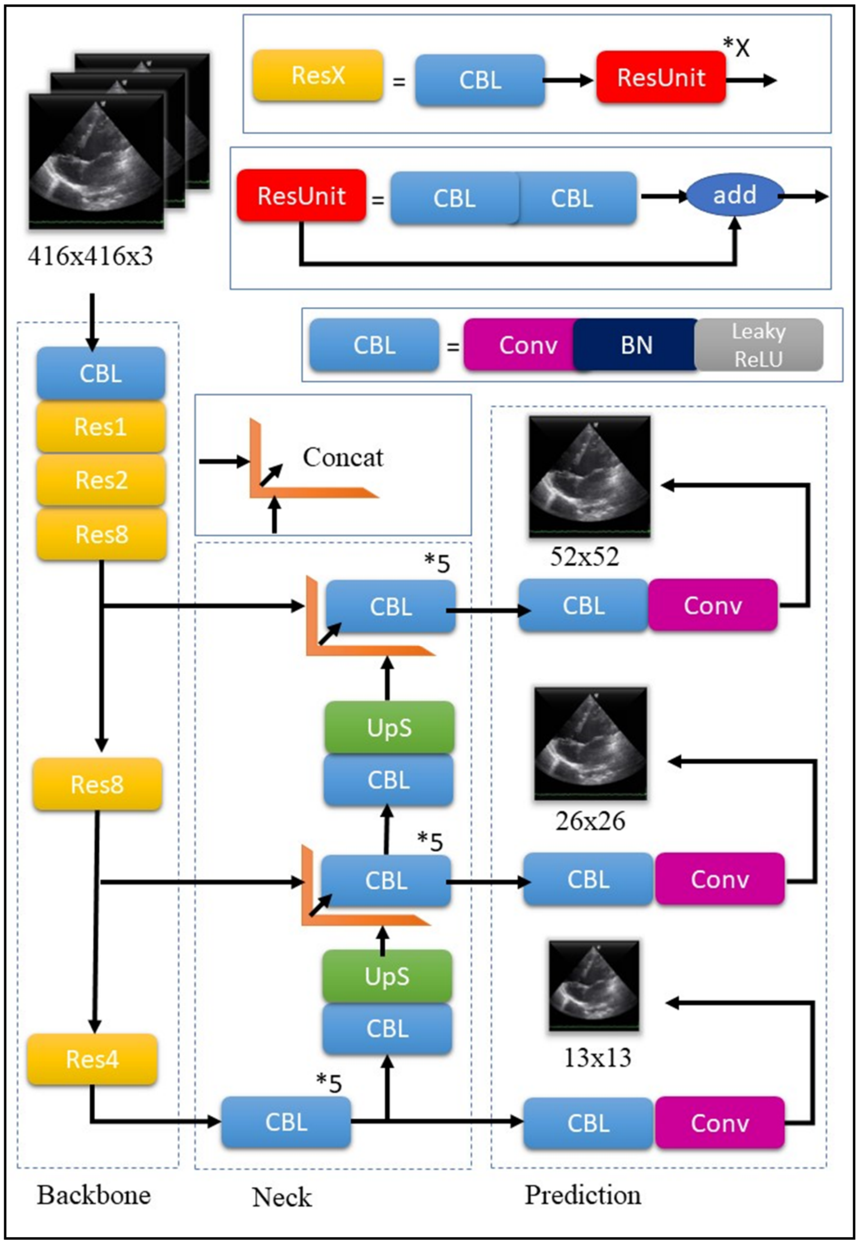


Figure 2: YOLOv3 Architecture

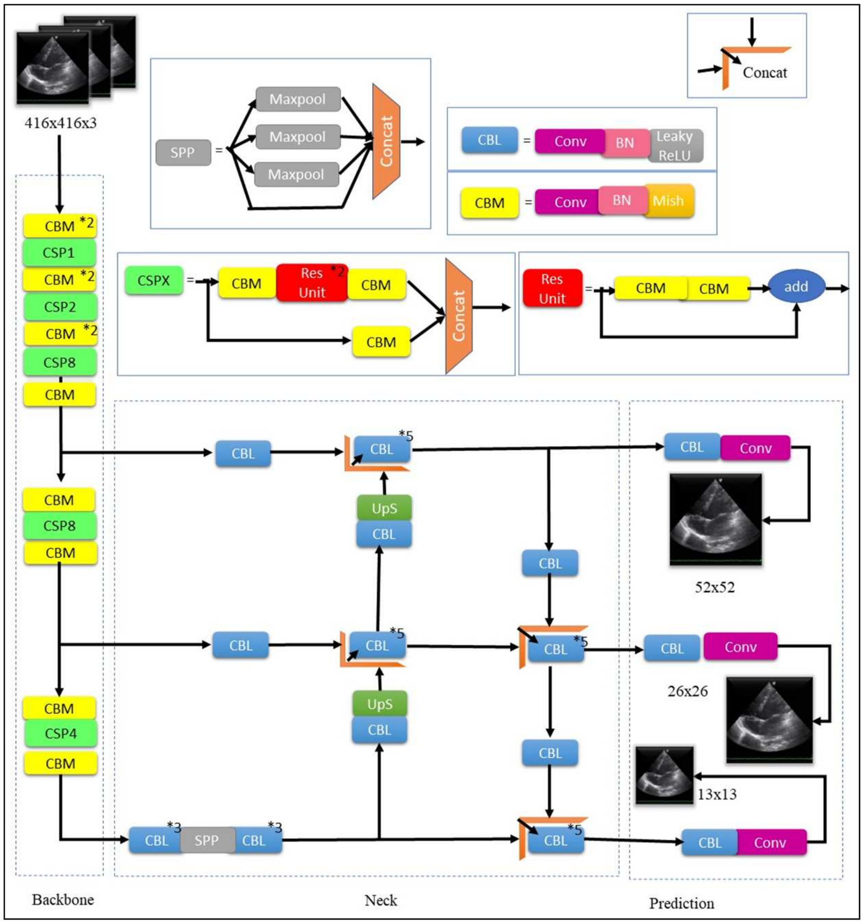
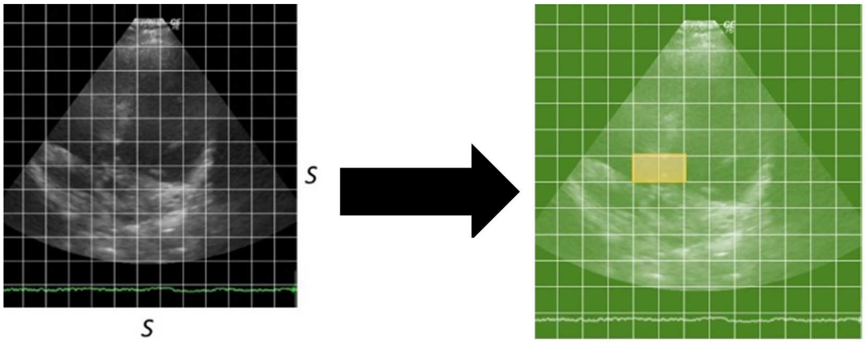


Figure 3: YOLOv4 Architecture



(a) (b)

Figure 4: (a) The Grid Cell, (b) Map Probability Class

Leaky ReLU and Equation (3) is the mish activation function.

$$FM_{a,b} = bias + \sum_{c}^{C} \sum_{d}^{D} Z_{c,d} \times X_{a+c-1,b+d-1}, \tag{1}$$

$$f(FM_{a,b}) = \max(0.01FM_{a,b}, FM_{a,b}), \tag{2}$$

$$f\left(FM_{a,b}\right) = FM_{a,b}.\tanh\left(\ln\left(1 + e^{FM_{a,b}}\right)\right). \tag{3}$$

(d) Detecting objects in each grid cell. Each grid cell containing the object will be responsible for detecting it. Furthermore, if there are objects in the grid cell, a bounding box will be drawn to indicate the existence of a predetermined object.

(e) Perform non max suppression to get a bounding box with a maximum confidence score. Furthermore, after all the grid cells have carried out the object detection process, several bounding boxes are obtained that mark the existence of objects. Among the bounding boxes, one will be chosen which will mark the object with the highest probability or confidence score, while the bounding box that contains the same object that has a lower confidence score will be removed.

2.2.3 Model Evaluation

Performance evaluation via mean average precision (mAP) determines how well the model is. The bounding box results, from an annotation data detection system, and ground truth that is annotated by researchers with the doctor validation [19], are then processed in the confusion matrix as in Table 1. The data is grouped as True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), which has been defined as Interest of Union (IoU). Performance Parameters Calculation Flow can be illustrated in Figure 4.

Table 1: Confusion Matrix

		Actual	
		Positive	Negative
Predict	Positive	TP	FP
	Negative	FN	TN

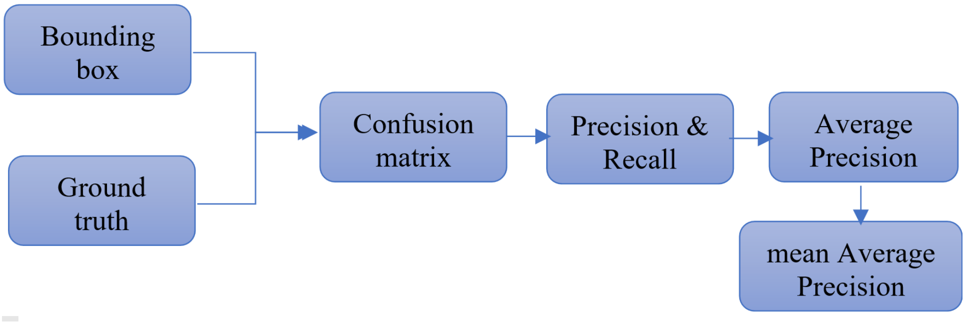


Figure 5: Performance Parameters Calculation Flow

As seen in Figure 5, several metrics of evaluation can be calculated based on the value in the confusion matrix. Precision and Recall formulation are in equations (4) and (5), respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{5}$$

Precision represents the model's ability to identify the related object as a percentage value of the correct prediction. Recall is the ability of the model to find all relevant objects as a true positive percentage value that can be detected in all ground truths. The mean Average Precision (mAP) is the average value of Average Precision (AP) which forms the performance evaluation matrix of object detection. The AP value is obtained from the Precision calculation in Equation (4) and the Recall in Equation (5) which is then calculated as in Equation (6).

$$\text{AP} = \sum_{n=0} (r_{n+1} - r_n)\, p_{\text{interp}}(r_{n+1}), \tag{6}$$

$$p_{\text{interp}}(r_{n+1}) = \max_{\tilde{r}:\tilde{r} \geq r_{n+1}} p(\tilde{r}),$$

where p is Precision, p_{interp} is Precision interpolation, while r is Recall and $p(\tilde{r})$ is Precision calculated on the Recall. AP is the area under curve of Recall and Precision. The curve is sampled at all unique recall values $(r_1, r_2, ..., r_n)$ so r_n and r_{n+1} are according to the Recall values [20].

3 Result and Discussion

The hyperparameter setting in this study is shown in Table 2. Image size for input resize 416×416 pixel, YOLOv4 predict 3 different classification scales and localization VSD, so 3 outputs will be produced. Classification and localization VSD with 3 scales do YOLOv4 has the aim of being carried out on various sizes of objects, both large, medium, and small objects. At each scale in YOLOv4, each grid cell predicts 3 bounding boxes using 3 anchors, bringing the total number of anchors used to 9 anchors. YOLOv4-tiny used 6 anchors. In the training process, a batch size of 64 is used and the maximum number of train steps is 2,000. The Batch size with a value of 64 causes 64 images to be taken as samples and updates the weight and bias in 1 train step. A Subdivisions or minibatch with a value of 16 means that in four images are taken in one subdivision or minibatch. Each train step/iteration includes weight and bias updates based on batch size, as well as checking average loss and accuracy for training data. The total number of images trained to obtain one model is 128,000 images.

Table 2: Hyperparameter of Study

Model	***Hyper-parameter***	**Value**
CSP-Darknet53	Image Size	416×416
	Batch size	64
	Subdivisions	16
	Training Step	2000
	Learning Rate	0.001

YOLOv3	Anchors	(10×13);(16×30);(33×23);
		(30×61);(62×45);(59×119);
		(116×90);(156×198);(373×326)

YOLOv4	Anchors	(12×16);(19×36);(40×28);
		(36×75);(76×55);(72×146);
		(142×110);(192×243);(459×401)

YOLOv4-tiny	Anchors	(10×14);(23×27);(37×58); (81×82);(135×169);(344×319)

The first evaluation value in this study is the confusion matrix, which is used to determine the results of the classification goodness. Then the precision and recall values can be calculated. The counting members of goodness of classification are presented in a confusion matrix in Table 3. The second evaluation value is mAP, which is the most popular evaluation value in object detection. From the confusion matrix, the Precision and Recall could be calculated. The value of Precision, Recall, mAP testing, and training are shown in Table 4. In addition, the visual analysis of the best result is YOLOv4 with 3 fps is exhibited in Figure 6.

Table 3: Confusion Matrix

YOLOv4		Predicted		YOLOv4-tiny		Predicted	
		Positive	Negative			Positive	Negative
Ground-Truth	Positive	123	1	Ground-truth	Positive	120	4
	Negative	1	22		Negative	3	11

Table 4: Performance YOLOv3, YOLOv4 and YOLOv4-tiny

Performance	YOLOv3	YOLOv4	YOLOv4-tiny
Recall	99.00%	99.00%	97.00%
Precision	99.00%	99.00%	98.00%
mAP training	97.25%	98.36%	95.35%
mAP testing	94.78%	95.56%	87.24%

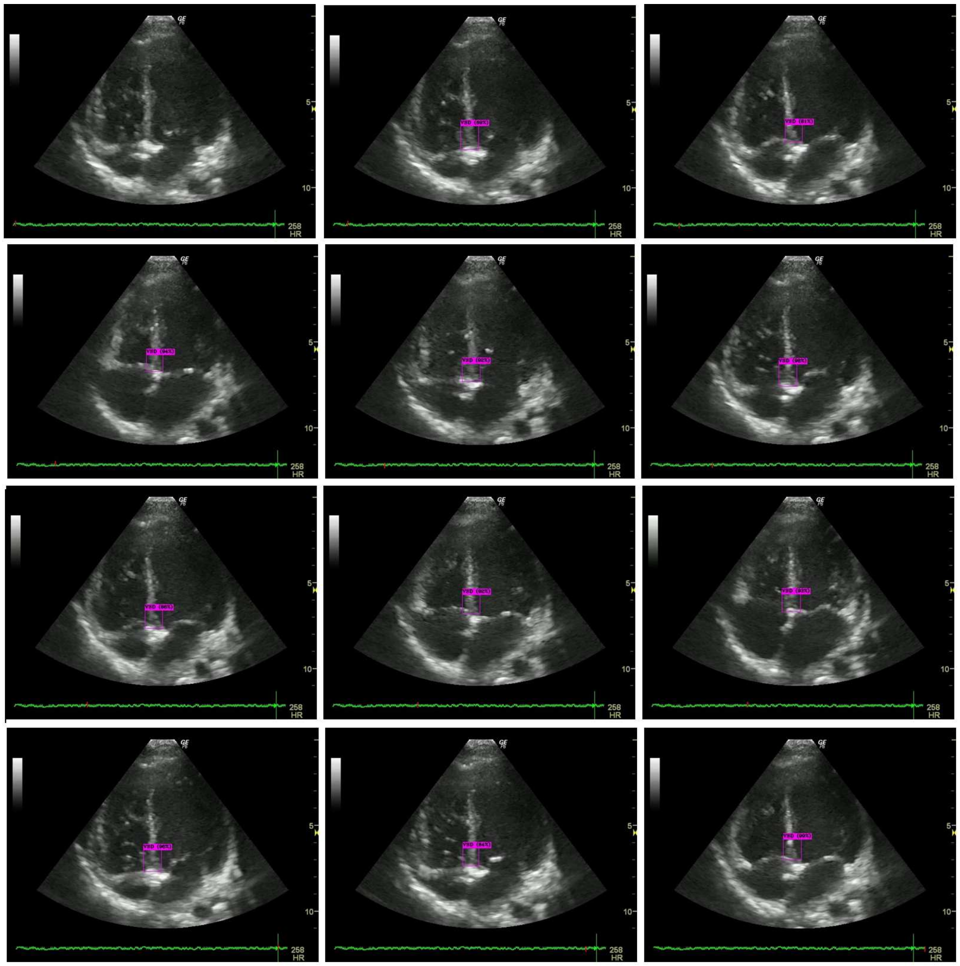


Figure 6: Example of VSD Detection Results in an Ultrasound Video

4 Conclusions

This paper has succeeded in demonstrating the work of YOLOv4 and YOLOv4-tiny in detecting VSD with mAP testing data 95.56% for YOLOv4 and 87.24% for YOLOv4- tiny. YOLOv4-tiny can detect until 132fps, which is better than YOLOv4 only detecting until 40fps. We have demonstrated the feasibility and effectiveness of proposed YOLOV4-based architectures for VSD detection in using ultrasound videos. Future work will be difficult to investigate the applicability of these algorithms for detecting the various CHD. The reliability of these algorithms with another model, on the other hand, will also be a good seed for further research.

Acknowledgements

The authors acknowledge Dr. Soetomo Hospital for providing data for this research in ethical clearance 0199/KEPK/V/2021.

References

[1] American Heart Association. 2018. Retrieved from https://www.heart.org/en/health-topics/congenital-heart-defects/ ventricular -septal-defect-vsd, accessed March 2021.

[2] Dakkak, W., & Oliver, T. *Ventricular septal defect. Treasure Island.* FL: STAT Pearls Publishing. 2019.

[3] Tripathi, R.R. Ventricular Septal Defect Echocardiography Evaluation. *J. Indian Acad Echocardiogr Cardiovasc Imaging*. 2020. 4:260-6. DOI-10.413/jiae.jiae_42_20

[4] Goodfellow, I., Bengio, Y., & Courville, A. *Deep Learning.* Cambridge: MIT Press. 2017

[5] Pujari, P., Karim, M. R., & Sewak, M. *Practical Convolutional Neural Networks.* Birmingham, UK: Packt Publishing. 2018.

[6] Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. *Dive into Deep Learning.* UC Berkeley: Spring. 2020.

[7] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. *You Only Look Once: Unified, Real-Time Object Detection.* 2016. Retrieved from https://arxiv.org/pdf/1506.02640.pdf.

[8] Liu, W, Anguelov D., Erhan, D. *et al.*. SSD: single shot multibox detector, in European Conference on Computer Vision. vol. 9905. *Lecture Notes in Computer Science.* 21–37. 2016,

[9] R. Girshick Fast R-CNN. In *Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV '15).* 2015. 1440–1448. doi: 10.1109/ICCV.2015.169

[10] S. Ren, K. He, R. Girshick, and Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2017. 39(6): 1137–1149, doi: 10.1109/TPAMI.2016.2577031.

[11] He, K., Zhang, X., Ren, S. and Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2015. 37(9): 1904–1916. doi: 10.1109/TPAMI.2015.2389824

[12] Abdurahman, F., Fante, K.A. & Aliy, M. Malaria parasite detection in thick blood smear microscopic images using modified YOLOV3 and YOLOV4 models. *BMC Bioinformatics*. 2021. 22. 112.

[13] Nurmain, S., Rachmatullah. M.N., Sapitri. A.I., Darmawahyuni. A., Jovandy. A., Firdaus. A., Tutuko. B., & Passarella. R. Accurate Detection of Septal Defects With Fetal Ultrasonography Images Using Deep Learning-Based Multiclass Instance Segmentation. In *IEEE Access, vol. 8* 196160-196174. 2020. doi: 10.1109/ACCESS.2020.3034367.

[14] Bridge. C.P., Ioannou. C., Noble. J.A. Automated annotation and quantitative description of ultrasound videos of the fetal heart. *Medical Image Analysis*. 2017. 36:147-161, DOI-10.1016/j.media.2016.11.006.

[15] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. *Yolov4: Optimal speed and accuracy of object detection.* arXiv preprint arXiv:2004.10934. 2020.

[16] Nguyen, N.D., Do, Tien., Ngo, T.D., Le, D.D. An Evaluation of Deep Learning Methods for Small Object Detection. *Journal of Electrical and Computer Engineering.* 2020. ID 3189691. DOI. 10.1155/2020/3189691.

[17] ImageNet. *ImageNet.* 2016. Retrieved from http://www.image-net.org/, accessed March 2021.

[18] Redmon, J. & Farhadi, A. *YOLO v3*. An Incremental Improvement. 2018. arxiv:1804.02767v1.

[19] Mohammed, Bazhdar. *Re: How to calculate Precision and Recall?*. 2018. Retrieved from: https://www.researchgate.net/ accessed April 2021

[20] Hui, J. *mAP (mean Average Precision) for Object Detection.* 2018. Retrieved medium.com accessed April 2021