

Vision Transformer-Based View Classification for Pediatric Cardiac Septal Defect

Anggun Islami^{123*}, Ermatita Ermatita¹, Ria Nova⁴, Ade Iriani Sapitri^{2,3}, Rahmatulloh Rahmatulloh^{2,3}

¹*Faculty of Computer Science, Sriwijaya University, 30139, Palembang, Indonesia*

²*Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia*

³*Artificial Intelligence-Medical Center of Excellence, Universitas Sriwijaya, 30139, Palembang, Indonesia*

⁴*Division of Pediatric Cardiology, Department of Child Health, Mohammad Hoesin General Hospital, Palembang, Indonesia*

**anggunislami2@gmail.com*

ABSTRACT

Congenital Heart Disease (CHDs), particularly Cardiac Septal Defects (CSDs), remains a major cause of morbidity in children. This study proposes a deep learning-based approach using Vision Transformer (ViT) to automatically classify view echocardiographic images as a fundamental step toward improving diagnostic accuracy in pediatric CSDs. A total of 158 videos for the four-chamber view (4CV), five-chamber view (5CV), and subcostal (SC) views; 161 videos for the parasternal long-axis (PLAX) view; and 165 videos for the parasternal short-axis (PSAX) view were used in this study and converted into 48,000 image frames. Four ViT variants (ViT-B/16, ViT-B/32, ViT-L/16, and ViT-L/32) were evaluated under identical experimental settings using 100 training epochs, the AdamW optimizer, and consistent input configurations. Model performance was assessed using accuracy, precision, and recall metrics. Experimental results showed that ViT-L/32 achieved the best performance with an accuracy of 92.20%, accompanied by superior precision and recall compared to other models. The proposed ViT-based framework provided an effective and reliable solution for automatic echocardiographic view classification, supporting early and accurate diagnosis of CSD.

Keywords: Artificial Intelligence, Congenital Heart Disease, Cardiac Septal Defects, Echocardiography, Healthy Lives, SDG3

1. INTRODUCTION

Congenital heart disease (CHDs) is the most common congenital anomaly, affecting 0.8% of the general population and ranking among the leading causes of morbidity and mortality [1][2][3]. These anomalies involve structural abnormalities of the heart, one of which is a cardiac septal defect (CSDs). CSDs occur when there is a hole in the septum separating the heart chambers [4][5][6]. Clinically, the most common CSDs can occur in three main locations: Atrial Septal Defect (ASD); Ventricular Septal Defect (VSD); and Atrioventricular Septal Defect (AVSD) [6]–[9]. CSDs significantly cause abnormal blood flow, such as to the wrong chamber or increased blood flow to the lungs [10][11][12]. Consequently, early detection and continuous monitoring are essential to reduce morbidity rates [13][14][15].

Echocardiography has become the most widely used noninvasive imaging technique for early detection and ongoing monitoring of cardiac structure and function in patients ranging from infants to adults [5][6]. Echocardiography provides visualization through various standard views on CSDs. However, echocardiographic interpretation remains highly operator-dependent and requires substantial expertise [16][17]. The acquisition and interpretation of standard views are often influenced by factors such as patient movement, probe position, image quality, and unstable cardiac motion [18]. Clinically, image quality variability can lead to inconsistencies in identifying standard views, particularly in pediatric cases where anatomical structures are smaller [19][20]. These challenges can result in longer diagnostic times, reduced reproducibility, and an increased risk of misdiagnosis of cardiac abnormalities [20]–[25].

To address these challenges, recent advancements in artificial intelligence (AI), particularly deep learning (DL), have been utilized. DL approaches have been widely applied in cardiac image analysis research, specifically for detecting CSDs. Various architectures have been developed; for instance, the study in [26] used a Temporal–Spatial Decoupled Network (TSDNet), and the study in [27] used VGG16, demonstrating that VSD classification accuracy is stable. However, these approaches focus only on localization in a single CSD condition, using only Computed Tomography (CT) image data. Furthermore, [26] has a high noise level, which has the potential to cause overfitting and reduce generalization ability on new data, and does not yet consider multi-view echocardiography images.

In addition, recent studies in [28][29], using transformer-based approaches such as the Vision Transformer (ViT), have demonstrated stable and promising capabilities in integrating information. Through the self-attention mechanism, the model is capable of capturing information from complex global spatial data. However, studies [28][29] still focus on direct diagnosis without considering initial stages such as view classification in echocardiography, which plays a crucial role in ensuring the quality of subsequent analysis [30]. Furthermore, accurate multi-view identification is a crucial first step in ensuring that further analysis is performed on the correct and clinically relevant views; however, the application of ViT remains relatively limited and has not been extensively explored [31][32].

Therefore, this study proposes a ViT-based approach for performing view classification on pediatric CSDs. This study evaluates several variants of the ViT architecture to identify the model with the best performance in distinguishing standard views. By conducting a systematic evaluation within a single experimental framework, this study aims to provide a comprehensive analysis of the models' ability to understand the spatial representation of echocardiographic images. The main contributions of this study are as follows:

- Proposing a Vision Transformer-based echocardiographic view classification approach as a fundamental step in the cardiac image analysis pipeline, and
- Conducting a comparative evaluation of several ViT architecture variants to determine the best model for supporting further analysis

2. MATERIAL AND METHODS

In this study, four stages were conducted, namely: (i) data preparation; (ii) data preprocessing; (iii) model training for view classification; and (iv) model evaluation to identify the best model. Each stage was designed systematically to ensure optimal data quality and improve the performance of the classification model. Figure 1 shows the workflow used for view classification for CSDs using the ViT.

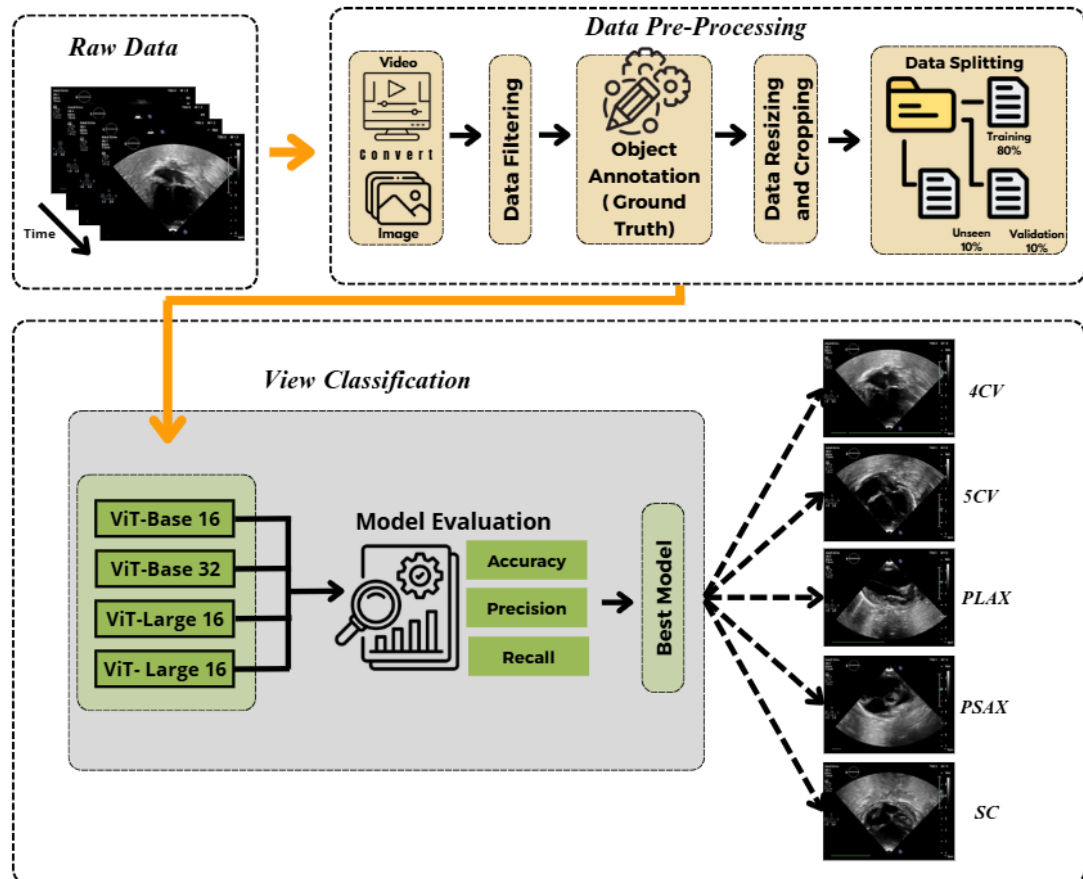


FIGURE 1. View classification pipeline of CSDs using ViT

2.1 DATA PREPARATION

In this stage, the data used in the study was prepared. The data consisted of echocardiography recordings in video format (.AVI) with a native resolution of 800 x 600 pixels and a duration of 1–5 seconds per video (30–60 MB). Echocardiography videos were collected from the Pediatric Cardiology Department of Dr. Mohammad Hoesin General Hospital, Palembang, Indonesia, using a Philips ultrasound machine.

In total, the dataset consisted of 800 echocardiography videos, representing various conditions and anatomical views of the heart. The standard views used were 158 videos for the four-chamber view (4CV), five-chamber view (5CV), and subcostal (SC) views; 161 videos for the parasternal long-axis (PLAX) view; and 165 videos for the parasternal short-axis (PSAX) view. The visual characteristics of each view

are shown in Figure 2, which was used as the feature set for view classification purposes.

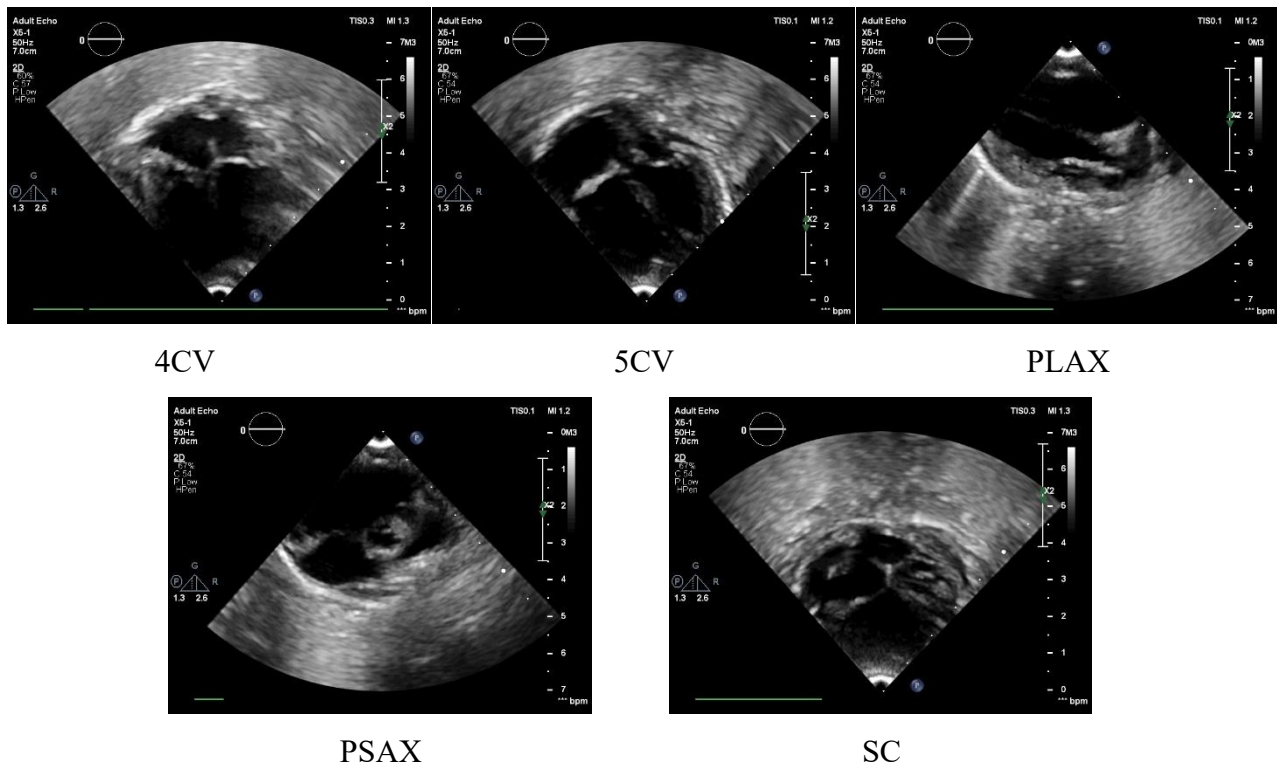


FIGURE 2. Sample image of CSDs per view

2.2 DATA PRE-PROCESSING

Before model training, several pre-processing steps were performed to ensure that the data was clean, uniform, and conforms to the input format required by the model. This stage included converting video data into images, filtering data quality, annotating objects for segmentation, labeling images for classification, transforming annotation formats, resizing images, and splitting the dataset.

First, echocardiography data in video format was converted into frame-by-frame images for image-based processing. Each video consisted of 150–251 frames, and data filtering was performed to remove non-representative frames, such as those with low quality or containing artifacts. Overall, 60 images were extracted per video. Second, labeling was performed by grouping images based on view type. This process yielded a total of approximately 48,000 images with a relatively balanced distribution across classes, thereby supporting more stable and unbiased model performance.

To meet the model training input requirements, all images were resized from 800x600 to 224x224 pixels. Additionally, a cropping process was performed to adjust the aspect ratio to 1:1 without removing important parts of the heart structure. This process was crucial for preserving diagnostic information while ensuring compatibility with the pre-trained model used (as depicted in Figure 3).

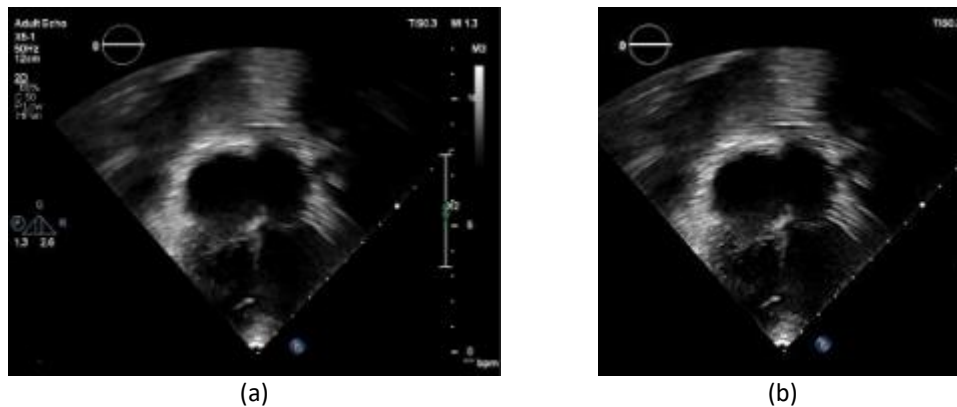


FIGURE 3. Cropping image: (a) before cropping; (b) after cropping

To ensure a reliable and unbiased evaluation, the dataset was divided into three subsets at the video level, such as training, validation, and testing data. The splitting was performed separately for each echocardiographic view while maintaining a balanced distribution of the view classes across the three subsets. Following the video-level splitting, 60 representative frames were extracted from each video according to the data filtering procedure described previously. All frames originating from the same video were therefore assigned exclusively to the subset to which that video belonged. Consequently, no frames from the same video were shared between the training, validation, and testing sets. This process was carried out for each view of every model, with the data distribution shown in Table 1.

TABLE 1.
Data Distribution for Classification After Splitting

View	Video			Frame		
	Train	Validation	Test	Train	Validation	Test
4CV	128	15	15	7680	900	900
5CV	128	15	15	7680	900	900
PLAX	129	16	16	7740	960	960
PSAX	133	16	16	7980	960	960
SC	128	15	15	7680	900	900
Total				38760	4620	4620

2.3 VIEW CLASSIFICATION MODEL TRAINING

During the model training phase, four pre-trained model architectures from the Vision Transformer (ViT) family were used: ViT-Base with a patch size of 16 (ViT-

B/16), ViT-Base with a patch size of 32 (ViT-B/32), ViT -Large with a patch size of 16 (ViT-L/16), and ViT-Large with a patch size of 32 (ViT-L/32). In this study, ViT was a deep learning (DL) architecture designed to perform classification tasks on image data. ViT adapted the Transformer Encoder concept from natural language processing (NLP) to computer vision. The Transformer Encoder featured a Multi-Head Self-Attention mechanism that enabled the model to learn relationships between different patches in an image. As shown in Figure 4, ViT operated by patching the input into N patches, as described in Equation (1).

$$N = \frac{HW}{P^2} \quad (1)$$

where, H is the image height, W is the image width, dan P^2 is the patch size. Next, the model flattened the $H \times W$ patch into a vector, performed embedding using a fully connected layer, added a positional encoder to capture information within each patch, encoded the data, and performed classification [33]

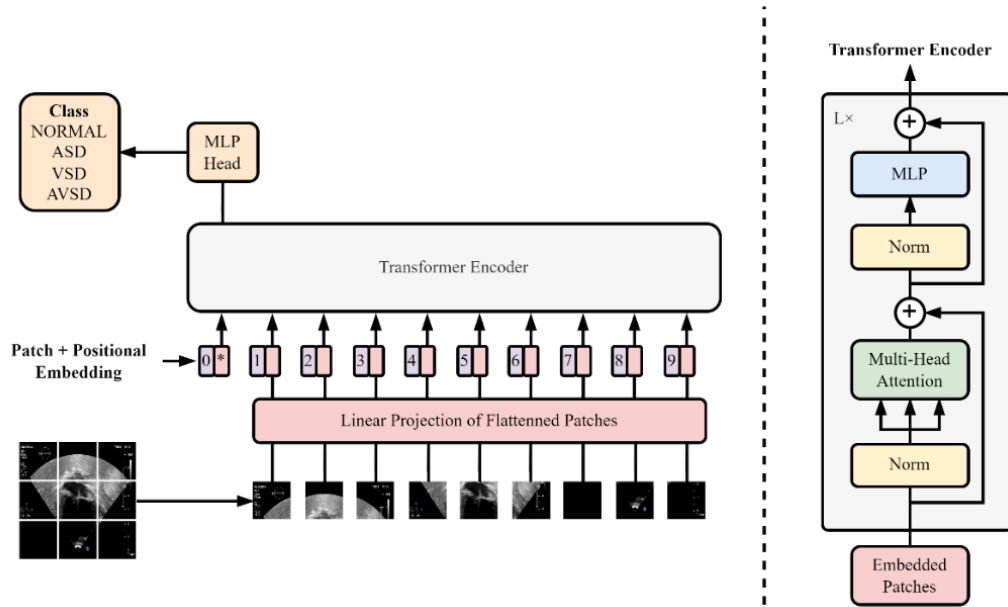


FIGURE 4. ViT Architecture for View Classification of CSDs Dataset

To maintain consistency in the evaluation process across models, each tested ViT architecture was given a hyperparameter configuration. Some of the hyperparameters used included a batch size of 4, a learning rate of 3×10^{-4} , and the use of the Adam with Weight Decay (AdamW) optimizer. Training was conducted over 100 epochs, with the expectation that the model would achieve convergence and learn deep data representations within that range.

The input provided to the model was also adjusted to be compatible with the ViT architecture, specifically with an image resolution of 224×224 pixels and 3 color channels (RGB), resulting in an input dimension of $224 \times 224 \times 3$. This was conducted because ViT processed inputs in the form of small patches divided from the original image, so that an appropriate input size ensured that the number of patches and their representations remained consistent across models

2.4 MODEL EVALUATION

To measure performance, this study used three metrics. Details regarding the definitions, formulas, and calculation methods for these metrics were provided in Equations (2)–(5), as follow

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

where, TP is the number of true positive predictions, TN is the number of true negative predictions, FP is the number of false positive predictions, and FN is the number of false negative predictions

2.5 PLATFORM

All experiments were conducted using an NVIDIA GeForce RTX 2080 GPU (32 GB of shared memory and 8 GB of dedicated memory), running Windows 11 Pro 64-bit and Python 3.9.18. The model implementation utilized PyTorch with support from the NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, and Roboflow libraries.

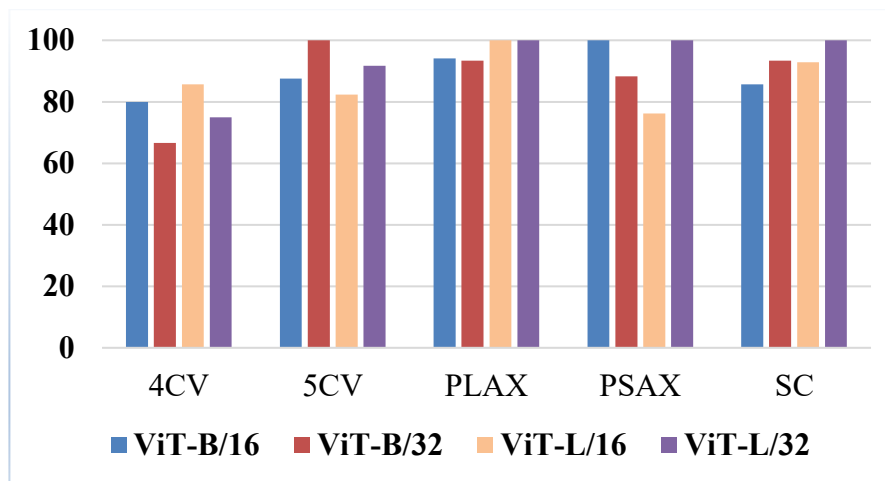
3. RESULTS AND DISCUSSIONS

This section presented the results of view classification experiments on pediatric echocardiography data using four variants of the ViT model: ViT-B/16, ViT-B/32, ViT-L/16, and ViT-L/32. Performance was evaluated based on the metrics of accuracy, precision, and recall. The classification performance results for the CSDs dataset were presented in Table 2. Based on the experiments conducted in this study, the ViT-L/32 model demonstrated the best performance compared to the ViT-B/16, ViT-B/32, and ViT-L/16 models, with an accuracy of 92.20%. Meanwhile, the ViT-B/16 model also demonstrated performance with an accuracy of 89.61%, indicating stable generalization capabilities despite its smaller architecture. However, the ViT-B/32 and ViT-L/16 models showed relatively lower performance, particularly in terms of recall, suggesting limitations in capturing discriminative features among specific classes.

Overall, these results demonstrated that the ViT-L model combined with a patch size of 32x32 enhanced the model's ability to capture global spatial features and was not only accurate in predicting the correct class but also consistent in minimizing both false positives and false negatives. Figure 5 presents a visual representation of the results shown in Table 2.

TABLE 2.
Performance Results of View Classification Using Testing Dataset of CSDs

Model	Class	Performance (%)		
		Precision	Recall	Accuracy
ViT-B/16	4CV	80.00	80.00	89,61
	5CV	87.50	93.33	
	PLAX	94.11	100	
	PSAX	100	93.75	
	SC	85.71	80.00	
ViT-B/32	4CV	66.66	93.33	85.71
	5CV	100	60.00	
	PLAX	93.33	87.50	
	PSAX	88.23	93.75	
	SC	93.33	93.33	
ViT-L/16	4CV	85.71	80.00	85.71
	5CV	82.35	93.33	
	PLAX	100	68.75	
	PSAX	76.19	100	
	SC	92.85	86.66	
ViT-L/32	4CV	75.00	100	92.20
	5CV	91.66	73.33	
	PLAX	100	100	
	PSAX	100	100	
	SC	100	86.66	



(a)

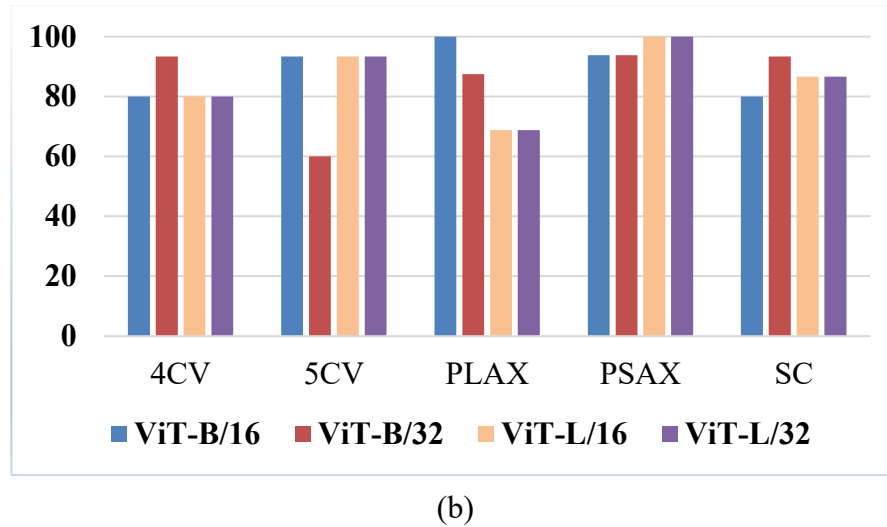


FIGURE 5. Visualization of (a) precision and (b) recall performance analysis results for view classification using the CSDs dataset

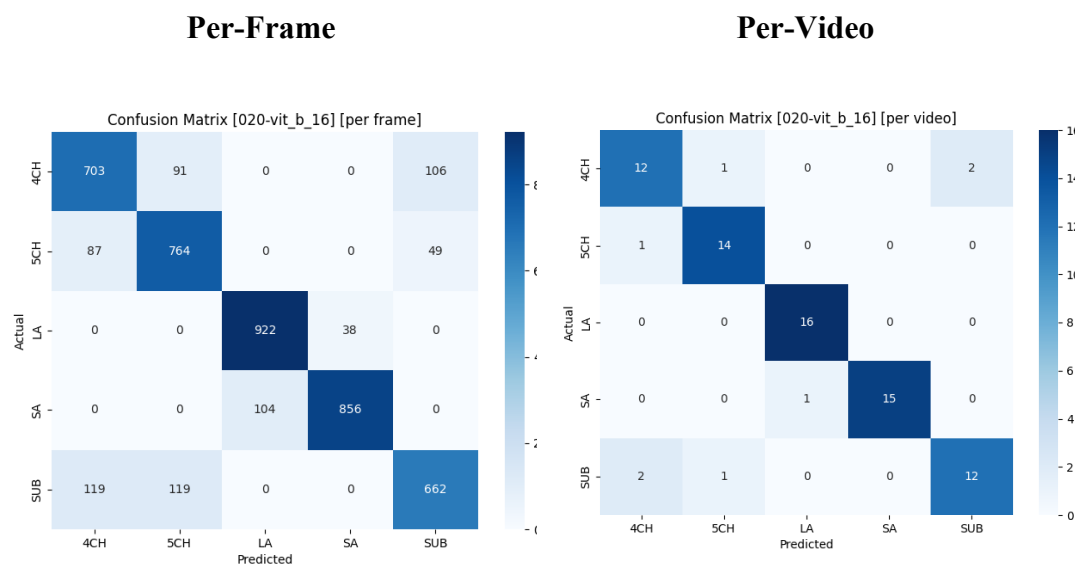
The performance results presented in Table 2 were then summarized by calculating the average performance across all classes to provide a more robust evaluation of each model. Table 3 showed that the ViT-L/32 model achieved the best overall performance with the highest accuracy, precision, and recall, demonstrating its strong ability to generalize across various CSDs viewpoints. Meanwhile, ViT-B/32 showed the lowest average results, particularly in terms of recall, indicating limitations in view classification. These findings highlight that larger model capacity combined with larger patch sizes significantly contributed to improved view classification performance.

Model	Accuracy (%)	Precision (%)	Recall (%)
ViT-B/16	89.61	89.46	89.41
ViT-B/32	85.71	88.31	85.58
ViT-L/16	85.71	87.42	85.75
ViT-L/32	92.20	93.33	91.99

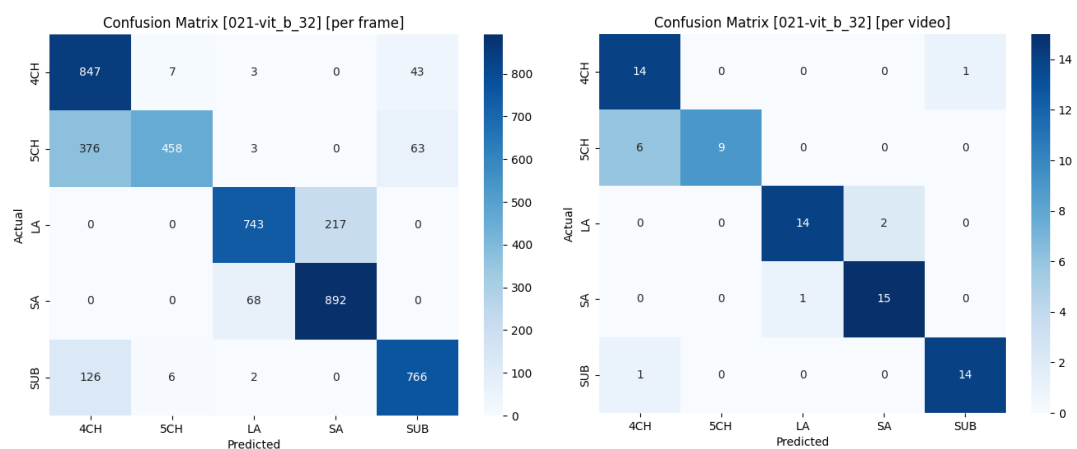
The confusion matrix (CM) heatmaps for four models were presented in Figure 6, illustrating both per-frame and per-video classification performance across performance acrossed the five standard views of CSDs (4CV, 5CV, PLAX, PSAX, and SC). Overall, the results showed a consistent pattern where LA and SA classes achieve the highest classification accuracy, as indicated by strong diagonal dominance in the confusion matrices across all models. In particular, the ViT-L/32 model demonstrated the most concentrated diagonal distribution, correctly classifying the majority of samples, such as 901 frames for LA and 908 frames for SA, with minimal classification errors.

However, significant confusion was observed between the 4CH and 5CH classes, particularly in per-frame evaluation, where a large number of samples were classification errors between these two categories (e.g., 259 5CH samples were predicted as 4CH on ViT-L/32). This behavior indicated a high degree of structural similarity between the two views, making them difficult to distinguish even by advanced models. Additionally, the SUB class exhibited moderate confusion, often being misclassified as 4CH or 5CH, indicating the presence of overlapping visual features.

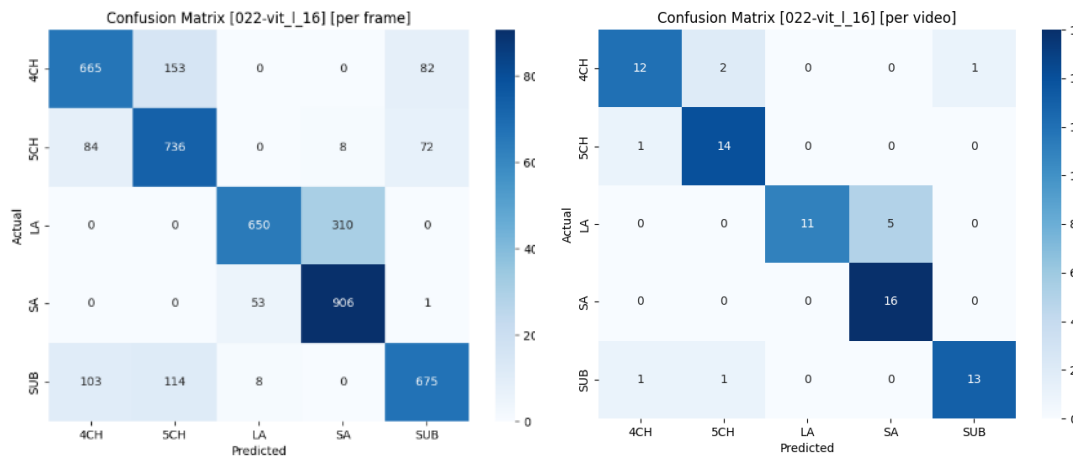
When comparing per-frame and per-video results, the per-video confusion matrix showed significantly better performance, with fewer classification errors. This indicated that temporal aggregation across frames enhances classification robustness. Overall, among all evaluated models, ViT-L/32 achieved the best performance, with the lowest inter-class confusion and the most stable predictions, confirming its superiority in capturing global contextual features for view classification of CSDs.



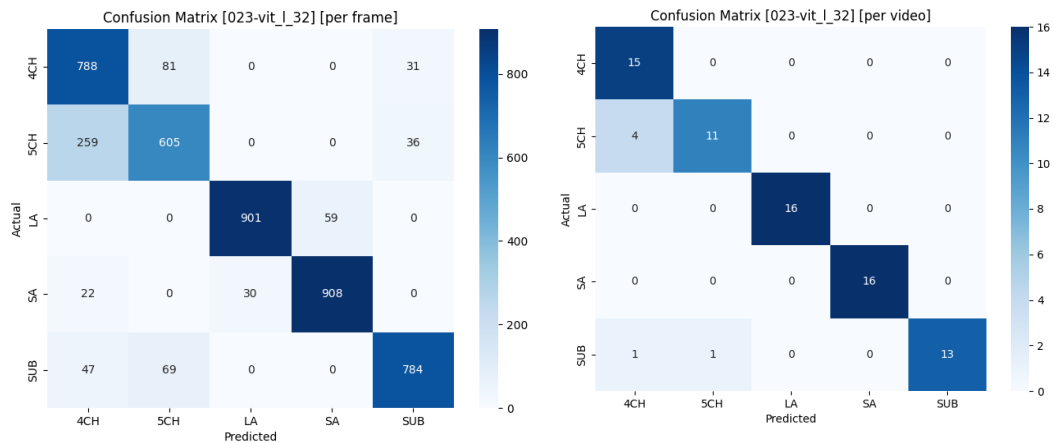
(a) ViT-B/16



(b) ViT-B/32



(c) ViT-L/16



(d) ViT-L/32

FIGURE 6. The CM heatmap of ViT-B/16, ViT-B/32, ViT-L/16, and ViT-L/32 model for CSDs view

3. CONCLUSION

This study developed and proposed a ViT-based approach for view classification of CSDs in children as a fundamental step in the CHD analysis pipeline. Using an echocardiography dataset, this study evaluated four ViT variants: ViT-B/16, ViT-B/32, ViT-L/16, and ViT-L/32. Experimental results showed that the ViT-L/32 model in view classification (4CV, 5CV, PLAX, PSAX, SC) delivered the best performance with an accuracy of 92.20%, as well as the highest precision and recall values compared to other models. Furthermore, this contribution was also relevant to global efforts to improve child health, in line with Sustainable Development Goal (SDG) 3. Moving forward, research can focus on developing end-to-end systems that integrate view classification with the detection and segmentation of cardiac abnormalities, the

utilization of more advanced transformer architectures, and clinical validation on a larger scale to ensure readiness for implementation in the real medical world.

ACKNOWLEDGEMENTS

This research was supported by the Intelligent System Research Group (IsysRG) and the Artificial Intelligence–Medical Center of Excellence (AIMed CoE) of Universitas Sriwijaya. The authors also express their sincere appreciation to Dr. Mohammad Hoesin Central General Hospital, Palembang, Indonesia, for their valuable contributions and support throughout this study.

AUTHOR’S CONTRIBUTIONS

AI: Wrote the manuscript, performed the analysis, analysis tools, and prepared the visualization. EE: formal analysis, wrote and edited the manuscript. RN: formal analysis dataset. AIS and RR: Methodology analysis.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

CONSENT FOR PUBLICATION

Not applicable

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest to report regarding the present study

REFERENCES

- [1] A. Egbe, S. Lee, D. Ho, S. Uppu, and S. Srivastava, “Prevalence of congenital anomalies in newborns with congenital heart disease diagnosis,” *Ann. Pediatr. Cardiol.*, vol. 7, no. 2, pp. 86–91, 2014, doi: 10.4103/0974-2069.132474.
- [2] Y. Liu *et al.*, “Global birth prevalence of congenital heart defects 1970-2017: updated systematic review and meta-analysis of 260 studies,” *Int. J. Epidemiol.*, vol. 48, no. 2, pp. 455–463, Apr. 2019, doi: 10.1093/ije/dyz009.
- [3] A. Marelli *et al.*, “Canadian Cardiovascular Society 2022 Guidelines for Cardiovascular Interventions in Adults With Congenital Heart Disease,” *Can. J. Cardiol.*, vol. 38, no. 7, pp. 862–896, 2022, doi: 10.1016/j.cjca.2022.03.021.
- [4] S. H. Chen *et al.*, “Optimizing Object Detection Algorithms for Congenital Heart Diseases in Echocardiography: Exploring Bounding Box Sizes and Data Augmentation Techniques,” *Rev. Cardiovasc. Med.*, vol. 25, no. 9, 2024, doi: 10.31083/j.rcm2509335.
- [5] Z. Jalal *et al.*, “Long-term Complications After Transcatheter Atrial Septal Defect Closure: A Review of the Medical Literature,” *Can. J. Cardiol.*, vol.

- 32, no. 11, pp. 1315.e11-1315.e18, Nov. 2016, doi: 10.1016/j.cjca.2016.02.068.
- [6] Y. LIU *et al.*, “Intelligent diagnosis of atrial septal defect in children using echocardiography with deep learning,” *Virtual Real. Intell. Hardw.*, vol. 6, no. 3, pp. 217–225, 2024, doi: 10.1016/j.vrih.2023.05.002.
- [7] R. Nova, S. Nurmaini, R. U. Partan, and S. T. Putra, “Automated image segmentation for cardiac septal defects based on contour region with convolutional neural networks: A preliminary study,” *Informatics Med. Unlocked*, vol. 24, p. 100601, 2021, doi: 10.1016/j.imu.2021.100601.
- [8] K. Miura *et al.*, “Deep learning-based model detects atrial septal defects from electrocardiography: a cross-sectional multicenter hospital-based study,” *eClinicalMedicine*, vol. 63, p. 102141, 2023, doi: 10.1016/j.eclinm.2023.102141.
- [9] A. Dozen *et al.*, “Image segmentation of the ventricular septum in fetal cardiac ultrasound videos based on deep learning using time-series information,” *Biomolecules*, vol. 10, no. 11, pp. 1–17, 2020, doi: 10.3390/biom10111526.
- [10] U. Doshi and E. Wang-Giuffre, “Ventricular Septal Defects: A Review,” in *Congenital Heart Defects*, P. S. Rao, Ed. Rijeka: IntechOpen, 2022. doi: 10.5772/intechopen.104809.
- [11] I. K. Owusu, G. M. Amponsah, and Y. A. Wiafe, “Severe Pulmonary Hypertension From Combined Ventricular Septal Defect and Rheumatic Mitral Valve Disease: A Case Report and Literature Review,” *Clin. Med. Insights Case Reports*, vol. 17, 2024, doi: 10.1177/11795476241231140.
- [12] W. Zhang, H. Y. Xu, Y. C. Zhang, and K. B. Liu, “Delayed diagnosis of critical congenital heart defects predicting risk factors and survival rate in newborns in Beijing: a retrospective study,” *J. Int. Med. Res.*, vol. 49, no. 7, 2021, doi: 10.1177/03000605211028028.
- [13] A. A. Karatza *et al.*, “Missed or Delayed Diagnosis of Heart Disease by the General Pediatrician,” 2024.
- [14] Z. Eyileten and A. Uysalel, “Isolated ventricular septal defect in infants,” *E J. Cardiovasc. Med.*, vol. 5, no. 2, pp. 27–33, 2017, doi: 10.15511/ejcm.17.00227.
- [15] R. Vullings, “Fetal Electrocardiography and Deep Learning for Prenatal Detection of Congenital Heart Disease,” *2019 Comput. Cardiol. Conf.*, vol. 45, pp. 1–4, 2019, doi: 10.22489/cinc.2019.072.
- [16] A. Iriani Sapitri *et al.*, “Deep learning-based real time detection for cardiac objects with fetal ultrasound video,” *Informatics Med. Unlocked*, vol. 36, no. November 2022, p. 101150, 2023, doi: 10.1016/j.imu.2022.101150.
- [17] J. A. Slivnick *et al.*, “Echocardiographic Detection of Regional Wall Motion Abnormalities Using Artificial Intelligence Compared to Human Readers,” *J. Am. Soc. Echocardiogr.*, vol. 37, no. 7, pp. 655–663, 2024, doi: 10.1016/j.echo.2024.03.017.
- [18] M. G. Elhassan, S. Grewal, and N. Nezarat, “Point-of-Care Ultrasonography in Internal Medicine: Limitations and Pitfalls for Novice Users,” *Cureus*, vol. 15, no. 8, 2023, doi: 10.7759/cureus.43655.
- [19] Y. Deng and K. Pushparajah, “Evaluation of Congenital Heart Disease using Machine Learning,” 2025.
- [20] B. Mohammad Nijres and A. M. Qureshi, “Pediatric Diagnostic Cardiac

- Catheterization,” in *Pediatric Cardiology: Fetal, Pediatric, and Adult Congenital Heart Diseases*, Springer, 2024, pp. 855–901.
- [21] A. Hauptmann, S. Arridge, F. Lucka, V. Muthurangu, and J. A. Steeden, “Real-time cardiovascular MR with spatio-temporal artifact suppression using deep learning—proof of concept in congenital heart disease,” *Magn. Reson. Med.*, vol. 81, no. 2, pp. 1143–1156, 2019.
 - [22] Q. Qi, X. Han, and Y. Liu, “Application status and prospect of deep learning in echocardiography,” *IEEE Access*, vol. 12, pp. 83405–83416, 2024.
 - [23] F. Viñals, P. Poblete, and A. Giuliano, “Spatio-temporal image correlation (STIC): A new tool for the prenatal screening of congenital heart defects,” *Ultrasound Obstet. Gynecol.*, vol. 22, no. 4, pp. 388–394, 2003, doi: 10.1002/uog.883.
 - [24] J. Wang *et al.*, “Automated interpretation of congenital heart disease from multi-view echocardiograms,” *Med. Image Anal.*, vol. 69, p. 101942, 2021.
 - [25] Y. S. Yoo, S. H. Lee, and S. H. Bae, “Effective Multi-Object Tracking via Global Object Models and Object Constraint Learning,” *Sensors (Basel)*, vol. 22, no. 20, 2022, doi: 10.3390/s22207943.
 - [26] Y. Liu *et al.*, “A Dual-Center Study on Intelligent Detection of Ventricular Septal Defect in Pediatric Echocardiographic Videos Using the TSDNet Model,” *Int. J. Imaging Syst. Technol.*, vol. 36, no. 2, p. e70304, 2026.
 - [27] O. Access, “Classification of Ventricular Septal Defect Disease Using Deep,” vol. 9, no. 1, pp. 34–54, 2025.
 - [28] A. Amangeldi, “Congenital Heart Disease recognition using Deep Learning / Transformer models,” 2025.
 - [29] B. Liu *et al.*, “A deep learning framework assisted echocardiography with diagnosis, lesion localization, phenogrouping heterogeneous disease, and anomaly detection,” *Sci. Rep.*, vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-022-27211-w.
 - [30] Q. Yang *et al.*, “A Deep Learning Framework for Fetal Heart Tracking in Ultrasound Videos: Toward Enhanced Congenital Heart Defects Detection,” in *International Conference on Functional Imaging and Modeling of the Heart*, 2025, pp. 219–230.
 - [31] Y. Liu *et al.*, “Atrial Septal Defect Detection in Children Based on Ultrasound Video Using Multiple Instances Learning,” *J. Imaging Informatics Med.*, vol. 37, no. 3, pp. 965–975, 2024, doi: 10.1007/s10278-024-00987-1.
 - [32] Y. Yang *et al.*, “Latent Motion Profiling for Annotation-free Cardiac Phase Detection in Adult and Fetal Echocardiography Videos,” 2025, [Online]. Available: <http://arxiv.org/abs/2507.05154>
 - [33] I. Fayad, P. Ciais, M. Schwartz, J. Wigneron, A. De Truchis, and F. Frappart, “Vision Transformers , a new approach for high-resolution and large- scale mapping of canopy heights,” pp. 1–42.