

CardioViT-CAM: Self-Supervised Vision Transformer with Class Activation Mapping for Interpretable Congenital Heart Disease Detection

Kiramath Ullah¹, Jabar Ali², Hee Cheol Kim³

¹Department of Computer Engineering, Inje University, Gimhae, South Korea.

²Assistant Professor, Lady Reading Hospital, Peshawar, Pakistan.

³Digital Anti-Aging Healthcare, College of AI Convergence, u-AHRC, Inje University.

kiramathullah541@gmail.com, jabbar.ali@lrh.edu.pk, heeki@inje.ac.kr

Abstract—Congenital heart disease (CHD) persists as a dominant source of infant disease and demise. Although transthoracic echocardiography (TTE) is essential for screening CHD, expert-level recognition remains limited in many clinical settings. We recommend CardioViT-CAM, a two-stage learning framework that integrates a self-supervised learning method, Mask Autoencoder (MAE) pre-training for feature representation, and supervised vision transformer (ViT-B/16) fine-tuning for automated CHD classification. Our approach leverages 74,000 unlabelled echocardiographic frames for pretraining and 6,000 labelled images for three-class classification (ASD, VSD, AVSD). Grad-Cam++ visualization provides understandable heatmaps that emphasize defects of diagnostic cardiac regions. CardioViT-CAM achieved 97.03% accuracy, 97.04% precision, 97.03% recall, and 97.03% F1-Score, representing strong diagnostic performance and interpretability for clinical decision support.

Keywords— Computer Vision, Self-Supervised Learning, Vision Transformer, Congenital Heart Disease, Grad-Cam++, Echocardiography.

I. INTRODUCTION

Congenital heart disease (CHD) remains a substantial worldwide health burden, affecting millions of newborns each year. It is the most common and frequently severe birth anomaly globally, with a prevalence of 7 to 12 in 1000 newborn babies (1, 2, 3 and can take a heavy health and financial load on patients. Universally, approximately 1.35 million children are born with CHD every year. CHDs are one of the world's most common congenital anomalies and, thus, a substantial international health issue. CHD occurrence is higher in Asia at 9.3 per 10,000 live births, compared with Europe at 8.2 and Africa at 1.9 (4). The updated diagnoses, treatment, and most common types of CHD include Atrial Septal Defects (ASD), Ventricular Septal Defects (VSD), and Atrioventricular Septal Defects (AVSD) (5). Two-dimensional (2D) and colour Doppler ultrasound methods provide individual echocardiographic information essential for understanding the positions of cardiac structures and the blood flow dynamics and function in a CHD. CHD has visible, significant features in both 2D and Doppler TTEs, easing the

diagnosis of CHD. However, for experienced cardiac sonographers, the recognition of ASD, VSD, and AVSD can be challenging to diagnose and may result in missing the first screening diagnosis. Recent research has proven the capability of computer vision procedures, such as Convolutional Neural Networks (CNNs), to advance CHD analysis and forecasting (6). Deep Learning (DL) is one of the most innovative technologies in Artificial Intelligence. DL algorithms running on high-performance computing systems deliver robust results in medical image analysis, such as segmentation and classification (7). Using AI to detect CHD presents challenges, such as the requirement for a huge amount of high-quality data and the complex validation of algorithms; however, AI is likely to be useful for improving the diagnosis, and representing the issue success in a clinical background, enhancing the usefulness of routine screening (8, 9, 10) to accurately classify ASD 2 closure and suggest treatment strategies, guide cardiologists in making decisions (11).

To address these challenges, the study proposes **CardioViT-CAM**, a novel two-stage deep learning pipeline for explainable CHD classification.

- Pretrain a Vision Transformer (ViT) applying Masked Autoencoder (MAE) on unlabeled echocardiographic frames to discover common cardiac feature representations.
- Fine-tune the pretrained ViT for supervised three-class CHD classification (ASD, VSD, AVSD).
- Apply Grad-CAM++ to generate class-specific heatmaps that localise and visualise diagnostic-relevant cardiac regions.

II. RELATED WORKS

This section reviews ongoing work in CHD recognition and summarizes related research in medical imaging and artificial intelligence. Liu et al. (12) proposed several instance-learning models for detecting only one class of atrial septal defects (ASD) in children's ultrasound videos, achieving an accuracy of 83.04%. At the same time, Wang et al. (13) developed a deep learning framework for automated clarification of CHD from an echocardiogram, achieving an accuracy of 92.3% for three-class classification: Atrial Septal Defects (ASD),

Ventricular Septal Defects (VSD), and normal. Hong et al. (14) developed a convolutional neural network model for secundum atrial septal defects in children, with a color echocardiography accuracy of 91.8%. Jiang et al. (15) introduced a deep learning-based method that employs seven standard pediatric echocardiographic views to detect CHD, accomplishing an accuracy of 92.3%. Day et al. (16) demonstrated that integrating AI into clinicians' echocardiographic ultrasound images could improve diagnostic accuracy for AVSD. They found that providing AI ideas significantly improved overall accuracy to 86.5%, including by adding a model confidence score and Grad-CAM. Szijarto et al. (17) suggested a masked autoencoder customized for medical ultrasound videos, focusing on clinically related regions, to improve pre-training efficiency. Dosovitskiy et al. (18) introduced Vision Transformer (ViT) and demonstrated that transformers can surpass CNNs in image identification. ViTs outperform by highlighting popular techniques such as attention maps and Grad-CAM (19), and Matsoukas et al. (20) questioned whether transformers can replace CNNs for medical image analysis, making them a strong option for medical imaging with partial-label data. Overall, these studies emphasize that, compared to conventional methodologies, the growing shift towards transformer-based and self-supervised learning frameworks is more competent in limited-data and poor-generalization cases for medical imaging tasks.

III. DATASET

In this study, we used a private transthoracic echocardiography (TTE) video dataset collected at Lady Reading Hospital, Pakistan, between 2022 and 2024, concentrating on three congenital heart disease (CHD) categories: Atrial Septal Defect (ASD), Ventricular Septal Defect (VSD), and Atrioventricular Septal Defect (AVSD). The dataset contains 1600 TTE videos, and we extracted 74,000 image frames. Two senior cardiologists labelled balanced subsets of 6000 images (2000 per class) for downstream supervised fine-tuning classification evaluation. Ethical agreement was taken from the Institutional Review Board, and all images were anonymized.

IV. METHODOLOGY

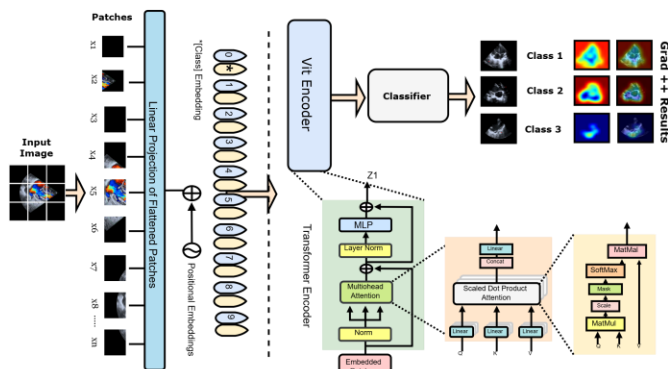


Figure 1. Overview of the methodology.

The suggested **CardioViT-CAM** framework consists of two main stages: a self-supervised pre-training using MAE and a supervised fine-tuning with a Vision Transformer (ViT-B/16) integrated with Grad-CAM++ for interpretability.

A. Self-Supervised Pre-training via Masked Autoencoder (MAE)

In the first stage, we employ a Vision Transformer-based Masked Autoencoder (MAE) for self-supervised pre-training on unlabelled echocardiographic images to learn generalized cardiac representations. The architecture consists of an asymmetric encoder-decoder structure. The ViT-Base encoder processes input images $x \in \mathbb{R}^{\wedge} (224 \times 224 \times 3)$ by splitting them into 196 non-overlapping 16×16 patches, each embedded into a 768-dimensional space. The encoder includes 12 transformer blocks with 12-head self-attention. The lightweight decoder (8 blocks, 512-dim, 16 heads) reconstructs the masked patches. We randomly mask 75% of patches and reconstruct them via a normalized MSE loss. Training uses AdamW (learning rate = 1.5×10^{-4} , weight decay = 0.05) with cosine annealing (40 warmup epochs, 600 total), gradient clipping (norm = 1.0), and medical image augmentations (random cropping, flips, contrast adjustment, Gaussian blur).

B. Fine-Tuning for Classification

In the second stage, the proposed **CardioViT-CAM** architecture (Figure 1) performs supervised fine-tuning of a ViT-B/16 with Grad-CAM++ for explainable CHD classification (ASD, VSD, AVSD). The pre-trained encoder is linked with a three-layer classifier: $768 \rightarrow 1024 \rightarrow 512 \rightarrow 3$, using LayerNorm, GELU activations, and progressive dropout ($0.3 \rightarrow 0.25 \rightarrow 0.2$). The model predicts. We use AdamW with differential learning rates (encoder learning rate = 1.5×10^{-5} , classifier learning rate = 1.5×10^{-4}), weight decay = 0.01, cosine scheduling, and early stopping (patience = 25). Data augmentation includes geometric transforms (flips, $\pm 15^\circ$ rotation) and color jitter. The dataset is split as 72% training, 8% validation, and 20% testing. For interpretability, we apply **Grad-CAM++**.

V. RESULTS

A. Comparison of the Visualization Methods

Table 1 represents a quantitative evaluation of gradient-based class activation mapping methods on the CHD dataset. Each model uses the identical ViT-B/16 architecture with MAE pretraining, differing only in the visualization method. Among the evaluated approaches, **Grad-CAM++** achieves the best performance across all metrics.

TABLE 1. COMPARISON OF VISUALIZATION METHODS ON THE CHD

	Acc.	Prec.	Rec..	F1 (%)
Grad-CAM	94.32	94.99	94.93	94.92
Score-CAM	96.15	96.18	96.15	96.15
Grad-Cam++	97.03	97.04	97.03	97.03

B. Performance Metrics

Figure 2 presents a bar chart comparing three Grad-CAM visualization methods (Grad-CAM, Score-CAM, Grad-CAM++) for CHD classification across four metrics. **Grad-CAM++** surpasses all methods and achieves 97.03% accuracy, 97.04% precision, 97.03% recall, and 97.03% F1-score.

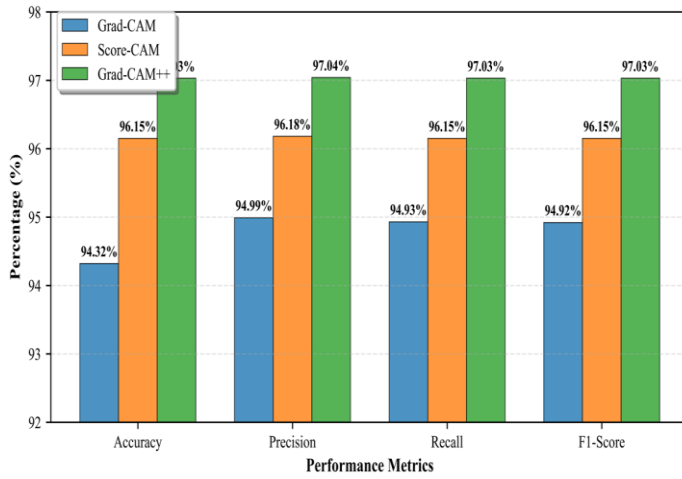


Figure 2. Comparison of Grad-CAM, and Score CAM Grad-CAM++, visualization methods for CHD classification across four evaluation metrics. Grad-CAM++ achieves the highest accuracy, precision, recall, and F1-score.

C. Confusion Matrix

Figure 3 shows the model's performance in classifying CHD with an overall accuracy of 97.03% across the three classes: ASD, VSD, and AVSD. True positives are high for ASD (428), VSD (399), and AVSD (283), while misclassification is minimal; for example, 16 VSD cases predicted as AVSD and 14 AVSD cases predicted as VSD.

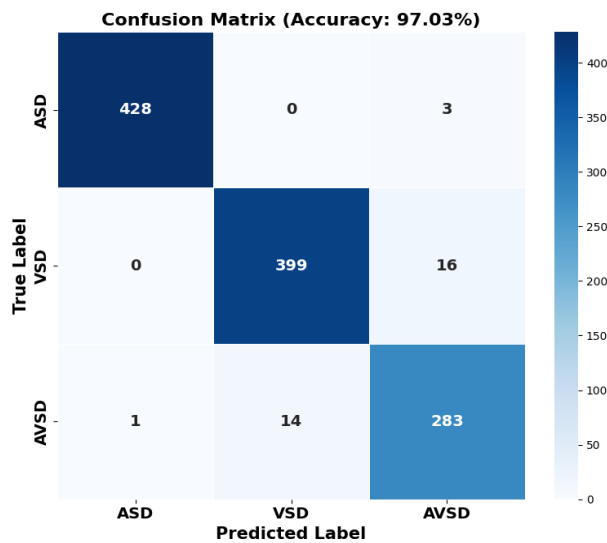


Figure 3. Confusion matrix of CHD classification using CardioViT-CAM showing strong diagonal dominance with minimal misclassification among ASD, VSD, and AVSD.

D. Interpretability of the Model Using Grad-CAM++ Visualization

The **Grad-CAM++** visualization highlights the most influential regions contributing to model predictions across distinct congenital heart defect classes. Warmer-colored regions correspond to higher model attention, indicating areas of diagnostic importance. This visualization approach enhances interpretability and demonstrates that the network focuses on clinically relevant cardiac structures during CHD classification.

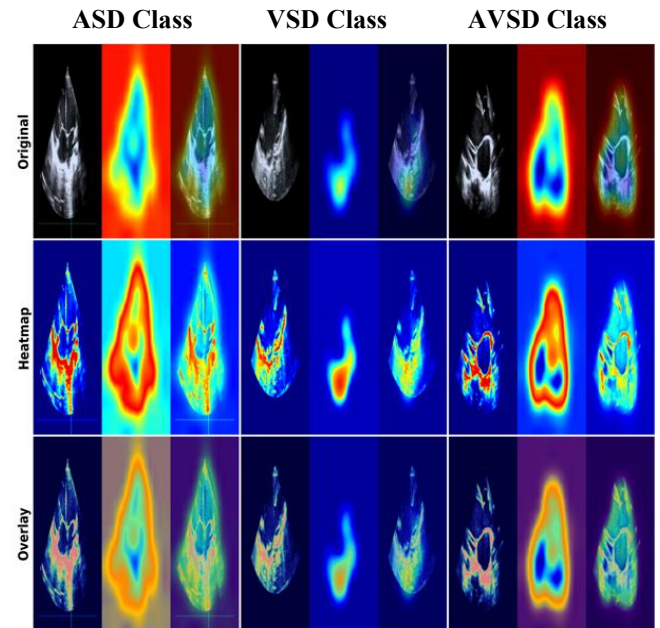


Figure 4. Grad-CAM++ visualization shows discriminative regions for ASD, VSD, and AVSD classification. Warmer colors denote areas of stronger model attention, confirming clinical relevance of detected cardiac features.

TABLE 2. COMPARISON OF VISUALIZATION METHODS ON THE CHD

Literature	Method	Classes	Dataset	Acc. %
Wang et al. [13]	CNN	ASD, VSD, Normal	4.5K Images	92.3
Hong et al. [14]	CNN	ASD only	3.2K Images	91.8
Jiang et al. [15]	DL-7 view Fusion	Multiple CHD	5.0K Images	92.3
Day et al. [16]	CNN + Clinician	AVSD	2.8K	86.5
CardioViT-CAM (Ours)	MAE + ViT + Grad-CAM++	ASD, VSD, AVSD	74K (pre) + 6K (fine-tuning)	97.03

E. Discussion and Comparison with Existing Models

Table 2 presents the comparative analysis of CHD classification models. In contrast to previous studies that focus on a single class or a limited number of defect types, the proposed **MAE + ViT + Grad-CAM++** framework achieves superior accuracy across multiple CHD categories (i.e., ASD, VSD, and AVSD). Leveraging large-scale pretraining and domain-specific fine-tuning enhances feature representation and diagnostic precision. The accuracy achieved of **97.03%** marks a significant improvement over existing CNN-based methodology.

VI. CONCLUSION

This study presents **CardioViT-CAM**, a two-stage learning framework that integrates MAE-based self-supervised pretraining, ViT-B/16 fine-tuning, and Grad-CAM++ visualization for interpretable CHD classification. Trained on echocardiographic data, the proposed model achieves an accuracy of **97.03%**, surpassing existing CNN-based methods. The Grad-CAM++ heatmaps enhance clinical interpretability by accurately highlighting diagnostic relevant cardiac regions associated with congenital defects. In future work, this framework will be extended to localize specific defect areas and quantitatively measure their corresponding sizes for improved diagnostic support.

ACKNOWLEDGMENT

This research was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT) under the Innovative Human Resource Development for Local Intellectualization program (IITP-2024-RS-2024-00436773).

REFERENCES

- [1] Julien IE Hoffman and Samuel Kaplan, "The incidence of congenital heart disease," *Journal of the American college of cardiology*, vol. 39, no. 12, pp. 1890–1900, 2002.J.
- [2] Elisabeth Leirgul, Tatiana Fomina, Kristoffer Brodwall, Gottfried Greve, Henrik Holmström, Stein Emil Vollset, Grethe S Tell, and Nina Øyen, "Birth prevalence of congenital heart defects in Norway 1994–2009—a nation wide study," *American heart journal*, vol. 168, no. 6, pp. 956–964, 2014.
- [3] Yingjuan Liu, Sen Chen, Liesl Z'uhlke, Graeme C Black, Mun-kit Choy, Ningxiu Li, and Bernard D Keavney, "Global birth prevalence of congenital heart defects 1970–2017: updated systematic review and meta-analysis of 260 studies," *International journal of epidemiology*, vol. 48, no. 2, pp. 455–463, 2019.
- [4] Samaa Akhtar and Assma Asif, "Covid-19 in children with congenital heart disease in asia: clinical features and recommendations for management," *International Journal of Contemporary Pediatrics*, vol. 9, no. 4, pp. 408, 2022.
- [5] Xiao Meng, Ming Song, Kai Zhang, Weida Lu, Yunyi Li, Cheng Zhang, and Yun Zhang, "Congenital heart disease: types, pathophysiology, diagnosis, and treatment options," *MedComm*, vol. 5, no. 7, pp. e631, 2024.
- [6] Khalil Khan, Farhan Ullah, Ikram Syed, and Irfan Ullah, "The role of machine learning in congenital heart disease diagnosis: Datasets, algorithms, and insights," *arXiv preprint arXiv:2501.04493*, 2025.
- [7] Nafiseh Ghaffar Nia, Erkan Kaplanoglu, and Ahad Nasab, "Evaluation of artificial intelligence techniques in disease diagnosis and prediction," *Discover Artificial Intelligence*, vol. 3, no. 1, pp. 5, 2023.
- [8] Haris Ejaz, Tarannum Thyib, Ahmed Ibrahim, Aroob Nishat, and Jhancy Malay, "Role of artificial intelligence in early detection of congenital heart diseases in neonates," *Frontiers in Digital Health*, vol. 5, pp. 1345814, 2024.
- [9] Haris Ejaz, Tarannum Thyib, Ahmed Ibrahim, Aroob Nishat, and Jhancy Malay, "Role of artificial intelligence in early detection of congenital heart diseases in neonates," *Frontiers in Digital Health*, vol. 5, pp. 1345814, 2024.
- [10] Xusheng Jiang, Jin Yu, Jingjing Ye, Weijie Jia, Weize Xu, and Qiang Shu, "A deep learning-based method for pediatric congenital heart disease detection with seven standard views in echocardiography," *World Journal of Pediatric Surgery*, vol. 6, no. 3, pp. e000580, 2023.
- [11] Siti Nurmaini, Radiyati Umi Partan, Nuswil Bernolian, Ade Iriani Sapitri, Bambang Tutuko, Muhammad Nau fal Rachmatullah, Annisa Darmawahyuni, Firdaus Firdaus, and Johannes C Mose, "Deep learning for improving the effectiveness of routine prenatal screening for major congenital heart diseases," *Journal of clinical medicine*, vol. 11, no. 21, pp. 6454, 2022.
- [12] Bohan Liu, Hao Chang, Dong Yang, Feifei Yang, Qiushuang Wang, Yujiao Deng, Lijun Li, Wenqing Lv, Bo Zhang, Liheng Yu, et al., "A deep learning frame work assisted echocardiography with diagnosis, lesion localization, phenogrouping heterogeneous disease, and anomaly detection," *Scientific Reports*, vol. 13, no. 1, pp. 3, 2023.
- [13] Yiman Liu, Qiming Huang, Xiaoxiang Han, Tongtong Liang, Zhifang Zhang, Xiuli Lu, Bin Dong, Jiajun Yuan, Yan Wang, Menghan Hu, et al., "Atrial septal defect detection in children based on ultrasound video using multiple instances learning," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 3, pp. 965–975, 2024.
- [14] WenjingHong, QiuyangSheng, BinDong, LanpingWu, Lijun Chen, Leisheng Zhao, Yiqing Liu, Junxue Zhu, Yiman Liu, Yixin Xie, et al., "Automatic detection of secundum atrial septal defect in children based on color doppler echocardiographic images using convolutional neural networks," *Frontiers in Cardiovascular Medicine*, vol. 9, pp. 834285, 2022.
- [15] Xusheng Jiang, Jin Yu, Jingjing Ye, Weijie Jia, Weize Xu, and Qiang Shu, "A deep learning-based method for pediatric congenital heart disease detection with seven standard views in echocardiography," *World Journal of Pediatric Surgery*, vol. 6, no. 3, pp. e000580, 2023.
- [16] TG Day, J Matthew, SF Budd, L Venturini, R Wright, A Farruggia, TV Vigneswaran, V Zidere, JV Hajnal, RRazavi, et al., "Interaction between clinicians and artificial intelligence to detect fetal atrioventricular septal defects on ultrasound: how can we optimize collaborative performance?," *Ultrasound in Obstetrics & Gynecology*, vol. 64, no. 1, pp. 28–35, 2024.
- [17] 'Ad'am Szij'art'o, B'alint Magyar, Thomas 'A Szeier, M'at'e Tolvaj, Alexandra F'abi'an, B'alint K Lakatos, Zsuzsanna Lad'anyi, Zsolt Bagyura, B'ela Merkely, Attila Kov'acs, et al., "Masked autoencoders for medical ultrasound videos using roi-aware masking," in *International Workshop on Advances in Simplifying Medical Ultrasound*. Springer, 2024, pp. 167–176.
- [18] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [19] Piotr Komorowski, Hubert Baniecki, and Przemyslaw Biecek, "Towards evaluating explanations of vision transformers for medical imaging," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 3726–3732.
- [20] Christos Matsoukas, Johan Fredin Haslum, Magnus S'oderberg, and Kevin Smith, "Is it time to replace cnns with transformers for medical images?," *arXiv preprint arXiv:2108.09038*, 2021.



KIRAMAT ULLAH is an accomplished professional with a solid foundation in software engineering and a robust background in academia. Master's student in Computer Engineering at Inje University, South Korea. He holds a BSc in Software Engineering from BUITEMS, Pakistan, and has served as a lecturer and project coordinator at IQRA National University. His research focuses on AI, ML, DL, NLP, and web development using Flask, and Django.



JABAR ALI is currently working as Assistant Professor of General and Interventional Cardiology at the Department of Cardiology MTI, LRH since December 2016. Previously, he worked as Senior Registrar in the Department of Cardiology MTI, LRH from January 2014 to December 2016.



HEE-CHEOL KIM received his BSc in Mathematics, MSc in Computer Science from SoGang University in Korea, and Ph.D. in Numerical Analysis and Computing Science from Stockholm University in Sweden in 2001. He is a Professor at the Department of Computer Engineering and the Head of the Institute of Digital Anti-Aging Healthcare at Inje University in Korea. His research interests include machine learning, deep learning, computer vision, and medical informatics.