Physics in Medicine &
Biology

PAPER

Assessment of valve regurgitation severity via contrastive learning and multi-view video integration

To cite this article: Sekeun Kim *et al* 2024 *Phys. Med. Biol.* **69** 045020

IOP Publishing Phys. Med. Biol. 69 (2024) 045020 https://doi.org/10.1088/1361-6560/ad22a4

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
10 July 2023

REVISED
19 January 2024

ACCEPTED FOR PUBLICATION
25 January 2024

PUBLISHED
12 February 2024

PAPER

Assessment of valve regurgitation severity via contrastive learning and multi-view video integration

Sekeun Kim[1], Hui Ren[1], Jerome Charton[1], Jiang Hu[1], Carola A Maraboto Gonzalez[2], Jay Khambhati[2], Justin Cheng[3], Jeena DeFrancesco[3], Anam A Waheed[3], Sylwia Marciniak[3], Filipe Moura[3], Rhanderson N Cardoso[3], Bruno B Lima[3], Suzannah McKinney[2], Michael H Picard[4], Xiang Li[1] and Quanzheng Li[1]

[1] Center of Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard Medical School, Boston, MA, United States of America
[2] Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States of America
[3] Brigham and Women's Hospital, Boston, MA, United States of America
[4] Cardiology Division, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States of America

E-mail: skim207@mgh.harvard.edu

Keywords: echocardiography, contrastive learning, multi-view video integration, deep learning

Abstract

Objective. This paper presents a novel approach for addressing the intricate task of diagnosing aortic valve regurgitation (AR), a valvular disease characterized by blood leakage due to incompetence of the valve closure. Conventional diagnostic techniques require detailed evaluations of multi-modal clinical data, frequently resulting in labor-intensive and time-consuming procedures that are vulnerable to varying subjective assessment of regurgitation severity. *Approach*. In our research, we introduce the multi-view video contrastive network, designed to leverage multiple color Doppler imaging inputs for multi-view video processing. We leverage supervised contrastive learning as a strategic approach to tackle class imbalance and enhance the effectiveness of our feature representation learning. Specifically, we introduce a contrastive learning framework to enhance representation learning within the embedding space through inter-patient and intra-patient contrastive loss terms. *Main results*. We conducted extensive experiments using an in-house dataset comprising 250 echocardiography video series. Our results exhibit a substantial improvement in diagnostic accuracy for AR compared to state-of-the-art methods in terms of accuracy by 9.60%, precision by 8.67%, recall by 9.01%, and F_1-score by 8.92%. These results emphasize the capacity of our approach to provide a more precise and efficient method for evaluating the severity of AR. *Significance*. The proposed model could quickly and accurately make decisions about the severity of AR, potentially serving as a useful prescreening tool.

1. Introduction

Aortic valve regurgitation (AR) is a complex valvular disease characterized by the retrograde flow of blood caused by the incompetance of the valve closure. Although this condition appears and impacts in various complex ways, it demands careful understanding and a strategic approach to diagnosis because it significantly affects patient health and care management. Particularly in severe instances of AR often require surgical intervention, such as aortic valve repair or replacement, making the accurate diagnosis of paramount importance (Otto *et al* 2021). Thus, ensuring an accurate and timely diagnosis is not merely a procedural necessity but a critical component in safeguarding patient health, optimizing intervention strategies, and enhancing post-operative recovery and management. The intricacies of diagnosing AR involve navigating through its varied manifestations and understanding its pathophysiological underpinnings, thereby making the diagnostic process both pivotal and challenging in the overarching management of the disease.

The established diagnostic standards, as outlined by the American Heart Association (AHA), involve a comprehensive analysis of both morphological and functional aspects of the heart valves. This analysis employs a multitude of imaging modalities, including color Doppler imaging, continuous wave Doppler (CWD), pulse wave Doppler (PWD), computed tomography (CT), and cardiovascular magnetic resonance (CMR) (Zoghbi *et al* 2017). Among these methods, echocardiography stands out as a vital tool that allows cardiologists to visualize the heart valves and identify potential defects, enabling them to assess valvular function. However, while transthoracic echocardiography (TTE) is a widely adopted protocol for evaluating suspected valvular diseases in clinical settings, it presents certain challenges. TTE requires the expertize of highly skilled cardiologists and is susceptible to inter-user variability. This limitation often leads to variations in assessing the severity of regurgitation.

Recent efforts to improve the precision and accuracy of aortic regurgitation (AR) diagnosis have led to the emergence of machine learning techniques. Edward *et al* (2023) introduced a machine learning approach that relies on a single systolic frame extracted from color Doppler videos as input to their model. While this approach proves effective in detecting regurgitation in pediatric patients, it falls short in accurately assessing the severity of aortic regurgitation due to its sole reliance on selected 2D frame interpretation, which limits its diagnostic capabilities. In response to this limitation, Cheng *et al* (2022) implemented a spatiotemporal convolution layer into their model and devised an aortic valve (AV) regurgitation model using B-mode videos. However, this method depends on a solitary standard view for video interpretation in order to classify the severity of aortic regurgitation (AR). Such an approach may not offer sufficient comprehensiveness for clinical practice since diagnosing AR severity typically incorporates the consideration of multiple views to achieve a comprehensive assessment.

Multi-view convolutional neural networks (MCNNs) aim to combine valuable insights from different perspectives, allowing for the creation of more complete representations that can lead to a more effective classifier (Seeland and Mäder 2021, Vyas *et al* 2020, Edwards *et al* 2023). Marco *et al* (2021) introduced a systematic analysis of utilizing 2D rendered multi-view images for 3D object classification with various fusion methods. However, critical factors such as physics, geometry, and semantics are often shared across all views. Recent studies have provided evidence that high-quality embeddings can yield strong classification performance even when there is a limited amount of labeled data available (Chen *et al* 2023, Tian *et al* 2020a). Chen *et al* (2023) proposed a framework called SimCLR, which generated augmentation-invariant embedding for input images by maximizing agreement between different augmentations of the same image. Experimental results show that the representations learned by SimCLR outperform SOTA in classification tasks. Prannay *et al* (Zhang *et al* 2022) expand the self-supervised contrastive methodology approach by effectively incorporating label information into a supervised contrastive learning (SCL). It facilitates the augmentation of data sharing the same label and results in the creation of efficient visual representation embeddings. This is achieved through the transformation of identical instances, resulting in the generation of a more extensive set of positive and negative pairs.

In this paper, we introduce the multi-view video contrastive network (MVCN), which utilizes a video encoder to capture high-level features and effectively learn a good visual representations by contrastive learning. We hyphothesize the potential use of SCL to enhance accuracy and exhibit good embedding properties, thereby enhancing the efficacy of our proposed method (Tian *et al* 2020b, Zhang *et al* 2022). To achieve this goal, we propose a semantically meaningful contrastive learning approach that assesses relevance across various standard echocardiographic views and among patients with identical levels of severity. To our knowledge, this is the first approach that uses multi-view video (figure 1) for the accurate assessment of AR severity.

Our contributions can be summarized in three main aspects:

- We present the MVCN, which is designed to incorporate multi-view videos for the classification of AR severity using a video feature encoder.
- We propose a contrastive loss specifically designed for echocaridography, which includes both intra-patient and inter-patient loss terms, allowing us to learn semantically relevent visual representations in embedding space.
- We extensively validate our methods using an in-house dataset, which has been annotated by two cardiologists. The results shows the remarkable capabilities of our models, outperforming the performance of state-of-the-art methods.

2. Methods

We propose the MVCN, which comprises three main components: supervised contrastive learning using label information, a shared pre-trained encoder for feature extraction, and the incorporation of inter-intra contrastive loss alongside classification loss in the optimization process as in figure 2.

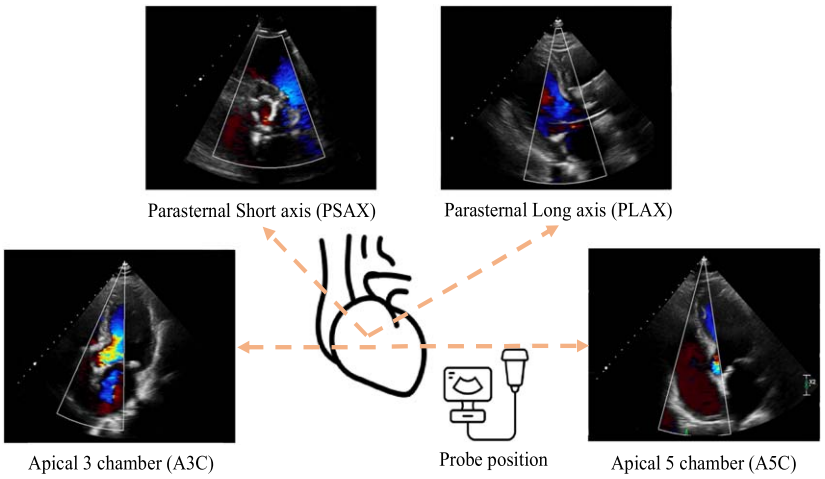

Figure 1. Acquisition of various color Doppler imaging in standard echocardiographic views from varying probe positions. The white region of interest (ROI) within the image represents a specific area that has been selected for analysis. The color represents both the speed and direction of blood flow within the ROI Blood flow towards the probe is visualized in red, while flow away from the probe is depicted in blue.

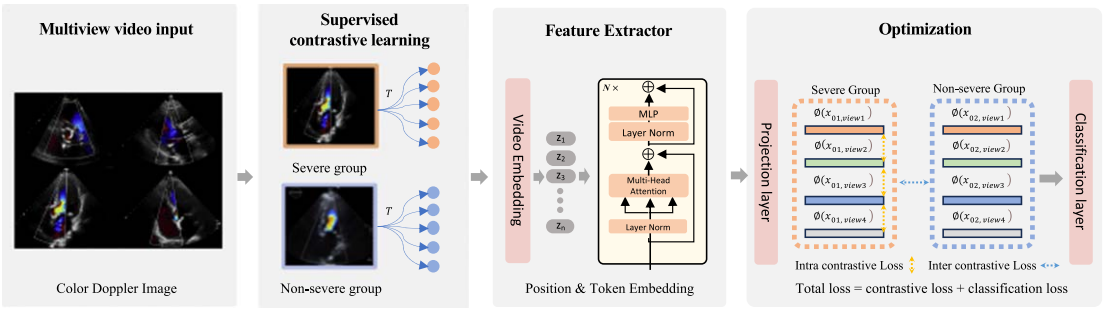

Figure 2. Overall architecture of the proposed method. A supervised contrastive learning to learn representations using a contrastive loss but uses label information to sample positives and negative pairs. A feature extractor which employs a pre-trained encoder shared across all input video. The optimization step involves the use of inter-intra contrastive loss in conjunction with classification loss.

2.1. Feature extractor

We employed a pre-trained video vision transformer (ViViT) (Arnab *et al* 2021) as our shared video encoder, as transformer-based networks demand a significant volume of training data, and ViViT has been pretrained on a large dataset. Given input video, the transformer embeds the positional and temporal information. Each transformer block consists of a layer normalization, multi-head attention layer (Vaswani *et al* 2017), and multi-layer perceptron (MLP). Straight forward choices of tokenization is dividing each frames in spatial domain along with width and height. However, such straightforward choices in generating videos with increasing frames is actually not feasible due to their high computational complexity.

To address this issues, we employed a Tubelet embedding (Arnab *et al* 2021) which divides spatio-temporal tube blocks from the input video with linear projection. Input video dimensions of $T \times h \times w \times 3$, tublet tokens size of $n_t = \left[\frac{T}{t}\right]$, $n_h = \left[\frac{H}{h}\right]$, and $n_w = \left[\frac{W}{w}\right]$ are extracted along the temporal, height, and width, respectively as shown in figure 2, the self-attention mechanism (Dosovitskiy *et al* 2023) calculated as follows:

$$\text{Attention}(Q,\ K,\ V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V, \tag{1}$$

where W_Q, W_k, and W_V are learnable matrices that project the inputs to query, key and value, respectively, and and d is the output dimension of key and query features. After embedding of each video input, these features are fed to projection layter which consists of linear projection, batch normalization (Ioffe and Szegedy 2023), and ReLU activation function (Agarap 2023).

We generates positive and negative pairs on the projected features to learn effective visual representation in embedding space. This is achieved by learning a good visual representation space where semantically relevant

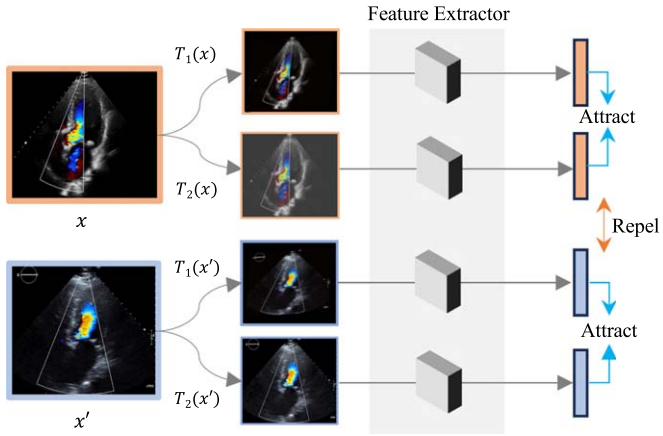

Figure 3. Illustration of supervised contrastive learning scheme: transform input video and fed to feature extraction network then fed to contrastive loss to attract features belonging to a same instance closer, whereas features belonging to different instances are separated. $T(\cdot)$ indicates augmentation operation.

features are attract and repel each other by contrastive losses. Details of contrastive loss term will be described in sections 2.2 and 2.3. We concatenate the projected each features from various video input in the classification layer for classification loss term.

2.2. Intra patient contrastive loss

Our approach draws inspiration from prior supervised contrastive learning (SCL) research (Khosla *et al* 2023) which uses a contrastive loss and data augmentation to learn visual representations in embedding space. Specifically, we group videos belonging to the same severity for augmentation purposes, while maintaining the separation of augmentations for different severity group. We augment input videos of a same severity are grouped together and augmentation in different severity are repel together with predefined augmentation. To mitigate the challenge posed by class imbalance, we strategically increase the frequency of augmentations applied to the severity group during training.

We defined two augmentations for SCL framework. Firstly, we use Gaussian noise and Poisson noise to mimic the speckle which is a multiplicative noise in ultrasound image. Secondly, we employ geometrical transformations, encompassing operations such as rotation ($\pm 30°$), translation ($\pm 5\%$), flipping, and shearing to simulate the various angles and positions to simulate the various angles and positions.

We define positive and negative pairs as described in equations (2), (3), called intra patient contrastive loss, as illustrated in figures 3 and 4(a)

$$\{E_{\text{pos}}\} : = \mathbb{1}_{[i=j]}\{\varnothing(x_{iv}),\ \varnothing(x_{jv})\} \tag{2}$$

$$\{E_{\text{intra,neg}}\} : = \mathbb{1}_{[i\neq j]}\{\varnothing(x_{iv}),\ \varnothing(x_{jv})\}, \tag{3}$$

where $\varnothing$ indicates the projection layer which projects input video features to the low dimensional embedding space. x_i indicate input data, i represents severity groups. v indicates types of standard echo view.

2.3. Inter patient contrastive loss

We propose an inter patient contrastive loss, normalized embeddings from the different views are push each other, as illustrated in figure 3(b). We observe that although various standard video scans the same anatomy and regurgitation blood flow, visually distinct structures are present in different standard views. Given N training data, denote as $D = \{\{x_{iv},\ y_i\}_{v=1}^{4}\}_{i=1}^{N}$, we define the negative pairs as follow:

$$\{E_{\text{inter,neg}}\} : = \mathbb{1}_{[k\neq v]}\{\varnothing(x_{ik}),\ \varnothing(x_{iv})\}, \tag{4}$$

where $\varnothing$ is the view-dependent decoder which projects input video features to the embedding space. i indicate patient index and k represent standard input (k = A2CH, A4CH, PSAX, and PLAX view). We demonstrate the impact of defining pairs in inter contrastive loss table 3.

To optimize intra and inter patient contrastive loss, we employ a mutual information loss, called InfoNCE (van den Oord *et al* 2023) for AV severity classification. The InfoNCE loss is defined as follows:

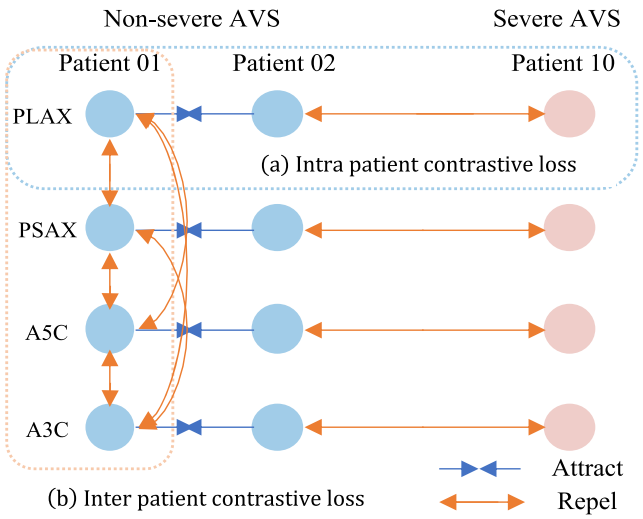

Figure 4. Illustration of inter and intra patient contrastive loss. Blue dots represent non-severe, and red represents severe groups. (a) The intra-patient contrastive loss involves the encoded input view features repelling in different severity groups, while attracting in the same severity group, and (b) the encoded heterogeneous features repelling each other in latent space.

$$L_C = -\log \frac{\sum_{\{Z_{\text{pos}}\}} \exp(\text{sim}(\{E_{\text{pos}}\})/\tau)}{\sum_{\{Z_{\text{pos}}\}} \exp(\text{sim}(\{E_{\text{pos}}\})/\tau) + \sum_{\{Z_{\text{neg}}\}} \exp(\text{sim}(\{E_{\text{neg}}\})/\tau)}, \tag{5}$$

where τ denotes the temperature parameter and is empirically set as 0.05.

2.4. Total loss

Our goal here is to jointly minimize the classification and contrastive losses. One natural is to minimize the following weighted loss:

$$\min_{\theta} \frac{1}{N} \sum_{x_i \in D} \lambda L_{cls} + (1 - \lambda) L_C \tag{6}$$

where L_{cls} denotes cross-entropy loss; L_c denotes the contrastive loss. λ are parameters controlling the relative weights of losses.

Alternatively, we restrict L_c within a certain range and then construct the truncated loss

$$\min_{\theta} \frac{1}{N} \sum_{x_i \in D} \lambda L_{cls} + (1 - \lambda) \max\,(L_C, \epsilon_C), \tag{6'}$$

where $\epsilon_C > 0$ is a user-defined value. Basically, such loss means that if the contrastive loss L_c is greater than the threshold value ϵ_C, it contributes the total loss and otherwise not. We note that (6') reduce to (6) by setting $\epsilon_C = 0$. As there is possibility in (6') that L_C exceeds ϵ_C. To ensure $L_C < \epsilon_C$ with a high probability, we can employ the dynamic loss.

$$\min_{\theta} \frac{1}{N} \sum_{x_i \in D} \lambda L_{cls} + (1 - \lambda) \text{K} \max\,(L_C, \epsilon_C), \tag{6''}$$

where K denotes the number of epochs. Hence, the weight of the contrastive loss in (6'') is getting larger when the iteration involves. If $L_C > \epsilon_C$, the second part in (6'') will dominate the first part for a large K. Hence, $L_C < \epsilon_C$ will be satisfied eventually.

3. Experimental settings

3.1. Dataset

We trained and evaluated our models on a dataset comprising of 183 echocardiography scans from 250 patients at Massachusetts general hospital (MGH), with the approval of the Institutional Review Board (IRB). Following the standard clinical protocols, sonographers localized valve regions and view selections. We incorporated four standard views: the parasternal long axis (PLAX), apical 5 chamber (A5C), apical 3 chamber (A3C), and parasternal short axis (PSAX) at the aortic valve (AV) level. The selection of these views was based on extensive

discussions with clinical experts and adhered to the guidelines provided by the American Society of Echocardiography (ASE).

We set two-step process to create labels from patient data. First, we examine the reports of cardiologists to label all patients. Subsequently, we have two cardiologists assessed and assigned severity grades to all patients, ranging from 0.0 to 4.0 in intervals of 0.5, aligning with the severity levels. However, our objective is to develop a tool for sonographers that enhances the efficiency of patient care and directs attention to more severe cases. To achieve this, we establish a threshold value to distinguish between severe and non-severe groups in clinical practice. After reaching an agreement of our clinical champions, we decided to use a threshold of 2.5 to differentiate between the non-severe and severe groups in our dataset.

In this study, a total of 183 cases which comprised 42 healthy cases and 141 diseased cases. 132(72.1%) were classified as non-severe cases, while the remaining 51(27.9%) were categorized as severe cases. We excluded 67 echocardiography scans where the quality of the focal view was poor, there was a non-standard view of interest, or no images were found for our four standard views. All patients underwent scanning using either GE Vingmed or Philips probes. From a total of 183 samples, we employed stratified random sampling to designate 110 studies for the training dataset, 10 cases for validation, and 63 for the test dataset. This sampling procedure was iterated three times for evaluation purposes. Notably, most patients include multiple scans, such as 2 scans for A2C, 3 scans for PSAX, and similarly for PLAX and A5C. During the training process, we adopted a strategy of randomly selecting one video from one view at each iteration, except in cases where image quality rendered scans non-interpretable. This approach allowed us to generate additional samples, even when working with a relatively small dataset.

To eliminate patient information and other irrelevant data, we carried out the extraction of the echocardiography scan region in the following steps. Our assumption was that pixel intensity fluctuations occurred exclusively within the confines of the scan region. We proceeded to calculate the standard deviation for each pixel location across the frames. We then applied a threshold to these images, creating a binary map where pixel values were set to 0 if they fell below the threshold and 1 if they exceeded the threshold. And we applied erosion, morphological operation that involves shrinking the white regions in the binary image, to remove small noise or isolated pixels, and applied dilation operation to help reconnected broken or fragmented regions in binary image. By applying these erosion and dilation operations sequentially, we aimed to enhance and refine the features of interest within the binary map for further analysis or processing. Then, Edge detection, especially canny edge filter, was used to extract image region, and we generate left and right line to cover fan-shaped image regions.

3.2. Implementation details

The input video is preprocessed by cropping outside of ultrasound scan region and eliminating all patient info. The preprocessed input dimension is $16 \times 3 \times 224 \times 224$ (frame, channel, width, and height), including one cardiac cycle. We include three augmentation methods including random cropping, flipping, and shearing to augment input video data during training network. The best model is selected by identifying the epoch with the best classification accuracy on a validation set. We initialized image encoder with pretrained ViViT model with tublet size of $2 \times 16 \times 16$ for fast convergence during training. We utilized Adam optimizer with the learning rate of 10^{-3}, and the batch size of 5. All experiments were implemented with Pytorch (1.13.0) and trained on A100 (NVIDIA, Santa Clara, CA) with 40 GB memory for 1500 epochs. To evaluate the performance of different methods, we use accuracy, precision, recall, and F_1-score.

4. Experimental results

We conducted an extensive comparison of our method with various state-of-the-art (SOTA) AR classification methods, including R(2+1)D (Tran *et al* 2018), which is a spatiotemporal convolution network based on ResNet (He *et al* 2016). R(2+1)D factorizes the 3D convolutional filter into separate spatial and temporal convolutions, originally designed for action recognition tasks. It is worth noting that Cheng *et al* (2022) utilized R(2+1)D for AR classification within the B-mode A4C view context. We initialized with pretrained weights from the Kinetics-400 dataset (Kay *et al* 2017). Additionally, we considered convLSTM (Shi *et al* 2015), which is based on convolutional long short-term memory networks, designed to capture spatiotemporal relations for downstream tasks. Lastly, we evaluated video vision transformer (ViViT) (Arnab *et al* 2021), a video transformer model that incorporates a video-based self-attention mechanism for classification tasks. In our study, we made modifications to the last layer of ViViT to adapt it for AR severity classification.

Table 1 presents a comparative analysis of the results for AR classification tasks. When using a single view video input, ViViT achieves the best results with an accuracy of 76.0, precision of 79.2, recall of 76.1, and F_1-score of 77.6. The ensemble approach applied to ViViT, which is trained on each of the single views, leads to

Table 1. Comparison of AR severity classification performance generated from our MVCN and other state-of-the-art methods. R(2+1)D and ViViT initialized with pre-trained weights. We employed an ensemble approach named Ensemble ViViT, which integrates the outcomes of four individually trained ViViT models.

Input		Methods	Accuracy	Precision	Recall	F_1-score
Single	Multi					
✓		ConvLSTM (X Shi *et al*)	72.7	76.1	73.0	74.5
✓		R(2+1)D (Tran *et al* 2018)	76.0	79.5	75.4	77.3
✓		ViViT (Arnab *et al* 2021)	76.0	79.2	76.1	77.6
✓		Ensemble ViViT	77.0	81.1	77.1	79.0
	✓	ViViT	77.1	79.5	77.0	78.2
	✓	ViViT + intra contrastive loss	81.4	84.6	80.2	82.3
	✓	ViViT + inter contrastive loss	82.5	85.2	81.2	83.2
	✓	ViViT + inter/intra cross-entropy	82.2	83.9	82.0	82.9
	✓	Proposed	83.3	86.4	82.2	84.2

improvements in accuracy, precision, recall, and F1-score by 1.0, 0.8, 1.0, and 1.3, respectively, when compared to ViViT trained on a single view (A3C). We also explored the combination of multiple video features for AS severity classification and achieved remarkable results. Our findings demonstrate that integrating features from multiple views results in superior classification performance on Doppler image data. Our proposed method achieves the best results with an accuracy of 83.3, precision of 86.4, recall of 82.2, and F_1-score of 84.2. More specifically, the accuracy for the severe class was 84.6, and for the non-severe class, it was 82.2 for our proposed method.

5. Ablation study

We conducted ablation studies to explore various critical elements of our MVCN components and their effects on model performance. This includes (1) the influence of individual components within the contrastive loss function, (2) the effects of losing partial video input, and (3) designing the inter patient contrastive loss.

5.1. Effectiveness of contrastive learning

We first validate the contribution of key components within our method, i.e. contrastive learning, which consists of intra and inter patient contrastive loss. With our dedicated designed contrastive loss, our method consistently outperforms other comparative methods with a significant margin. We have observed notable enhancements in accuracy by 5.3, precision by 5.6, recall by 4.1, and F_1-score by 4.9. It is worth highlighting that we observed a higher contribution from the inter-contrastive loss, resulting in a higher performance in accuracy of 3.1, precision of 3.9, recall of 2.0, and an F_1-score of 2.9 compared to contribution of intra contrastive loss. To further assess the efficacy of contrastive loss, we conducted a comparison with cross-entropy loss, specifically applying cross-entropy classification loss for both inter and intra-patient loss. In this evaluation, we observed an accuracy of 82.2, a precision of 83.9, a recall of 82.0, and an F_1-score of 82.9.

Our method has consistently achieved significantly superior results, demonstrated the advantages of learning semantically relevant information and employed self-supervised learning for AR severity. Our approach leads to substantial improvements, enhancing accuracy by 6.1, precision by 6.9, recall by 5.2, and F_1-score by 6.0 when compared to the multiple ViViT model.

5.2. Impact of partial video loss on model performance

We analyze the influence of partially missing video data on the model's performance, considering variations in echo protocols across different hospitals. In table 2, we deliberately excluded specific views from the input and evaluated the model's performance. Our findings clearly indicate that as the number of missing views increased, the model's performance suffered. When more than two videos were omitted, the model's accuracy deteriorated significantly compared to the single-view ViViT baseline. The accuracy dropped by approximately in terms of accuracy by 2.22%, precision by 5.29%, recall by 4.79%, and F_1-score by 0.99% for missing two inputs, and 5.01%, 7.45% 5.86%, and 1.47% for three input loss.

5.3. Design of inter patient contrastive loss

We conduct ablation experiments to explore the design relationship between pairs, such as positive and negative pairs within the inter-patient contrastive loss. We compare the model performance on inter patient contrastive

Table 2. Evaluation of performance of MCVN method in the presence of partial input loss. # missing view indicates missing input number.

# Missing view	Accuracy	Precision	Recall	F_1-score
One	78.0	80.0	76.7	78.5
Two	74.3	75.0	72.5	76.8
Three	72.2	73.3	71.6	76.5

Table 3. Comparison of model performance with different types for inter patient contrastive loss.

Type	Accuracy	Precision	Recall	F_1-score
Positive	79.5	81.2	78.0	80.2
Negative	81.4	84.6	80.2	82.3

as 'attract' and 'repel' pairs, respectively. Table 3 shows the improvement of negative pairs exceed positive pairs in terms of accuracy by 1.9, precision by 3.3, recall by 2.1, and F1-score by 2.0.

6. Conclusion

In this research, we introduce an innovative MVCN designed to achieve precise diagnosis of aortic valve regurgitation using standard echocardiography scan views, including PLAX, PSAX-AV, A3C, and A5C. Our approach draws inspiration from the diagnostic methods employed by clinicians, who utilize multiple echocardiogram views to assess the severity of AV regurgitation accurately. Our model has demonstrated outstanding performance in AR severity classification tasks. Our method holds significant promise as an AR diagnosis framework that can be applied to various valvular disease for both mitral valve and tricuspid valve.

In clinical implication, our method can provide bedside aid in screening high-risk patients with regurgitation to identify severe cases that may require intervention. Our method serves as a method for primary care physicians to identify patients who should be referred to a specialist. The significance of this procedure is related to the clinical decision workflow of AR diagnosis. Specifically, the current workflow of AR assessment is one step with limited time involving measurements on echocardiography such as ventricle and atrium size, motion, speed, and pattern of colored blood flows in Doppler videos from multiple views, then making a comprehensive decision. On the other hand, the proposed model in this work can make a quick and accurate decision about the severity condition of AR, potentially useful as a prescreening tool.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Code availability

The code for our paper is available at https://github.com/kimsekeun/Assessment-of-Valve-Regurgitation-Severity-via-Contrastive-Learning-and-Multi-view-Video-Integration.

ORCID iDs

Sekeun Kim https://orcid.org/0000-0003-4196-6242
Jerome Charton https://orcid.org/0000-0003-0405-4543
Xiang Li https://orcid.org/0000-0002-9851-6376

References

Agarap A F 2019 Deep learning using rectified linear units (ReLU) arXiv:1803.08375
Arnab A, Dehghani M, Heigold G, Sun C, Lučić M and Schmid C 2021 *ViViT: a video vision transformer Proc. IEEE/CVF international conference on computer vision. 2021* arXiv:2103.15691

Chen T, Kornblith S, Norouzi M and Hinton G 2020 A Simple framework for contrastive learning of visual representations arXiv:2002.05709

Cheng L *et al* 2022 Revealing unforeseen diagnostic image features with deep learning by detecting cardiovascular diseases from apical 4-chamber ultrasounds *J. Am. Heart Assoc.* **11** e024168

Dosovitskiy A *et al* 2021 An image is worth 16 × 16 words: transformers for image recognition at scale arXiv:2010.11929

Edwards L A *et al* 2023 Machine learning for pediatric echocardiographic mitral regurgitation detection *J. Am. Soc. Echocardiogr.* **36** 96–104.e4

He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE) pp 770–8

Ioffe S and Szegedy C 2023 Batch normalization: accelerating deep network training by reducing internal covariate shift *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning* vol 37, pp 448–456

Kay W *et al* 2017 The kinetics human action video dataset arXiv:1705.06950

Khosla P *et al* 2021 Supervised contrastive learning *34th Conference on Neural Information Processing Systems (Vancouver, Canada)* pp 18661–73 https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf

Otto C M *et al* 2021 2020 ACC/AHA guideline for the management of patients with valvular heart disease: a report of the american college of cardiology/American Heart Association joint committee on clinical practice guidelines *Circulation* **143** e72–e227

Seeland M and Mäder P 2021 Multi-view classification with convolutional neural networks *PLoS One* **16** e0245230

Shi X, Chen Z, Wang H, Yeung D-Y, Wong W and Woo W 2015 Convolutional LSTM network: a machine learning approach for precipitation nowcasting *Advances in Neural Information Processing Systems 28 (NIPS 2015)* https://papers.nips.cc/paper_files/paper/2015/hash/07563a3fe3bbe7e3ba84431ad9d055af-Abstract.html

Tian Y, Wang Y, Krishnan D, Tenenbaum J B and Isola P 2020a Rethinking few-shot image classification: a good embedding is all you need? arXiv:2003.11539

Tian Y, Krishnan D, Sun C, Schmid C, Poole B and Isola P 2020b What makes for good views for contrastive learning? arXiv:2005.10243

Tran D, Wang H, Torresani L, Ray J, LeCun Y and Paluri M 2018 A Closer look at spatiotemporal convolutions for action recognition arXiv:1711.11248

van den Oord A, Li Y and Vinyals O 2019 Representation learning with contrastive predictive coding arXiv:1807.03748

Vaswani A *et al* 2017 Attention is all you need arXiv:1706.03762

Vyas S, Rawat Y S and Shah M 2020 Multi-view action recognition using cross-view video prediction *Computer Vision—ECCV* 12372 (*Lecture Notes in Computer Science*) ed A Vedaldi *et al* (Springer International Publishing) pp 427–44

Zhang J *et al* 2022 A class-aware supervised contrastive learning framework for imbalanced fault diagnosis *Knowl. Syst.* **252** 109437

Zoghbi W A *et al* 2017 Recommendations for noninvasive evaluation of native valvular regurgitation *J. Am. Soc. Echocardiogr.* **30** 303–71