Self-Supervised Learning for Detection of Posterior Systolic Curling from Echocardiographic Motion

Nicolò Bergo[1], Andrea Colli[2], Gastone Ciuti[1]

[1] The BioRobotics Institute and Department of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Pisa, Italy
[2] Department of Surgical, Medical and Molecular Pathology and Critical Care Medicine, University of Pisa, Pisa, Italy

Abstract

Posterior systolic curling (PSC) is a dynamic regional abnormality of the posterior mitral annulus and adjacent inferobasal left ventricular wall, associated with arrhythmogenic mitral valve prolapse. Its clinical assessment remains visual and operator-dependent, limiting reproducibility.

We propose an automated framework for patient-level PSC detection from parasternal long-axis echocardiographic cine-loops. A retrospective cohort of 100 subjects (40 PSC, 60 noPSC) was analysed, with labels assigned by three independent clinicians. The posterior annulus–ventricular wall region was localised by deep learning-based segmentation, and motion was encoded using TV-L1 optical flow. Optical-flow clips were used to pretrain a self-supervised motion encoder through contrastive learning, masked motion modelling, and temporal ordering. Downstream diagnosis used a weakly supervised Top-K multiple-instance learning classifier, and performance was evaluated using nested patient-level stratified cross-validation.

The model achieved a mean AUC of 0.910 ± 0.087 and accuracy of 0.86 ± 0.07. Temporal attention highlighted intermediate and late-systolic motion clips, supporting self-supervised optical-flow multiple-instance learning as a promising strategy for automated PSC detection.

1. Introduction

Mitral Valve Prolapse (MVP) is generally considered a benign valvular condition, although a subset of patients may develop an arrhythmogenic phenotype associated with ventricular arrhythmias and, rarely, sudden cardiac death [1]. Among the abnormalities linked to arrhythmic MVP, Posterior Systolic Curling (PSC) has received increasing attention. PSC is characterised by abnormal systolic hypermobility of the posterior mitral annulus and adjacent inferobasal left ventricular wall, producing a regional, time-dependent deformation pattern that may generate repetitive mechanical stretch on the annular–myocardial complex [2,3].

Despite its clinical relevance, PSC diagnosis still relies mainly on qualitative visual assessment of echocardiographic or Cardiac Magnetic Resonance (CMR) cine images, limiting reproducibility and standardisation. Quantitative approaches, including CMR-based measurements and echocardiographic geometric indices such as MIRA and MAIBA, have been proposed to objectify PSC assessment [4,5]. However, these methods remain based on manually or semi-automatically derived descriptors and only partially capture PSC-related spatiotemporal motion. Recent automated motion-analysis approaches further support objective quantification, but still largely depend on predefined or handcrafted features rather than learned motion representations [6].

This study introduces an automated framework for patient-level PSC detection from parasternal long-axis (PLAX) echocardiographic cine-loops. The pipeline combines deep learning-based localisation of the posterior annulus–ventricular wall region, optical-flow motion encoding, self-supervised motion pretraining, and weakly supervised Top-K multiple instance learning (MIL) for patient-level prediction. This approach aims to reduce dependence on manual feature engineering and dense temporal annotations while focusing the classifier on localised PSC-related motion patterns.

2. Methods

An overview of the proposed automated pipeline for patient-level PSC detection is shown in Figure 1.

2.1. Study population

This retrospective study included 100 subjects (median age, 58 years; 44 women) who underwent transthoracic echocardiography at the Cardiothoracic and Vascular Department of the Azienda Ospedaliero-Universitaria

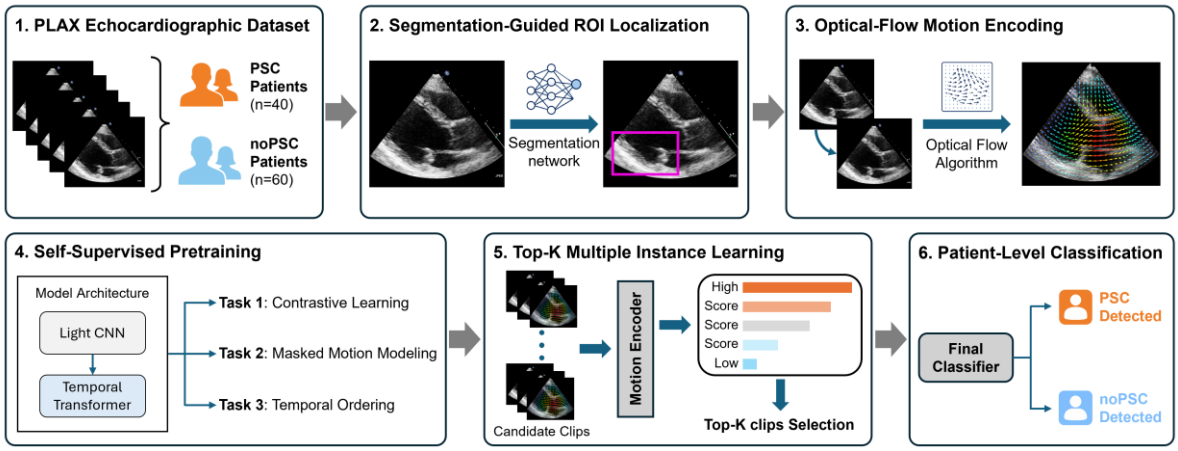

Figure 1. Overview of the proposed automated pipeline for PSC detection from PLAX echocardiographic cine-loops.

Pisana (Pisa, Italy). For each subject, one PLAX cine-loop was analysed. Cine-loops were acquired using Philips iE33 and EPIQ 7C systems (Koninklijke Philips N.V., Amsterdam, Netherlands), with a median frame rate of 58 frames/s. The cohort included 40 PSC and 60 noPSC subjects. PSC status was assigned at the patient level by three independent clinical experts, blinded to clinical data, through full-cycle visual assessment of PLAX cine-loops. Diagnostic criteria included posterior mitral annular hypermobility, the end-systolic "ballerina-foot" pattern, and abnormal displacement with clockwise rotation of the posterior annulus [4,7]. Agreement was reached in 92% of cases, with the remaining cases classified by majority consensus (Fleiss' $\kappa = 0.89$). All patients provided written informed consent for anonymised retrospective post-processing analysis, conducted according to institutional guidelines and Italian/EU data protection regulations.

2.2. Segmentation-guided ROI localisation

To focus motion analysis on the anatomical substrate of PSC, a segmentation-guided module identified the posterior annulus and adjacent ventricular wall in PLAX cine-loops. The module used a DeepLabv3+ network with a ResNet-50 backbone, initialised with ImageNet-pretrained weights and fine-tuned on 1000 manually annotated PLAX frames, obtained by sampling 10 frames from each subject [8]. Data were split at patient level into training, validation, and test sets using a 70%-15%-15% partition. For each cine-loop, the network generated anatomical masks that were used to define a rectangular region of interest (ROI) encompassing the posterior annulus–posterior wall complex. The ROI was cropped from the original sequence and resized to 112 × 112 pixels.

2.3. Optical-flow representation

Optical flow was computed within the segmentation-guided ROI between consecutive PLAX frames to encode regional cardiac motion. After patient-level intensity normalisation and light denoising, dense motion fields were estimated using the TV-L1 algorithm [9]. Each motion map was represented by three channels: horizontal displacement (u), vertical displacement (v), and motion magnitude, thereby capturing local displacement patterns of the posterior annulus–ventricular wall complex.

To reduce optical-flow artefacts, motion maps underwent standardised post-processing, including frame-wise global drift correction, adaptive clipping of extreme displacement values, normalisation of u and v, temporal smoothing, and soft sector masking. The cleaned optical-flow sequence of each patient was divided into fixed-length clips of 16 consecutive optical-flow frames, each representing a local temporal window of regional cardiac motion. Since PSC labels were available only at the patient level, each subject was modelled as a bag of candidate motion clips for downstream weakly supervised learning. Each clip tensor had size 16×3×112×112.

2.4. Self-supervised motion encoder

A self-supervised motion encoder was pretrained on ROI-derived optical-flow clips to learn compact representations of regional cardiac motion without using frame- or clip-level PSC annotations. The encoder combined a lightweight convolutional frame encoder, mapping each optical-flow frame to a 256-dimensional feature vector, with a two-layer, four-head temporal Transformer operating on the 16-frame sequence. Learned positional embeddings were added before the Transformer, and clip-level motion embeddings were obtained by temporal average pooling.

Pretraining used three complementary objectives. A contrastive NT-Xent loss promoted invariance between augmented views of temporally related motion clips while separating unrelated clips [10]. A masked motion modelling task required the network to recover local motion descriptors from masked temporal tokens,

encouraging spatial and temporal coherence of the optical-flow representation [11]. A temporal ordering task trained the model to distinguish the chronological order of paired clips, promoting sensitivity to cardiac motion progression [10]. No clinical labels were used during this pretraining stage. Flow-aware augmentations, including geometrical transformations of displacement fields, temporal and intensity perturbations, and spatial regularisation, were applied to improve robustness, while preserving motion consistency. Pretraining was performed for 200 epochs using AdamW optimisation, a learning rate of 5×10^{-4}, weight decay of 1×10^{-4}, batch size of 32, gradient accumulation of 2, and cosine learning-rate annealing.

2.5. Multiple instance learning classifier

After self-supervised pretraining, the frame encoder and temporal Transformer were transferred to the downstream PSC classifier. Since PSC labels were available only at the patient level, each subject *i* was formulated as an MIL bag of N_i candidate optical-flow clips. For each clip, the encoder produced a 256-dimensional motion embedding, yielding a set of clip-level representations $\{z_{i1}, z_{i2}, \ldots, z_{iNi}\}$ [12].

A Top-K MIL aggregation head was used to select the most informative temporal windows. Each clip embedding was assigned a relevance score by a lightweight attention network with a 128-unit hidden layer, Tanh activation, and scalar output. For each patient, only the K_i highest-scoring clips were retained, with $K_i=\max(1, \lfloor N_i \cdot r \rfloor)$, where r is the Top-K ratio selected within the inner cross-validation loop. The selected embeddings were aggregated using softmax-normalised attention weights to obtain a patient-level representation, followed by a final linear classifier for PSC prediction. During training, 4–12 candidate clips were randomly sampled per patient, and optimisation used binary cross-entropy with logits and class weighting computed only from training subjects. The SSL-pretrained frame encoder was initially frozen during warm-up and subsequently unfrozen for fine-tuning. At inference, $N_{inf}=20$ clips were sampled per patient, and predictions were averaged across 5 deterministic sampling ensembles.

2.6. Training and evaluation strategy

Model evaluation used nested patient-level cross-validation, with all clips from the same subject kept within the same fold to prevent leakage. The outer loop consisted of 5 stratified folds for performance estimation, while a 3-fold inner stratified group cross-validation within each outer-training set was used to select the MIL head learning rate and Top-K ratio. Inner-loop early stopping was allowed only for hyperparameter selection; the held-out outer fold was never used for checkpoint selection.

After hyperparameter selection, each final outer model was trained on the full outer-training set for a fixed budget of 30 epochs, including a 10-epoch warm-up phase. Class imbalance was handled using a positive-class weight computed only from training subjects. Platt calibration and threshold selection were performed exclusively on outer-training out-of-fold predictions. The sensitivity-oriented decision threshold was selected to achieve sensitivity ≥0.80 on these out-of-fold training predictions while maximising specificity and was then applied to the held-out outer fold. Patient-level performance was evaluated using AUC, accuracy, sensitivity, specificity, balanced accuracy, and F1-score. Results were reported across the five outer folds as mean ± standard deviation.

3. Results

3.1. Patient-level diagnostic performance

The proposed model achieved a mean AUC of 0.910 ± 0.087 across the 5 outer cross-validation folds. At the predefined sensitivity-oriented operating point, the model reached an accuracy of 0.86 ± 0.07. Additional patient-level performance metrics are reported in Table 1.

Table 1. Patient-level diagnostic performance.

Metric	Mean ± SD
AUC	0.910 ± 0.087
Accuracy	0.86 ± 0.07
Sensitivity	0.83 ± 0.11
Specificity	0.88 ± 0.14
Balanced accuracy	0.85 ± 0.07
F1-score	0.83 ± 0.08

3.2. Ablation analysis

A compact ablation analysis was performed to assess the contribution of the three main methodological components: self-supervised pretraining, optical-flow motion representation, and Top-K MIL temporal aggregation. Removing SSL, replacing optical flow with raw-image clips, or replacing Top-K MIL with mean pooling reduced performance (Table 2).

Table 2. Compact ablation study.

Model variant	AUC	Accuracy
Proposed model	0.910 ± 0.087	0.86 ± 0.07
No SSL	0.703 ± 0.111	0.58 ± 0.15
Raw-image SSL	0.782 ± 0.112	0.60 ± 0.17
Mean pooling	0.780 ± 0.128	0.74 ± 0.13

3.3. Temporal relevance analysis

MIL attention weights were projected onto the normalized cardiac cycle to qualitatively assess clip relevance. As shown in Figure 2, high-relevance clips were predominantly located before end-systole, with the main attention peak in intermediate and late systole.

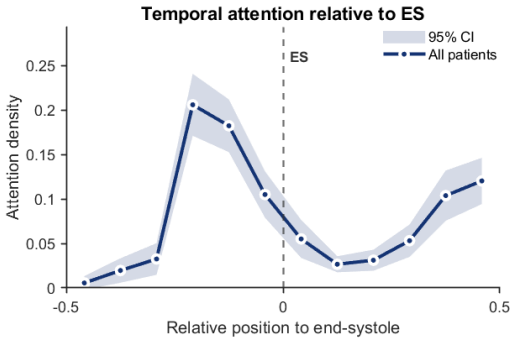

Figure 2. Temporal distribution of MIL attention weights relative to end-systole (ES).

4. Discussion and Conclusion

This study presents an automated framework for detection of PSC from PLAX echocardiographic cine-loops. By combining segmentation-guided anatomical localization, optical-flow motion representation, self-supervised pretraining, and Top-K MIL, the proposed model achieved promising performance across nested cross-validation, with a mean AUC of 0.910 and an accuracy of 0.86. These results support the feasibility of detecting PSC from learned regional motion patterns rather than from manually engineered measurements.

The proposed design is aligned with the dynamic and regional nature of PSC. Optical flow explicitly represents local displacement of the posterior annulus–ventricular wall complex, while self-supervised pretraining enables the encoder to learn cardiac motion representations without dense temporal annotations. Top-K MIL further addresses the weakly supervised setting by representing each patient as a bag of candidate clips and selectively aggregating the most informative temporal windows, which is appropriate for a phenotype expected to be expressed only during specific phases of the cardiac cycle.

The ablation analysis supported the contribution of SSL pretraining, optical-flow representation, and Top-K aggregation. Temporal relevance analysis showed that high-relevance clips were mainly concentrated before end-systole, with a late-systolic attention peak consistent with the expected dynamic expression of PSC. Although attention weights should be interpreted as clip-level relevance scores rather than causal explanations, these findings support self-supervised optical-flow Top-K MIL as a promising strategy for automated patient-level PSC detection. Limitations include the single-center retrospective design, modest cohort size, lack of external validation, and expert-consensus reference standard.

Acknowledgments

This research was supported by the Department of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà 33, Pisa 56127, Italy.

References

[1] Basso C, et al. Mitral Valve Prolapse, Ventricular Arrhythmias, and Sudden Death. Circulation. 2019 Sep 10;140(11):952-964.

[2] Perazzolo Marra M, et al. Determinants of Ventricular Arrhythmias in Mitral Valve Prolapse. JACC Clin Electrophysiol. 2024 Apr;10(4):670-681.

[3] Van der Bijl P, et al. Mitral Annular Disjunction in the Context of Mitral Valve Prolapse: Identifying the At-Risk Patient. JACC Cardiovasc Imaging. 2024;17(10):1229-1245.

[4] Perazzolo Marra M, et al. Morphofunctional Abnormalities of Mitral Annulus and Arrhythmic Mitral Valve Prolapse. Circ Cardiovasc Imaging. 2016 Aug;9(8):e005030.

[5] Bergo N, et al. Development of Semi-automatic Algorithms for the Echocardiographic Diagnosis of Posterior Systolic Curling. Ultrasound Med Biol. 2025 Oct;51(10):1781-1788.

[6] Bergo N, et al. CURL-AID: Automated Echocardiographic Motion Analysis for Quantitative Assessment of Posterior Systolic Curling. Ultrasound Med Biol. Published online August 12, 2026.

[7] Fiore G, et al. Prevalence of Diastolic and Systolic Mitral Annular Disjunction in Patients With Mitral Valve Prolapse. J Am Soc Echocardiogr. 2025;38(1):1-11.

[8] Zhu Z, et al. Automated ejection fraction and risk stratification in cardiomyopathy patients with diverse LV geometry using 2D echocardiography. Sci Rep. 2025 Jul 1;15(1):22332.

[9] Alvén J, et al. A deep multi-stream model for robust prediction of left ventricular ejection fraction in 2D echocardiography. Sci Rep. 2024 Jan 24;14(1):2104.

[10] Holste G, et al. Efficient deep learning-based automated diagnosis from echocardiography with contrastive self-supervised learning. Commun Med (Lond). 2024 Jul 6;4(1):133.

[11] Zhang Z, et al. Echo-Vision-FM: a pre-training and fine-tuning framework for echocardiogram video vision foundation model. Nat Commun. 2025 Dec 11;17(1):19.

[12] Chen J, et al. Weakly supervised video-based cardiac detection for hypertensive cardiomyopathy. BMC Med Imaging. 2023 Oct 19;23(1):163.

Address for correspondence:

Nicolò Bergo
The BioRobotics Institute, Scuola Superiore Sant'Anna
Viale Rinaldo Piaggio, 34 – 56025 Pontedera (Pisa), Italy
nicolo.bergo@santannapisa.it