Segmentation-Guided Multiple Instance Learning for Cardiac MRI Disease Classification

Keno Bressem[1,2⋆], Celine Dürner[1], Zeineb Ben Chaaben[1], Alessa Chami[1], Eva Hendrich[1], Lisa C. Adams[2], and Martin Hadamitzky[1]

[1] German Heart Center, TUM University Hospital, School of Medicine and Health, Technical University of Munich, Munich, Germany
[2] Klinikum rechts der Isar, TUM University Hospital, School of Medicine and Health, Technical University of Munich, Munich, Germany
keno.bressem@tum.de

Abstract. Cardiac magnetic resonance imaging (CMR) examinations have variable sequence availability and slice coverage. SegMIL is presented as a segmentation-guided multiple instance learning framework that represents each examination as a bag of slices. A U-Net branch predicts myocardial probability maps that supervise representation learning and guide decoder-feature pooling before patient-level aggregation. Evaluation comprised 934 internal patients with 12 primary diagnoses and two label-harmonized cine datasets. Five-checkpoint patient ensembles achieved macro AUCs of 0.709 (95% CI 0.671–0.746) internally, 0.702 (0.657–0.745) on ACDC, and 0.701 (0.619–0.776) on Sunnybrook. The strongest respective baselines achieved 0.703, 0.658, and 0.604, although paired confidence intervals included zero. On the common 119-case ACDC subset, replacing reference-mask with predicted-mask preprocessing reduced SegMIL AUC by 0.140, compared with reductions of 0.195–0.359 for mask-input baselines. Performance varied markedly by diagnosis, and macro F1 remained low. SegMIL therefore provides a research framework for anatomy-guided patient aggregation, but the results do not establish robustness to arbitrary missing data or clinical readiness.

Keywords: Cardiac MRI · Multiple instance learning · Segmentation · Domain shift · External validation

1 Introduction

Cardiac magnetic resonance imaging (CMR) supports assessment of myocardial structure, function, and tissue characteristics [13]. Clinical protocols combine complementary sequences, including T2-weighted imaging and T2 mapping for edema, late gadolinium enhancement (LGE) for replacement fibrosis, and native T1 mapping for diffuse fibrosis and infiltration [5]. Examinations nevertheless differ in sequence availability, slice coverage, scanner, and acquisition protocol.

⋆ Corresponding author.

Table 1. Internal cohort and sequence availability. Values denote training/validation/test cases for one split.

Primary diagnosis	T2-w	T2 map	LGE	Native T1 map	Patients
Acute myocardial infarction	23/5/10	24/5/10	24/5/10	24/5/10	24/5/10
Acute myocarditis	47/10/20	34/10/20	48/10/20	34/10/20	48/10/20
Amyloidosis	14/5/10	9/5/10	15/4/10	9/5/10	15/5/10
Dilated cardiomyopathy	77/10/20	77/10/20	77/10/20	77/10/20	77/10/20
Hypertrophic cardiomyopathy	70/10/19	32/10/19	68/10/19	32/10/19	70/10/19
Non-specific myocardial changes	93/10/19	93/10/19	93/10/19	93/10/19	93/10/19
Normal findings	85/8/19	85/8/19	84/8/19	85/8/19	85/8/19
Old myocardial infarction	32/6/16	30/6/16	33/6/16	30/6/16	33/6/16
Old myocarditis	142/10/20	143/10/20	143/10/20	143/10/20	143/10/20
Pericarditis	31/5/10	19/5/10	32/5/10	20/5/10	32/5/10
Sarcoidosis	8/5/5	1/4/5	8/5/4	1/4/5	9/5/5
Specific cardiomyopathy	29/6/17	19/6/17	30/6/17	19/6/17	30/6/17
Overall	651/90/185	566/89/185	655/89/184	567/89/185	659/90/185

Fixed-volume classifiers therefore require rigid resampling or separate sequence predictions.

In multiple instance learning (MIL), an examination is represented as a bag of images with one study-level label [4,7]. Anatomical priors may additionally constrain features to clinically relevant structures and reduce shortcut learning [3]. A common implementation concatenates a myocardial mask with the image, but classification can then depend on mask errors and crop localization. Such upstream dependencies should be evaluated under realistic data shifts [9].

SegMIL combines MIL with segmentation-supervised representation learning. Myocardial probabilities guide multi-level decoder pooling, while a transformer aggregates slices into one patient prediction. Evaluation includes explicit cine harmonization and bag construction, original external probabilities, patient-level uncertainty, controlled sequence removal, mask-shift analysis, and clinical classification metrics.

2 Methods

2.1 Internal cohort, labels, and sequence availability

The internal cohort comprised 934 patients examined in clinical routine at the German Heart Center Munich on the same 1.5-T Siemens Avanto scanner. Available sequences included T2-weighted imaging, T2 mapping, LGE, and native pre-contrast T1 mapping. A fixed patient-level split reserved 185 cases for testing; 749 development cases were divided into 659 training and 90 validation cases in each of five stratified random splits. The observed missing-sequence distribution is reported in Table 1. Use of the cohort was approved by the local institutional review board.

One primary study-level diagnosis was derived from the routine report after review of the complete examination by a radiology resident with 4 years of experience under supervision of a cardiovascular radiologist with 20 years

of experience. Representative slices were not selected to establish the diagnosis. Mixed findings and comorbidities were collapsed into this single primary label. Per-pixel myocardium masks were created by a medical student under supervision of a cardiovascular radiologist with 9 years of experience and served as segmentation targets.

2.2 Localization, segmentation, and classifier inputs

A standalone two-dimensional MONAI FlexibleUNet with an EfficientNet-B2 encoder generated myocardium masks [12,2,15]. Inputs were histogram-normalized, Gaussian-smoothed ($\sigma = 0.1$), scaled to $[-1, 1]$, and center-cropped to 196×196 pixels. Augmentations comprised flips, rotations, scaling, noise, bias fields, Gibbs artifacts, and intensity perturbations. The model was trained for 250 epochs on slices pooled across the internal sequences using Dice plus cross-entropy loss, AdamW, and a OneCycleLR schedule. Validation Dice of the model was 0.860.

All classifier families used a mask for localization or center crops. For 2D and 3D models, a foreground crop with a 16-pixel in-plane and one-slice margin was resampled to $1 \times 1 \times 8\,\text{mm}$, padded to $128 \times 128 \times 16$, or center-cropped to $128 \times 128 \times 12$. For MIL and SegMIL, each valid slice was either center cropped to 128×128 or cropped to the mask bounding box with a 12-pixel margin and minimum 8×8 box, then resized to 128×128. Reference masks localized internal examinations, whereas predicted masks localized the primary external evaluations. The predicted-mask foreground was retained by construction, although an erroneous mask could mislocalize the true myocardium. Within each architecture family, "+mask" and image-only variants received cropped images; only the former concatenated the mask as a second classifier channel. SegMIL consequently did not need an additional mask channel during inference.

2.3 Classifier variants and implementation

Seven variants were evaluated: 2D and 3D EfficientNet-B2 classifiers, conventional MIL classifiers, each with and without a mask input channel, and SegMIL. The 2D model encoded every slice and mean-pooled slice features within one sequence. The 3D model encoded the standardized sequence volume. Available sequence softmax vectors from either non-MIL model were averaged to obtain one patient vector. Conventional MIL and SegMIL directly generated one patient vector through classification-token transformer aggregation [14].

MIL bags contained slices from all available sequences and were capped at 48 instances. All instances were retained at or below this cap. Larger bags were randomly subsampled without replacement, after which the selected order was preserved. Shorter bags were zero-padded to the longest bag in a batch. A validity mask was passed to every bag-transformer layer as a key-padding mask after 2D encoding, preventing the classification token from attending to padded instances. This implementation accepts variable-length bags, but no controlled bag-size experiment was performed.

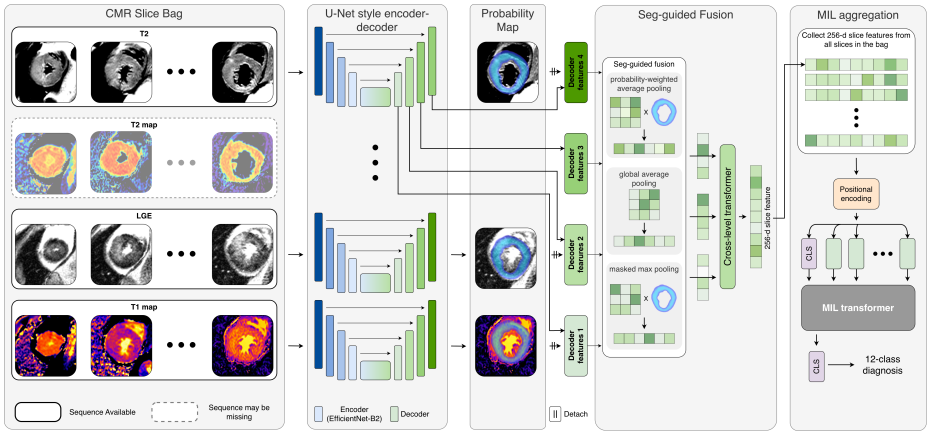

Fig. 1. SegMIL architecture. Localized slices from available sequences form a patient bag. A shared U-Net predicts detached myocardial probabilities that guide pooling of four decoder levels. Slice descriptors are aggregated by a classification-token transformer. Upstream mask-based localization, used by every model, is described in Sect. 2.2.

The 2D, 3D, and conventional MIL models were randomly initialized without ImageNet weights. SegMIL's U-Net was initialized from the standalone segmentation checkpoint, whereas its fusion, transformer, and classification modules were randomly initialized. Image-only and mask variants had 7,717,326 and 7,717,614 trainable parameters for 2D, 8,731,854 and 8,732,718 for 3D, and 10,040,666 and 10,040,954 for MIL; SegMIL had 14,000,662. In the recorded PyTorch 2.5.1 and MONAI 1.5.0 environment, inference cost was 2.559 and 2.573 giga multiply-accumulate operations (GMAC) for one 2D sequence, 0.469 and 0.491 GMAC for one 3D sequence, 10.350 and 10.406 GMAC for a 48-instance MIL bag, and 33.398 GMAC for a 48-instance SegMIL bag.

2.4 SegMIL architecture and training

SegMIL replaced the conventional per-slice encoder with a joint segmentation-classification branch (Fig. 1). The U-Net predicts a myocardial probability map for each slice. The map is supervised by the reference segmentation during training and detached before classification pooling, so classification gradients do not update the segmentation head through the map. At four decoder levels, probability-weighted average, global average, and probability-masked maximum pooling were projected to 256-dimensional representations and mixed by a cross-level transformer. The resulting slice descriptors were aggregated by a two-layer bag transformer.

All classifiers were trained with class-weighted cross-entropy, label smoothing 0.15, AdamW optimization, learning rate 5×10^{-5}, weight decay 0.04, cosine warm restarts, batch size 8, and 50 epochs. MIL losses additionally included top-

5 instance clustering and attention-entropy regularization. SegMIL combined slice-level segmentation and patient-level classification losses; classification was introduced after epoch 5 and reached full weight by epoch 10, after which the segmentation-loss weight was 0.5. Four ablations removed the auxiliary segmentation loss, segmentation-guided pooling, learned input-normalization block, or segmentation-checkpoint initialization.

2.5 External harmonization and statistical analysis

ACDC and Sunnybrook were evaluated using end-diastolic short-axis cine volumes [1,11]. For ACDC, the dataset-specified end-diastolic frame and all short-axis slices were used. DCM, HCM, MINF, and NOR were mapped to dilated cardiomyopathy, hypertrophic cardiomyopathy, old myocardial infarction, and normal findings; the unmatched RV group was excluded. For Sunnybrook, the frame with the lowest trigger time at each slice location was treated as end-diastole, and all available locations were stacked. Heart failure without infarction, hypertrophy, heart failure with infarction, and normal were mapped to the same four classes, respectively. The infarction group was mapped to old infarction because its dataset definition includes LGE evidence. One ACDC and one Sunnybrook examination could not be converted to a usable volume and were excluded, leaving 119 and 44 patients. Each cine volume was treated as one sequence, and each valid slice became one MIL instance.

All reported classification metrics were calculated at patient level. Five split-specific softmax vectors were averaged per patient for primary analysis. Internal macro AUC averaged 12 one-versus-rest AUCs. External macro AUC averaged the four overlapping outputs using their original, non-renormalized 12-class probabilities. Renormalization over the overlapping classes was not applied because a patient-specific denominator can change between-patient score order and thus AUC. Classification metrics comprised macro F1, balanced accuracy, and sensitivity. External class predictions used the argmax among the four overlapping outputs.

Uncertainty was estimated with 2,000 class-stratified patient bootstrap resamples. Percentile 95% confidence intervals were calculated for aggregate and per-class metrics. SegMIL was compared with the strongest non-SegMIL ensemble using paired resamples. A controlled internal analysis removed one sequence type at a time. ACDC mask shift was evaluated by replacing reference-mask with predicted-mask preprocessing on the common 119-patient subset.

3 Results

3.1 Patient-level classification and class variation

SegMIL produced the highest ensemble macro AUC on all three datasets (Table 2). The paired differences from the strongest baselines were 0.006 internally (95% CI −0.030 to 0.045), 0.044 on ACDC (−0.021 to 0.106), and 0.098

Table 2. Five-checkpoint patient-ensemble macro AUC (95% patient-bootstrap CI). External AUCs use original, non-renormalized probabilities.

Model	Internal, $n = 185$	ACDC, $n = 119$	Sunnybrook, $n = 44$
2D	0.683 [0.648, 0.719]	0.658 [0.593, 0.719]	0.595 [0.490, 0.693]
2D+mask	0.692 [0.655, 0.729]	0.546 [0.484, 0.607]	0.512 [0.400, 0.618]
3D	0.690 [0.657, 0.723]	0.642 [0.581, 0.706]	0.604 [0.501, 0.708]
3D+mask	0.703 [0.665, 0.741]	0.574 [0.519, 0.629]	0.533 [0.443, 0.613]
MIL	0.607 [0.563, 0.649]	0.564 [0.516, 0.609]	0.499 [0.418, 0.580]
MIL+mask	0.611 [0.574, 0.647]	0.523 [0.473, 0.575]	0.521 [0.461, 0.586]
SegMIL	**0.709 [0.671, 0.746]**	**0.702 [0.657, 0.745]**	**0.701 [0.619, 0.776]**

on Sunnybrook (−0.022 to 0.214). Each interval included zero. SegMIL macro F1/balanced accuracy was 0.216/0.280 internally, 0.317/0.376 on ACDC, and 0.288/0.391 on Sunnybrook, indicating that rank discrimination did not translate into reliable multiclass decisions.

Internal class-specific AUCs ranged from 0.381 for sarcoidosis to 0.991 for amyloidosis (Fig. 2A). Performance was also comparatively high for dilated and hypertrophic cardiomyopathy, with AUCs of 0.916 and 0.857, respectively. Lower AUCs were observed for acute infarction (0.524), non-specific changes (0.553), old myocarditis (0.607), and sarcoidosis (0.381). Overall, diseases with distinct morphologic phenotypes appeared easier to discriminate than those characterized by diffuse or overlapping findings. The characteristic post-contrast phenotype of amyloidosis likely aided differentiation, although sample size and label separability may also have contributed. At the ensemble argmax, sensitivity was zero for six internal classes, including acute infarction, old myocarditis, non-specific changes, and sarcoidosis. These classwise failures remain a major limitation despite moderate AUCs.

Removing LGE produced the largest performance reduction, from 0.709 to 0.688, consistent with LGE being the only contrast-enhanced sequence. Removing T2 mapping yielded an AUC of 0.695, compared with 0.711 after removal of native T1 mapping and 0.719 after removal of T2-weighted imaging (Fig. 2B). These findings indicate that the model accommodates varying sequence availability, although only removal of LGE or T2 mapping reduced macro AUC.

3.2 Mask shift and component ablations

The external segmentation model achieved mean ACDC Dice 0.425±0.210 across the 120 cases with reference masks; 12 had Dice below 0.1. On the common 119-patient subset, replacement of reference-mask with predicted-mask preprocessing reduced SegMIL AUC from 0.841 to 0.702, a paired change of −0.140 (95% CI −0.187 to −0.094; Fig. 2C). The corresponding reductions were −0.359 for 2D+mask, −0.306 for 3D+mask, and −0.195 for MIL+mask. Thus, SegMIL showed a smaller observed reduction than mask-input baselines, but every model remained dependent on mask-guided localization and the analysis cannot isolate classifier-channel error from crop mislocalization.

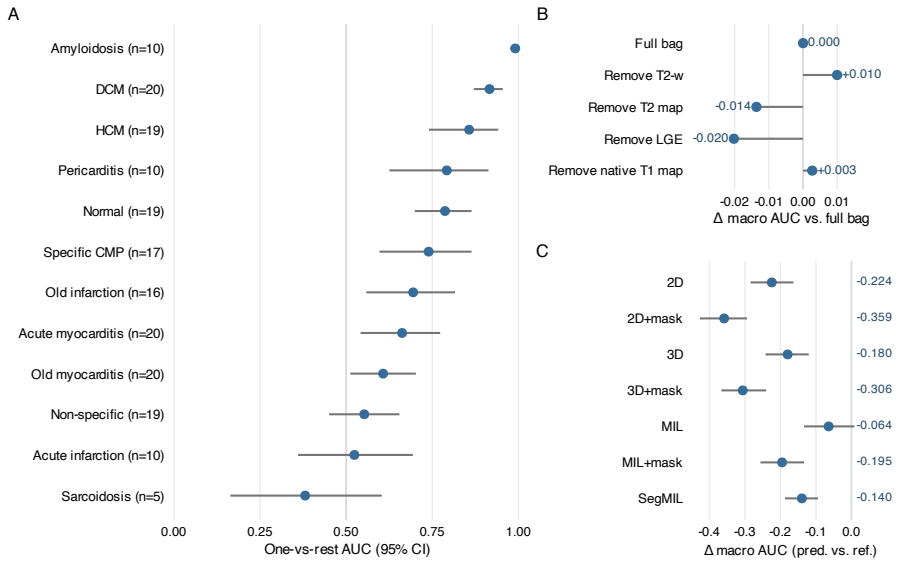

Fig. 2. Patient-level analyses. (A) Internal SegMIL per-class AUC with class-stratified bootstrap intervals. (B) Change in internal SegMIL macro AUC after removing one available sequence type (full-bag AUC 0.709). (C) Paired change in ACDC macro AUC when predicted-mask replaced reference-mask preprocessing. Error bars denote 95% patient-bootstrap intervals.

The five-checkpoint ensemble ablations are summarized in Table 3. Removing segmentation initialization caused the largest external reductions. Removing auxiliary segmentation loss also reduced both external AUCs, whereas removal of guided pooling or input normalization had smaller and dataset-dependent effects.

4 Discussion

SegMIL combines segmentation-supervised learning with MIL for CMR disease classification under variable sequence availability and variable slice coverage. The MIL formulation treats each study as a bag of slice-level instances, so examinations are processed regardless of how many slices or which sequences are available, and whatever data exist for a patient can be passed to the model directly without resampling to a fixed input. Ablations indicate that segmentation initialization and the auxiliary segmentation loss, rather than guided pooling, drove the external stability of the model.

SegMIL yielded the highest point estimate internally and on both cine datasets, although paired intervals against the strongest baselines were statistically not significant. Compared to other models, SegMIL did not consume a mask as a

Table 3. SegMIL component ablations using patient-ensemble macro AUC.

Variant	Internal	ACDC	Sunnybrook
Full SegMIL	0.709	**0.702**	0.701
No auxiliary segmentation loss	0.710	0.618	0.600
No segmentation-guided pooling	**0.715**	0.696	0.683
No input-normalization block	0.711	0.700	**0.707**
No segmentation initialization	0.695	0.548	0.551

classifier channel and degraded less than the three mask-input baselines when ACDC preprocessing was changed. Nevertheless, the low external Dice and the common localization step show that none of the models was independent of mask quality. A SAM- or MedSAM-like model could improve localization and mask baselines if it produced more accurate myocardium masks [8,10]. Such systems are substantially larger, generally require point or box prompts or an automatic prompting strategy, and would still require validation on cine CMR.

SegMIL performed well for diagnoses with strong anatomical or tissue-pattern priors, but lower AUCs were observed for old myocarditis, non-specific myocardial changes, acute infarction, and sarcoidosis. These diagnoses have overlapping or diffuse CMR appearances, and chronic inflammatory findings can be difficult to separate even clinically [6]. Lower performance therefore likely reflects not just algorithmic failure but also limited sample size, phenotype overlap, and the difficulty of assigning study-level labels from routine examinations. Further limitations include the single-center, single-scanner internal cohort, unavailable demographic summaries, and external evaluation on only four overlapping cine labels. The cine datasets tested modality and institution shift but not multi-sequence missingness. The controlled removal analysis did not cover combinations of missing sequences, and neither controlled bag-size experiments nor simulated mask perturbations were available. External validation of the segmentation model was limited to ACDC. Architectural support for variable bags should therefore not be interpreted as evidence of robustness.

5 Conclusion

SegMIL integrates segmentation-supervised decoder features with patient-level MIL. It produced the highest external point estimates and smaller ACDC mask-shift degradation than mask-input baselines. These findings motivate joint segmentation-classification as a training strategy, although differences from baselines were not statistically significant.

Prospects of Application: SegMIL is a research-stage framework for anatomy-guided patient-level CMR classification. Clinical use requires multicenter full-label-space validation, stronger reference standards, and prospective evaluation.

Acknowledgments. This work was supported by the German Heart Foundation/ German Foundation of Heart Research (grant F/28/24 to K.B.).

Disclosure of Interests. K.B. has received grants from the European Union, Bayern Innovativ, the German Federal Ministry of Research, Technology and Space, the Else Kröner-Fresenius Foundation, the Max Kade Foundation, and the Wilhelm Sander Foundation. K.B. serves as an advisor to the EU Innovative Health Initiative project IMAGIO (grant agreement No. 101112053) and has received payments from Novartis. M.H. has received a research grant from Cleerly. The remaining authors declare no competing interests.

References

1. Bernard, O., Lalande, A., Zotti, C., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (2018). https://doi.org/10.1109/TMI.2018.2837502
2. Cardoso, M.J., Li, W., Brown, R., et al.: MONAI: an open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022). https://doi.org/10.48550/arXiv.2211.02701
3. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11**(1), 3673 (2020). https://doi.org/10.1038/s41467-020-17478-w
4. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence **89**(1–2), 31–71 (1997). https://doi.org/10.1016/S0004-3702(96)00034-3
5. Ferreira, V.M., Schulz-Menger, J., Holmvang, G., Kramer, C.M., Carbone, I., Sechtem, U., Kindermann, I., Gutberlet, M., Cooper, L.T., Liu, P., Friedrich, M.G.: Cardiovascular magnetic resonance in nonischemic myocardial inflammation: expert recommendations. Journal of the American College of Cardiology **72**(24), 3158–3176 (2018). https://doi.org/10.1016/j.jacc.2018.09.072
6. Gutberlet, M., Spors, B., Thoma, T., Bertram, H., Denecke, T., Felix, R., Noutsias, M., Schultheiss, H.P., Kühl, U.: Suspected chronic myocarditis at cardiac MR: diagnostic accuracy and association with immunohistologically detected inflammation and viral persistence. Radiology **246**(2), 401–409 (2008). https://doi.org/10.1148/radiol.2461062179
7. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136 (2018)
8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023). https://doi.org/10.1109/ICCV51070.2023.00371
9. Lekadir, K., Frangi, A.F., Porras, A.R., et al.: FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ **388**, e081554 (2025). https://doi.org/10.1136/bmj-2024-081554
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024). https://doi.org/10.1038/s41467-024-44824-z
11. Radau, P., Lu, Y., Connelly, K., Paul, G., Dick, A.J., Wright, G.A.: Evaluation framework for algorithms segmenting short axis cardiac MRI. In: MIDAS Journal

– Cardiac MR Left Ventricle Segmentation Challenge (2009), http://hdl.handle.net/10380/3070
12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Schulz-Menger, J., Bluemke, D.A., Bremerich, J., et al.: Standardized image interpretation and post-processing in cardiovascular magnetic resonance: 2020 update. Journal of Cardiovascular Magnetic Resonance **22**(1), 19 (2020). https://doi.org/10.1186/s12968-020-00610-6
14. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In: Advances in Neural Information Processing Systems. vol. 34, pp. 2136–2147 (2021)
15. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114 (2019)