M(otion)-mode Based Prediction of Ejection Fraction using Echocardiograms

Ece Ozkan*[1][0000−0002−9889−6348] (✉), Thomas M. Sutter*[2][0000−0001−7503−4473] (✉), Yurong Hu[3][0009−0008−8997−0543], Sebastian Balzer[4], and Julia E. Vogt[2][0000−0002−6004−7770]

[1] Department of Brain and Cognitive Sciences, MIT, USA
[2] Department of Computer Science, ETH Zurich, Switzerland
[3] Department of Information Technology and Electrical Engineering, ETH Zurich, Switzerland
[4] Department of Biosystems Science and Engineering, ETH Zurich, Switzerland
✉ eoezkan@mit.edu, suttetho@inf.ethz.ch
* Shared first authorship.

Abstract. Early detection of cardiac dysfunction through routine screening is vital for diagnosing cardiovascular diseases. An important metric of cardiac function is the left ventricular ejection fraction (EF), where lower EF is associated with cardiomyopathy. Echocardiography is a popular diagnostic tool in cardiology, with ultrasound being a low-cost, real-time, and non-ionizing technology. However, human assessment of echocardiograms for calculating EF is time-consuming and expertise-demanding, raising the need for an automated approach. In this work, we propose using the M(otion)-mode of echocardiograms for estimating the EF and classifying cardiomyopathy. We generate multiple artificial M-mode images from a single echocardiogram and combine them using off-the-shelf model architectures. Additionally, we extend contrastive learning (CL) to cardiac imaging to learn meaningful representations from exploiting structures in unlabeled data allowing the model to achieve high accuracy, even with limited annotations. Our experiments show that the supervised setting converges with only ten modes and is comparable to the baseline method while bypassing its cumbersome training process and being computationally much more efficient. Furthermore, CL using M-mode images is helpful for limited data scenarios, such as having labels for only 200 patients, which is common in medical applications.

Keywords: Echocardiography · M-mode Ultrasound · Ejection Fraction · Computer Assisted Diagnosis (CAD)

1 Introduction

Cardiovascular diseases (CVD) are the leading cause of death worldwide, responsible for nearly one-third of global deaths [29]. Early assessment of cardiac dysfunction through routine screening is essential, as clinical management and

behavioral changes can prevent hospitalizations and premature deaths. An important metric for assessing cardiac (dys)function is the left ventricular (LV) ejection fraction (EF), which evaluates the ratio between LV end-systolic and end-diastolic volumes [3, 21].

Echocardiography is the most common and readily available diagnostic tool to assess cardiac function, ultrasound (US) imaging being a low-cost, non-ionizing, and rapid technology. However, the manual evaluation of echocardiograms is time-consuming, operator-dependent, and expertise-demanding. Thus, there is a clear need for an automated method to assist clinicians in estimating EF.

M(otion)-mode is a form of US, in which a single scan line is emitted and received at a high frame rate through time to evaluate the dynamics to assess different diseases [23]. M-mode is often utilized in clinical practice e.g. in lung ultrasonography [1, 25] or echocardiography [6, 7, 26, 10]. Since cardiac function assessment relies on heart dynamics, M-mode images can be an excellent alternative to B(rightness)-mode image- or video-based methods. However, little effort is directed toward exploiting M-mode images in an automated manner.

Data collection and annotation are expensive for most applications. Therefore, learning from limited labeled data is critical in data-limited problems, such as in healthcare. To overcome this data bottleneck, self-supervised learning (SSL) methods have been recently proposed to learn meaningful high-level representations from unlabeled data [16, 24].

Related Work A few existing works [14, 18] reconstruct M-mode images from B-mode videos to detect pneumothorax using CNNs. Furthermore, authors in [27] propose an automatic landmark localization method in M-mode images. A more related method using M-mode images in an automated manner to estimate EF is [22], which uses single M-mode images in parasternal long-axis view to measure chamber dimensions for calculating EF.

For automated EF prediction, some previous works exploit either still-images [17, 31, 8] or spatio-temporal convolutions on B(rightness)-mode echocardiography videos [21]. However, still-image-based methods have a high variability [20], and video-based methods rely on a complex pipeline with larger models. Furthermore, [19] uses vision transformers and CNNs to tackle the problem of estimating the LV EF, and [15] uses geometric features of the LV derived from ECG video frames to estimate EF. The authors in [28] evaluate ML-based methods in a multi-cohort setting using different imaging modalities. In the SSL setting, [5] propose a contrastive learning framework for deep image regression, which consists of a feature learning branch via a novel adaptive-margin contrastive loss and a regression prediction branch using echocardiography frames as input.

Our Contribution We propose to extract images from readily available B-mode echocardiogram videos, each mimicking an M-mode image from a different scan line of the heart. We combine the different artificial M-mode images using off-the-shelf model architectures and estimate their EF to diagnose cardiomyopathy in a supervised regime. Using M-mode images allows the model to naturally observe the motion and sample the heart from different angles while bypassing cumbersome 3D models. Secondly, we propose an alternative scheme for pre-

dicting EF using generated M-mode images in a self-supervised fashion while extending contrastive learning. We design a problem-specific contrastive loss for M-mode images to learn representations with structure and patient awareness. We evaluate both regimes on the publicly available EchoNet-Dynamic dataset ([20]) and demonstrate both models' effectiveness.

To the best of our knowledge, this is the first work on image-based and temporal information incorporating cardiac function prediction methods to estimate EF. Furthermore, our method can easily be applied to other problems where cardiac dynamics play an essential role in the diagnosis. To ensure reproducibility, we made the code available: https://github.com/thomassutter/mmodeecho.

2   Methods

This work aims to create a pipeline with as little intervention as possible; thus, our method consists of two parts, as shown in Figure 1. The first part is extracting M-mode images from readily available B-mode videos. The second part includes representation learning, which are lower-level information that preserves more information of the input image and are used to predict EF from M-mode images, including two schemes: supervised and self-supervised learning.

2.1   From B-mode Videos to M-mode Images

Assume our dataset contains N patients. For each patient $i = \{1, 2, \cdots, N\}$, the label y_i indicates its EF. Furthermore, the B-mode echocardiogram video of each patient i is given of size $h \times w \times t$ with h being height, w width, and t number of frames of the video. The m-th M-mode image of patient i is given as $\boldsymbol{x}_i^m$ with $m = \{1, 2, \cdots, M\}$. It is a single line of pixels through the center of the image with an angle θ_m over frames, assuming LV is around the center throughout the video, as in Figure 1(a). This image, corresponding to θ_m, is then of size $s_m \times t$, with s_m as the length of the scan line. For simplicity, we set $s_m = h \; \forall \; m$ independent of its angle θ_m. For generating multiple M-mode images, a set of M angles $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_M]$ is used to generate M M-mode images, where the angles $\boldsymbol{\theta}$ are equally spaced between $0°$ and $180°$.

While the proposed approach for generating M-mode images is intuitive and works well (see Section 3.3), other approaches are also feasible. For instance, the center of rotation in the middle of the image in our M-mode generation process could be changed. Like that, we could mimic the behavior of the data collection process as every generated M-mode image would resemble a scan line of the US probe. However, the main goal of this work is to highlight the potential of M-mode images for the analysis of US videos. Given our convincing results, we leave the exploration of different M-mode generation mechanisms for future work.

2.2   Learning Representations from M-mode Images

Supervised Learning for EF Prediction We aim to learn supervised representations using off-the-shelf model architectures to estimate EF. Instead of
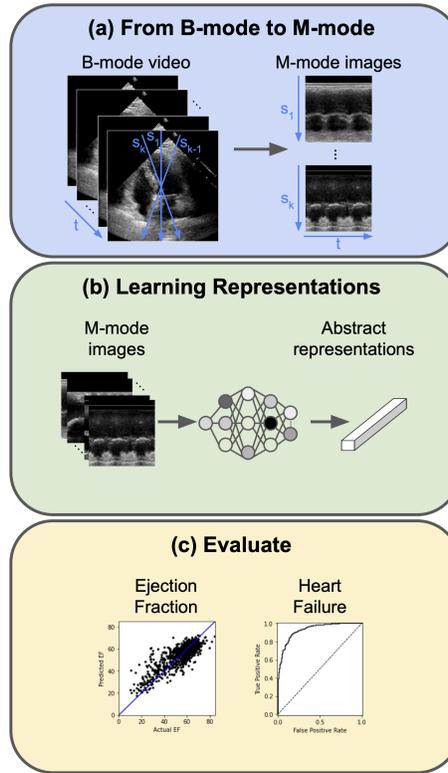
Fig. 1: Overview of our proposed method. (a) Generate M-mode images from B-mode echocardiography videos at different scan lines. (b) Learn representations from the generated M-mode images using supervised and self-supervised learning schemes. (c) Evaluate EF prediction to diagnose cardiomyopathy.

using a single M-mode, one can aggregate the information of M-mode images from the same patient to increase robustness. We evaluate two fusion methods for aggregating information among the M M-mode images: early-fusion and late-fusion [2]. With early fusion, we construct a $M \times s \times t$ image with the M M-mode images being the M channels of the newly created image. In late-fusion, we exploit three different methods. For all of the late-fusion schemes, we first infer an abstract representation z_i^m for every M-mode image x_i^m. The representations z_i^m are then aggregated to a joint representation $\tilde{z}_i$ using an LSTM cell [11], averaging, or concatenating.

We utilize a standard ResNet architecture [9] with 2D-convolutional layers independent of the fusion principle. With 2D-convolutions, we assume a single M-mode image as a 2D gray-scale image with two spatial dimensions, s and t.

(a)

(b)
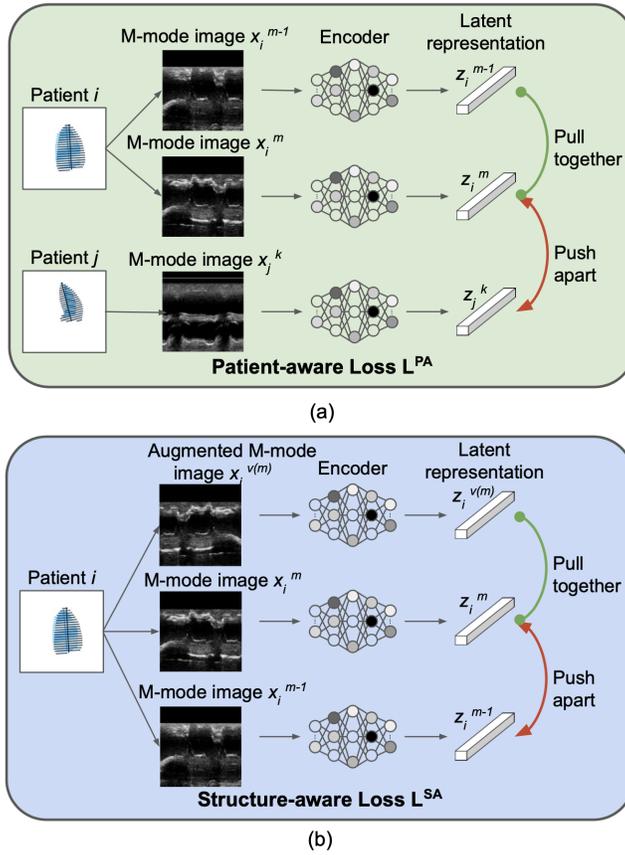
Fig. 2: Overview of our proposed SSL method. The contrastive loss includes (a) patient awareness to attract similarity between data from the same patient while discouraging it between different patients and (b) structure awareness to take the (possible) dissimilarity from the same patient into account.

Self-Supervised Learning for EF Prediction This part aims to learn meaningful representations from unlabeled data to estimate EF using echocardiograms. To this end, we propose an SSL scheme for M-mode images based on contrastive learning, where M-mode images from the same patient can naturally serve as positive pairs since they share labels for many downstream tasks. As discussed by [30], bio-signal data is inherently highly heterogeneous; thus, when applying learning-based methods to patient data, we need to consider both the similarity and the difference between samples originating from the same patient. Thus, we propose a problem-specific contrastive loss with patient and structure awareness, as shown in Figure 2.
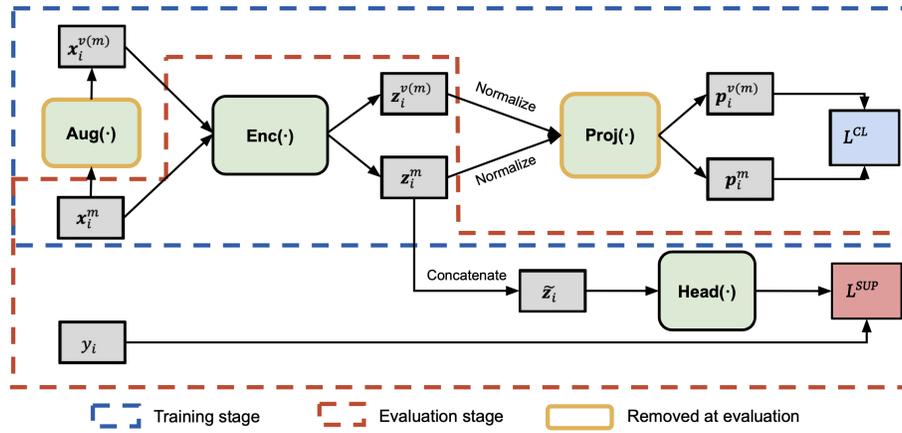
Fig. 3: Schema of the contrastive learning framework with training and evaluation stages. The training stage exploits the contrastive loss to learn a representation leveraging the unlabelled images. The evaluation stage exploits these learned representations in a supervised manner to predict EF.

Contrastive Learning Framework The framework contains training and evaluation stages and the overview is illustrated in Figure 3. In the training stage, we optimize the model with the contrastive loss leveraging the information from underlying structures of the unlabeled images. In the evaluation stage, a multi-layer perceptron (MLP) head is trained on top of the learned representations in a supervised manner.

For each generated M-mode image $\boldsymbol{x}_i^m$, we generate its augmented view $\boldsymbol{x}_i^{v(m)}$ using the $Aug(\cdot)$ module. So the augmented dataset is represented as $\{(\boldsymbol{x}_i^m,\ \boldsymbol{x}_i^{v(m)},\ y_i)\}$. The encoder network $Enc(\cdot)$ maps each image $\boldsymbol{x}_i^m$ to a feature vector $\boldsymbol{z}_i^m$. We utilize a standard ResNet architecture [9].

In the training stage, $\boldsymbol{z}_i^m$ is normalized to the unit hyper-sphere before being passed to the projection network. Following the work [4], we introduce a learnable non-linear projection network between the representation and the contrastive loss. The projection network $Proj(\cdot)$ takes the normalized lower-level representation $\boldsymbol{z}_i^m$ as input and outputs the higher-level representation $\boldsymbol{p}_i^m$. We use a two-layer MLP with ReLU activation as $Proj(\cdot)$ in this work.

In the evaluation stage, we initialize the parameters of the encoder network $Enc(\cdot)$ with the model obtained from contrastive learning and add an MLP head $Head(\cdot)$ to the top. For each patient i, we have M feature vectors $\boldsymbol{z}_i^m \in \mathbb{R}^K$. The M vectors are then fused to get the joint representation $\tilde{\boldsymbol{z}}_i \in \mathbb{R}^{K \times M}$ and passed to $Head(\cdot)$. One can have different fusion methods for aggregating information among the M vectors, e.g. using an LSTM cell [11], averaging, or concatenating.

Contrastive Loss for M-mode Images To account for (dis)similarities, we design two loss functions for learning both patient- and structure-awareness.

(a) Patient-aware loss: The goal is to attract the representations from the same patient to be similar while pushing apart representations from different patients (see Figure 2 (a)). This enforces two M-mode images to be considered similar if they are from the same patient and dissimilar if they are from different patients. The patient-aware loss is given as:

$$L^{PA} = -\frac{1}{M-1} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{l \neq m} \log \frac{\exp(\boldsymbol{p}_i^m \cdot \boldsymbol{p}_i^l / \tau)}{\sum_{j,k} \exp(\boldsymbol{p}_i^m \cdot \boldsymbol{p}_j^k / \tau) - \exp(\boldsymbol{p}_i^m \cdot \boldsymbol{p}_i^m / \tau)} \quad (1)$$

where N is the number of patients in one batch, M is the number of original M-mode images used for each patient, and τ is the temperature scaling parameter. The term $\boldsymbol{p}_i^m$ represents the output of $Proj(\cdot)$.

Inspired by [30], we tried defining a neighborhood function to limit the similarity of M-mode images from the same patient. However, incorporating neighbourhood to patient-awareness did not further improve the results; thus, we used all M-mode images per patient to define the patient-aware loss.

(b) Structure-aware loss: If we only use patient-aware loss L^{PA}, there exists a risk that all images from the same patient collapse to a single point [30]. So we propose the structure-aware loss to introduce some diversity (see Figure 2 (b)). To incorporate this into the learned representations, we construct positive pairs from each M-mode image with its augmentation and consider other combinations as negative pairs. It is then defined as:

$$L^{SA} = -\sum_{i=1}^{N} \sum_{m=1}^{2M} \log \frac{\exp(\boldsymbol{p}_i^m \cdot \boldsymbol{p}_i^{v(m)} / \tau)}{\sum_{l \neq m} \exp(\boldsymbol{p}_i^m \cdot \boldsymbol{p}_i^l / \tau)} \quad (2)$$

If image m is an original image, then $v(m)$ represents its augmented view; if image m is an augmented image, then $v(m)$ represents the original image. Minimizing L^{SA} drives the representation pairs from the augmented images in the numerator close while pushing the representations in the denominator far away, where the denominator contains M-mode images from the same patient.

Finally, we combine the two losses to get structure-aware and patient-aware contrastive loss for M-mode images using the hyperparameter α to control the trade-off between the awareness terms:

$$L^{CL} = \alpha L^{PA} + (1 - \alpha) L^{SA}. \quad (3)$$

3 Experiments and Results

3.1 Dataset

We use the publicly available EchoNet-Dynamic dataset [20]. It contains $10'030$ apical-4-chamber echocardiography videos from individuals who underwent imag-

ing between 2016-2018 as part of routine clinical care at Stanford University Hospital. Each B-mode video was cropped and masked to remove information outside the scanning sector and downsampled into standardized 112×112 pixel videos. For simplicity, we used videos with at least 112 frames. We use the official splits with 7465 training, 1289 validation, and 1282 test set samples.

3.2   Experimental Setup

We evaluate the models' performance using classification accuracy for five random seeds and report the mean performance and standard deviation. During training, all supervised models optimize the estimation of EF as a regression task. For testing, we use a constant threshold τ for classifying cardiomyopathy. In all experiments, we set $\tau = 0.5$. Hence, an estimation of $\hat{\tau} < 0.5$ results in classifying a sample as cardiomyopathic.

We evaluate all models using the area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC) with respect to whether a patient is correctly classified as healthy or cardiomyopathic. Additionally, we report the mean absolute error (MAE) and the root mean squared error (RMSE) of the predicted EF with respect to the true EF in the Supplementary Material. We report the mean performance, including standard deviations over five random seeds for all results.

We use the training set from EchoNet for pre-training (SSL), and apply a linear learning rate scheduler during the first 30 epochs as warm-up. For the supervised fine-tuning, we select different proportions of the training set in the limited labeled data scenario. All M-mode models are trained for 100 epochs using Adam optimizer [12] with an initial learning rate of 0.001 and a batch size of 64. For image augmentation, we apply random horizontal flip and Gaussian noise. For the fusion method of the the M-mode representations we used concatenation. For the EchoNet model, we use the same model and parameters as in [21]. The model is trained for 45 epochs with a learning rate of 0.0001 and a batch size of 20. We do not use test-time augmentation for any of the models. We report the full set of hyperparameters used in our experiments in Table 1.

3.3   Results and Discussion

Evaluating M-mode Images in Supervised Setting We train and evaluate models with different numbers of M-modes for $M \in \{1, 2, 5, 10, 20, 50\}$. We use the complete training set, including labels, as we are interested in the performance of the models depending on the number of available M-modes. Figure 4 shows the results for different numbers of M-modes. We see that late fusion models benefit from an increasing number of modes, whereas the early fusion method overfits quickly and never achieves a comparable performance.

Evaluating Limited Data Regime We evaluate the accuracy of the different models introduced in Section 2 for different amount of labeled training samples.

(a) AUPRC

(b) AUROC

(c) RMSE

(d) MAE

(e) R^2
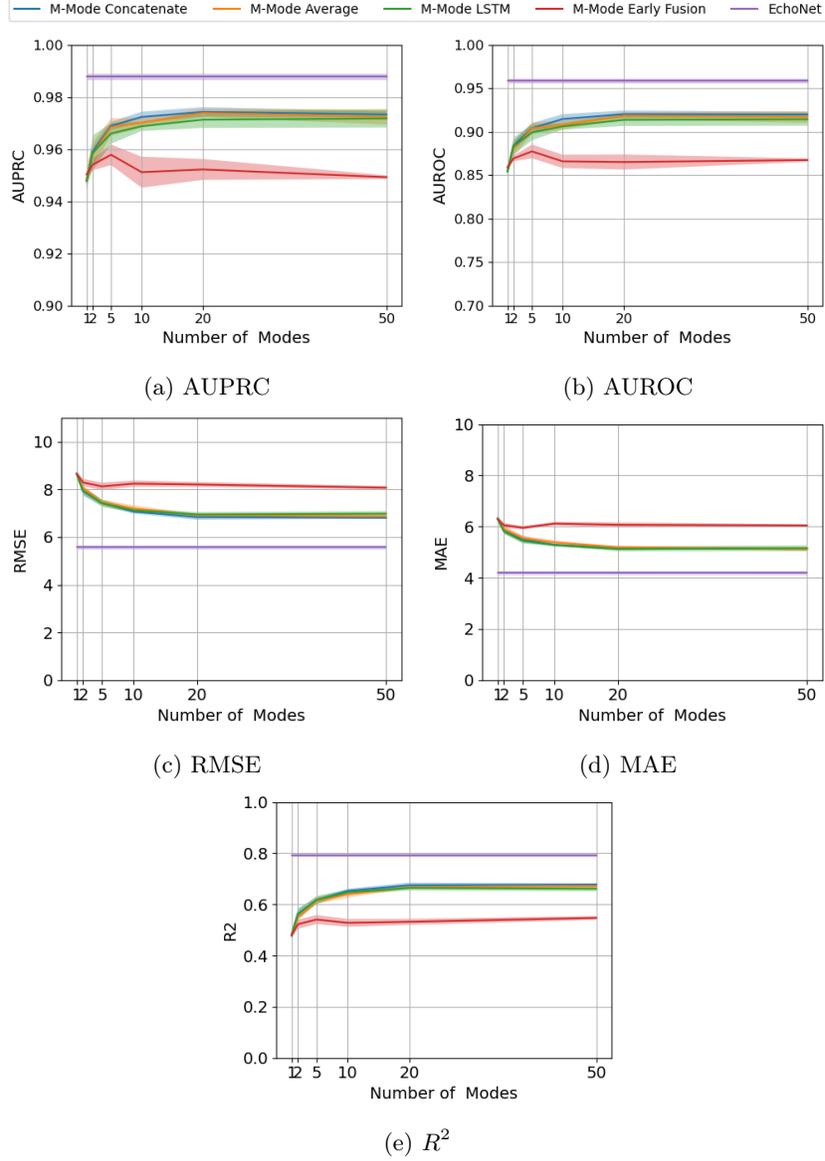
Fig. 4: Performance for different numbers of M-mode images using early and late-fusion methods. In (a), we evaluate the classification performance with respect to AUPRC and AUROC in (b), the regression performance with respect to RMSE in (c), MAE in (d), and R^2-score in (e).

Table 1: List the hyperparameters used in our experiments. We use the same hyper-parameters for E2E setup and the fine-tuning stage of SSL setup (denoted as "_sup" in Table 1). "_cl" denotes the hyper-parameters used in the SSL pre-training stage.

Parameter	Value	Description
lr_sup	0.001	learning rate for supervised training
lr_cl	1.0	learning rate for SSL training
opt	Adam	optimizer for SSL and supervised training
bsz_sup	64	batch size for supervised training
bsz_cl	256	batch size for SSL training
epoch_sup	100	epochs for supervised training
epoch_cl	300	epochs for SSL training
epoch_warm	30	warm-up epochs for SSL training
α	0.8	loss trade-off
τ	0.01	temperature scaling
Dim_e	512	$Enc(\cdot)$ output dimension
Dim_ph	2048	$Proj(\cdot)$ hidden layer dimension
Dim_po	128	$Proj(\cdot)$ output dimension
Dim_lstm	256	LSTM output dimension

As most medical datasets do not have the size of EchoNet-Dynamic [13], methods for medical machine learning should perform best in the limited labeled data regime. We use *E2E* for the supervised and *CL* for the self-supervised setting.

Additionally, we introduce *E2E+* and *CL+*, which, inspired by EchoNet [21], uses random short clips for each training epoch. Both models use M-mode images of 32 frames with a sampling period of 2. We train and evaluate models using $p\%$ of the full training set for $p \in \{1, 2, 3, 5, 10, 20, 30, 50, 75, 100\}$. All M-mode methods are trained with $M = 10$.

Figure 5 shows the limited labeled data experiment results. Although we are not able to reach the performance of the EchoNet model for any number of modes (see Figure 4b) if the number of labeled training samples is high (see Figure 5a), both supervised and self-supervised learning methods using M-mode instead of B-mode can outperform the EchoNet model in the low labeled data regime ($p < 5\%$, Figure 5b). Also, we observe that using shorter clips is useful for the self-supervised learning methods, with *CL+* being able to achieve an AUROC over 0.85 with only around 200 labeled samples.

Computational Cost Furthermore, we compare the number of parameters and computational costs for different models in Table 2, where we used a multi-GPU setup with four NVIDIA GeForce RTX 2080 Ti GPUs. We report the computation time in seconds per batch (sec/B) and milliseconds per sample (msec/sample), and the memory requirements in gigabytes per batch (GB/B).

Our proposed M-mode image based models require around six times less time and ten times less memory to train and run inference per sample. Given the used
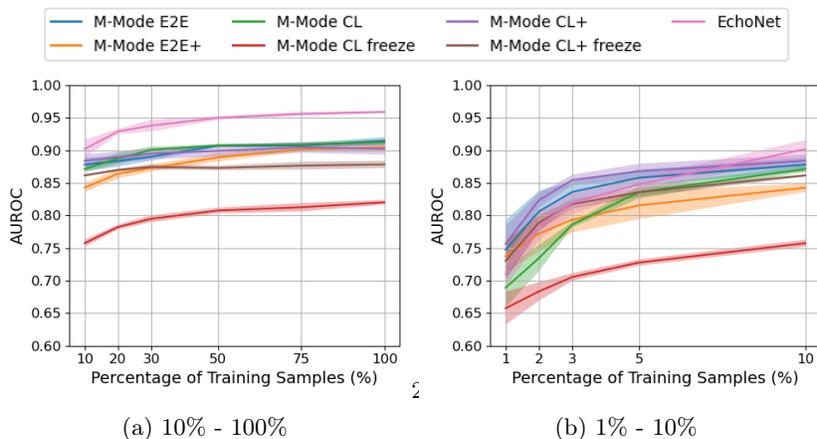
Fig. 5: Results for different training set sizes using the proposed end-to-end supervised (E2E) and contrastive learning (CL) approaches. In (a), we train and evaluate the models on 10%-100% labeled training samples, in (b) only on 1%-10% of the samples. E2E and CL models are trained using a fixed long clip with length 112; E2E+ and CL+ are trained using random short clips with length 32. CL freeze and CL+ freeze are fine-tuned with the encoder parameters frozen.

memory per batch, we could increase the batch size for the M-mode methods, lowering the computation time per sample even further, whereas the baseline model is already at the limit due to its architecture.

4   Discussion and Conclusion

In this work, we propose to generate M-mode images from readily available B-mode echocardiography videos and fuse these to estimate EF and, thus, cardiac dysfunction. Our results show that M-mode-based prediction methods are comparable to the baseline method while avoiding its complex training routine and reducing the computational cost and the need for expensive expert input.

Table 2: Computational costs. We evaluate the EchoNet and the proposed M-mode methods with respect to the number of parameters, the computation time, and the memory requirements. All M-mode models are evaluated using $M = 10$. E2E defines the end-to-end supervised and CL the contrastive learning approach.

Model	BS	#Params (Mio.)	Time (sec/B)		Time (msec/sample)		Memory (GB/B)	
			Train	Test	Train	Test	Train	Test
EchoNet	20	31.5	2.898	2.474	144.9	123.7	5.294	1.187
E2E & CL	64	11.7	1.568	1.330	24.5	21.1	1.013	0.120

Conventional M-mode images have a very high sampling rate, which results in a high temporal resolution so that even very rapid motion can be recorded. The generated M-mode images have significantly less temporal resolution than the conventional M-mode images from US machines. However, our results indicate that exploiting generated M-mode images does not limit the performance for EF estimation. As we do not use the M-mode images collected directly from the US machines, there is no need for an additional data collection step.

Additionally, we show the potential of pre-trained methods. In scenarios where expensive expert labels are not readily available, pre-training using unlabeled M-mode images outperforms more complicated pipelines highlighting the potential of M-Mode based pipelines for clinical use cases. In our future work, we want to investigate the use cases for M-mode on different diseases and further improve the performance of the proposed pre-training pipeline.

5   Acknowledgements

Bibliography

[1] Avila, J., Smith, B., Mead, T., Jurma, D., Dawson, M., Mallin, M., Dugan, A.: Does the Addition of M-Mode to B-Mode Ultrasound Increase the Accuracy of Identification of Lung Sliding in Traumatic Pneumothoraces? Journal of Ultrasound in Medicine **37**(11), 2681–2687 (2018)

[2] Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2018)

[3] Bamira, D., Picard, M.H.: Imaging: Echocardiology–Assessment of Cardiac Structure and Function. In: Encyclopedia of Cardiovascular Research and Medicine, pp. 35–54. Elsevier (2018)

[4] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607 (2020)

[5] Dai, W., Li, X., Chiu, W.H.K., Kuo, M.D., Cheng, K.T.: Adaptive Contrast for Image Regression in Computer-Aided Disease Assessment. IEEE Transactions on Medical Imaging **41**(5), 1255–1268 (5 2022)

[6] Devereux, R.B., Lutas, E.M., Casale, P.N., Kligfield, P., Eisenberg, R.R., et al.: Standardization of M-mode echocardiographic left ventricular anatomic measurements. J Am Coll Cardiol. **4**(6), 1222–1230 (1984)

[7] Gaspar, H.A., Morhy, S.S., Lianza, A.C., de Carvalho, W.B., Andrade, J.L., et al.: Focused cardiac ultrasound: a training course for pediatric intensivists and emergency physicians. BMC Medical Education **14**(1) (2014)

[8] Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., et al.: Deep learning interpretation of echocardiograms. npj Dig Med (2020)

[9] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[10] Hensel, K.O., Roskopf, M., Wilke, L., Heusch, A.: Intraobserver and interobserver reproducibility of M-mode and B-mode acquired mitral annular plane systolic excursion (MAPSE) and its dependency on echocardiographic image quality in children. PLOS ONE **13**(5), e0196614 (2018)

[11] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)

[12] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[13] Kiryati, N., Landau, Y.: Dataset growth in medical image analysis research. Journal of imaging **7**(8), 155 (2021)

[14] Kulhare, S., Zheng, X., Mehanian, C., Gregory, C., Zhu, M., Gregory, K., et al.: Ultrasound-Based Detection of Lung Abnormalities Using Single Shot Detection Convolutional Neural Networks. In: Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation (2018)

[15] Lagopoulos, A., Hristu-Varsakelis, D.: Measuring the Left Ventricular Ejection Fraction using Geometric Features. In: IEEE International Symposium on Computer-Based Medical Systems. pp. 1–6. IEEE (7 2022)

[16] LeCun, Y., Misra, I.: Self-supervised learning: The dark matter of intelligence. Meta AI **23** (2021)

[17] Madani, A., Ong, J.R., Tibrewal, A., Mofrad, M.R.K.: Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. npj Digital Medicine **1**(1) (2018)

[18] Mehanian, C., Kulhare, S., Millin, R., Zheng, X., Gregory, C., Zhu, M., et al.: Deep Learning-Based Pneumothorax Detection in Ultrasound Videos. pp. 74–82 (2019)

[19] Muhtaseb, R., Yaqub, M.: EchoCoTr: Estimation of the LV Ejection Fraction from Spatiotemporal Echocardiography. pp. 370–379 (2022)

[20] Ouyang, D., He, B., Ghorbani, A., Lungren, M.P., Ashley, E.A., et al.: Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In: NeurIPS ML4H Workshop (2019)

[21] Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., et al.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)

[22] Sarkar, P.G., Chandra, V.: A Novel Approach for Detecting Abnormality in Ejection Fraction Using Transthoracic Echocardiography with Deep Learning. Int J of Online and Biomed Eng **16**(13), 99 (2020)

[23] Saul, T., Siadecki, S.D., Berkowitz, R., Rose, G., Matilsky, D., Sauler, A.: M-Mode Ultrasound Applications for the Emergency Medicine Physician. The Journal of Emergency Medicine **49**(5), 686–692 (2015)

[24] Shurrab, S., Duwairi, R.: Self-supervised learning methods and applications in medical imaging analysis: A survey. PeerJ Comp Sci **8**, e1045 (2022)

[25] Singh, A.K., Mayo, P.H., Koenig, S., Talwar, A., Narasimhan, M.: The Use of M-Mode Ultrasonography to Differentiate the Causes of B Lines. Chest **153**(3), 689–696 (2018)

[26] Skinner, H., Kamaruddin, H., Mathew, T.: Tricuspid Annular Plane Systolic Excursion: Comparing Transthoracic to Transesophageal Echocardiography. Journal of Cardiothoracic and Vascular Anesthesia **31**(2), 590–594 (2017)

[27] Tian, Y., Xu, S., Guo, L., Cong, F.: A Periodic Frame Learning Approach for Accurate Landmark Localization in M-Mode Echocardiography. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2021)

[28] Tromp, J., Seekings, P.J., Hung, C.L., Iversen, M.B., Frost, M.J., et al.: Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. The Lancet Digital Health **4**(1) (1 2022)

[29] WHO: Cardiovascular diseases (CVDs). https://www.who.int/newsroom/fact-sheets/detail/cardiovascular-diseases-(cvds) (2022)

[30] Yèche, H., Dresdner, G., Locatello, F., Hüser, M., Rätsch, G.: Neighborhood contrastive learning applied to online patient monitoring. In: International Conference on Machine Learning. pp. 11964–11974 (2021)

[31] Zhang, J., Gajjala, S., Agrawal, P., Tison, G.H., Hallock, L.A., et al.: Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation **138**(16), 1623–1635 (2018)