Self-supervised Learning to Predict Ejection Fraction using Motion-mode Images

Conference Paper

Author(s):
Hu, Yurong; Sutter, Thomas M.; Oezkan, Ece; Vogt, Julia E.

Publication date:
2023-03-20

Permanent link:
https://doi.org/10.3929/ethz-b-000648766

SELF-SUPERVISED LEARNING TO PREDICT EJECTION FRACTION USING MOTION-MODE IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Data scarcity is a fundamental problem since data lies at the heart of any ML project. For most applications, annotation is an expensive task in addition to data collection. Thus, learning from limited labeled data is very critical for data-limited problems, such as in healthcare applications, to have the ability to learn in a sample-efficient manner. Self-supervised learning (SSL) can learn meaningful representations from exploiting structures in unlabeled data, which allows the model to achieve high accuracy in various downstream tasks, even with limited annotations. In this work, we extend contrastive learning, an efficient implementation of SSL, to cardiac imaging. We propose to use generated M(otion)-mode images from readily available B(rightness)-mode echocardiograms and design contrastive objectives with structure and patient-awareness. Experiments on EchoNet-Dynamic show that our proposed model can achieve an AUROC score of 0.85 by simply training a linear head on top of the learned representations, and is insensitive to the reduction of labeled data.

1 INTRODUCTION

Early assessment of cardiac dysfunction with routine screening is essential for diagnosing cardiovascular diseases, the leading cause of death worldwide (WHO, 2022). 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 -diastolic volumes (Bamira & Picard, 2018; Ouyang et al., 2020). Echocardiography is a widely-adopted imaging modality, with ultrasound being a low-cost, real-time, and non-ionizing technology (Sarkar & Chandra, 2020). However, the manual evaluation of echocardiograms is an expensive and operator-dependent task; thus, there has been a clear interest in automated EF prediction methods. Some recent works have applied deep learning techniques to EF prediction using echocardiograms (Sarkar & Chandra, 2020; Tian et al., 2021; Madani et al., 2018; Ghorbani et al., 2020; Mehanian et al., 2019), which exploit either still-images or spatio-temporal convolutions relying on fully labelled data. However, data collection and annotation are expensive for most applications, such as in healthcare. Therefore, learning from limited labeled data plays a key role in data-limited problems. To overcome this data bottleneck, self-supervised learning (SSL) methods have been proposed, which aim to learn meaningful high-level representations from unlabeled data (LeCun & Misra, 2021; Shurrab & Duwairi, 2022).

Our contribution In this work, we propose an SSL scheme for predicting EF using echocardiograms through extending contrastive learning, an efficient implementation of SSL. Instead of using conventional B-mode videos, we leverage generated M-mode images ((Avila et al., 2018; Singh et al., 2018)) as the input modality. Sutter et al. (2022) recently showed the effectiveness of M-mode images for assessing cardiac dysfunction while bypassing larger 3D models. For the contrastive learning part, M-mode images from the same patient can then naturally serve as positive pairs since they share labels for many downstream tasks. As discussed by (Yèche et al., 2021), 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. To remedy this problem, we design a ContrAstive Loss for M-mode images (CALM) to learn unsupervised representations with structure and patient awareness. We evaluate the learnt representation on the publicly available EchoNet-Dynamic dataset ((Ouyang et al., 2020)) and demonstrate the robustness of our models in the limited labeled-data scenario.

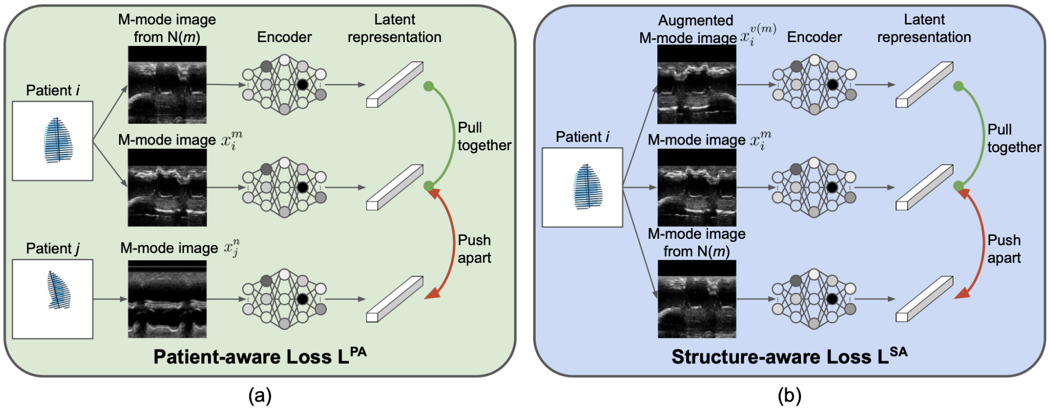

Figure 1: Overview of our proposed CALM method. The contrastive loss includes (a) patient awareness to attract similarity between data from the same patient and to discourage between different patients (b) structure awareness to take the (possible) dissimilarity from the same patient into account.

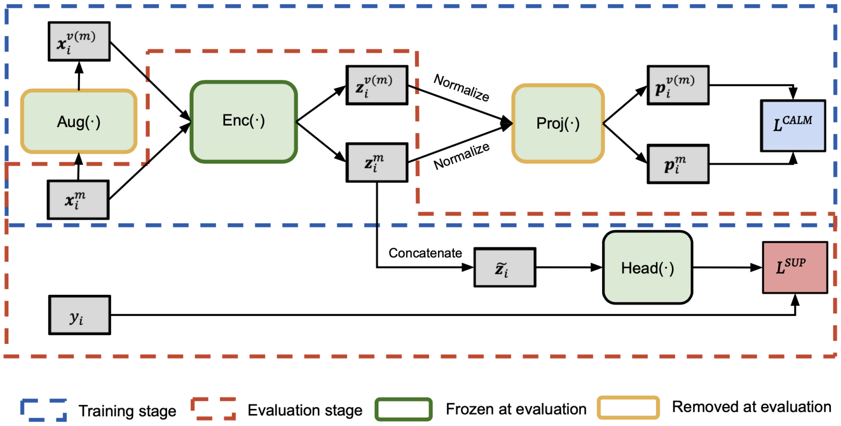

Figure 2: Schema of the contrastive learning framework with training and evaluation stages. The training stage exploits the contrastive loss to learn a representation leveraging the unlabelled images. The evaluation stage exploits these learned representations in a supervised manner to predict EF.

2 Method

This work 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, while extending it with patient and structure awareness as shown in Figure 1.

Contrastive Learning Framework It contains training and evaluation stages, as shown in Figure 2. In the training stage, the model is trained with the contrastive loss leveraging the information from underlying structures of the unlabeled images. In the evaluation stage, a multilayer perceptron (MLP) head is trained on top of the learned representations in a supervised manner. Assume our dataset contains N patients. For each patient $i = \{1, 2, \cdots, N\}$, the label y_i indicates its EF. We then generate M M-mode images $\boldsymbol{x}_i^m$ with $m = \{1, 2, \cdots, M\}$ with evenly-spaced angles from a single B-mode echocardiogram for each patient i. For more information, please refer to Appendix A. Furthermore, for each $\boldsymbol{x}_i^m$, 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 $\boldsymbol{z}_i^m$.

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 from (Chen et al., 2020), 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 $\boldsymbol{p}_i^m$.

In the evaluation stage, we freeze the encoder network $Enc(\cdot)$ 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 concatenated to get the joint representation $\tilde{\boldsymbol{z}}_i \in \mathbb{R}^{K \times M}$ for patient i, which is the input of $Head(\cdot)$. $Head(\cdot)$ is trained with labeled data using supervised loss.

ContrAstive Loss for M-mode images (CALM) To account for (dis)similarities from M-mode images, we design two loss functions for learning both patient- and structure-aware representations.

Patient-aware loss: The goal of the loss is to attract the representations from the same patient to be similar while pushing apart representations from different patients (see Figure 1 (a)). Inspired by Yèche et al. (2021), we define the neighborhood function $n(\boldsymbol{x}_i^m, \boldsymbol{x}_j^l) = (\mathbb{1}[i = j] \times \mathbb{1}[|m - l| \leq \lambda])$, which enforces two M-mode images to be considered neighbors if they are from the same patient and the distance of their mode ordinal numbers are within a certain threshold λ. Here we introduce a neighborhood threshold λ because M-mode images with large angle distance may contain quite different structural information, and we should not simply consider them from the same neighborhood. The neighborhood of the m-th M-mode image from patient i is defined as $N(i, m) = \{l \neq m | n(\boldsymbol{x}_i^m, \boldsymbol{x}_j^l) = 1\}$. Using the definition of neighborhood function, $N(i, m) = \{l \neq m | |m - l| \leq \lambda\} \triangleq N(m)$, the patient-aware loss is given as:

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

Structure-aware loss: If we only use patient-aware loss L^{PA}, there exists a risk that all images from the same neighborhood collapse to a single point (Yèche et al., 2021). So we propose the structure-aware loss, to introduce some diversity among neighbors (see Figure 1 (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:

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

where N is the number of patients in one batch, M is the number of M-mode images used for each patient, and τ is the temperature scaling parameter. As shown in Figure 2, $\boldsymbol{p}_i^m$ represents the output of $Proj(\cdot)$. 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 but from different modes in the neighbourhood.

Finally, we combine the two losses to get ContrAstive Loss for M-mode images (CALM). The hyper-parameter α is used to control the trade-off between patient and structure awareness:

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

3 Experiments and Results

Dataset We use the publicly available EchoNet-Dynamic (Ouyang et al., 2020) dataset containing $10,030$ apical-4-chamber echocardiography videos provided by Stanford University Hospital. Each video was processed to 112×112 pixel grayscale image sequences. In addition to the videos, the dataset provides the clinical measurement of the EF of LV for each patient. For each echocardiogram video, we extract $M = 50$ M-mode images. For convenience, we use a fixed video length $T = 112$. We apply the combination of random horizontal flip and Gaussian noise as data augmentations. The dataset split is the same as original (except that shorter ones (less than 112 frames) are discarded): $6,966$ videos in the training set, $1,230$ videos in the validation set, and $1,190$ videos in the test set.

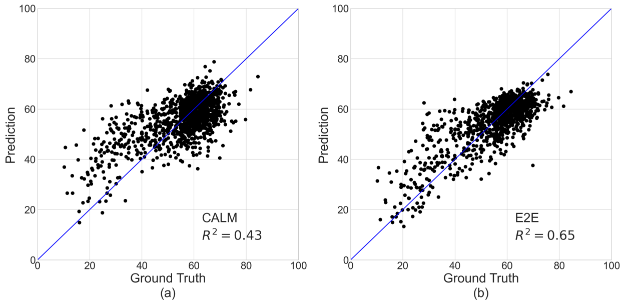

Figure 3: Estimated EF compared to actual EF for (a) CALM and (b) E2E models. R^2 scores for both models are also reported.

Labels	MSE		MAE		AUROC		AUPRC	
	E2E	CALM	E2E	CALM	E2E	CALM	E2E	CALM
100%	47.2±2.0	79.8±5.9	5.2±0.1	6.8±0.3	0.92±0.00	0.85±0.02	0.97±0.00	0.95±0.01
70%	50.5±1.2	82.4±6.3	5.3±0.1	6.9±0.3	0.91±0.00	0.85±0.01	0.97±0.00	0.95±0.00
50%	55.3±0.7	84.4±6.0	5.5±0.1	7.0±0.2	0.90±0.01	0.85±0.01	0.97±0.00	0.95±0.00
30%	59.3±0.6	87.8±5.5	5.8±0.0	7.2±0.2	0.89±0.00	0.84±0.01	0.96±0.00	0.94±0.00
10%	70.0±1.3	100.81±7.1	6.3±0.1	7.7±0.3	0.87±0.00	0.80±0.02	0.96±0.00	0.93±0.01

Table 1: Model performance on EchoNet-Dynamic validation set. AUROC and AUPRC are calculated with EF threshold 50. For each metric, mean ± std of 3 runs are reported.

Experiments We evaluate the proposed framework with two models: (i) CALM model is trained with 2-stage contrastive learning; (ii) E2E has the same architecture as CALM model in the evaluation stage, but is trained end-to-end in a supervised manner. We report the performance of the models using classification accuracy for three random seeds. The predictions of EF are shown in Figure 3. Both models generate predictions close to the ground truth, with CALM reaching $R^2 = 0.43$ and E2E achieving $R^2 = 0.65$. We set an EF threshold to 50 and use the EF predictions to detect potential heart failures, e. g. cardiomyopathy. CALM achieves mean AUROC score of 0.85 and AUPRC score of 0.95, whereas E2E 0.92 and 0.97, respectively. Note that our E2E model achieves similar performance with (Sutter et al., 2022), which leverages CNN and LSTM to predict EF from M-mode images, and achieves an AUROC score of 0.89 and AUPRC score of 0.96. We also evaluate the performance in the limited labeled-data scenario. As shown in Table 1, we gradually reduce the fraction of labeled training data from 100% to 10%, and observe only small degradation on all the metrics for both of the models (less than 0.02 for AUROC and AUPRC).

4 Discussion and Conclusion

In this work, we proposed a contrastive learning scheme tailored for predicting EF from M-mode images, where we leveraged the trade-off between structure-aware loss and patient-aware loss to tackle the heterogeneity problem in patient data. Furthermore, we showed that M-mode echocardiography is a good modality for learning SSL representation and predicting cardiac function. Admittedly, it is surprising that the end-to-end model consistently performs better than the 2-stage contrastive framework, even in the limited labeled-data scenario. There are several reasons that may explain these results: (i) the videos in the EchoNet-Dynamic dataset are well aligned, which can make it an easier task for the supervised learning; (ii) the dataset for SSL is not large enough only containing around 7k training images compared to 50k datapoints e. g. in (Yèche et al., 2021), iii) M-mode images contain time information compared to common images in contrastive learning settings. However, both structure- and patient-aware losses only deal with variations at the spatial, but not at the temporal level. Overall, our work provides an insight of using SSL to learn representations from M-mode echocardiograms. We believe that enriching the contrastive loss with other objectives to learn time-aware representation is a promising direction for future work.

REFERENCES

Jacob Avila, Ben Smith, Therese Mead, Duane Jurma, Matthew Dawson, Michael Mallin, and Adam Dugan. 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.

D Bamira and M H Picard. Imaging: Echocardiology–Assessment of Cardiac Structure and Function. In *Encyclopedia of Cardiovascular Research and Medicine*, pp. 35–54. Elsevier, 2018. doi: 10.1016/b978-0-12-809657-4.10953-6.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607, 2020.

Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H Chen, Robert A Harrington, David H Liang, Euan A Ashley, and James Y Zou. Deep learning interpretation of echocardiograms. *npj Digital Medicine*, 3(1), 2020.

Yann LeCun and Ishan Misra. Self-supervised learning: The dark matter of intelligence. *Meta AI*, 23, 2021.

Ali Madani, Jia Rui Ong, Anshul Tibrewal, and Mohammad R K Mofrad. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *npj Digital Medicine*, 1(1), 2018.

Courosh Mehanian, Sourabh Kulhare, Rachel Millin, Xinliang Zheng, Cynthia Gregory, Meihua Zhu, Hua Xie, James Jones, Jack Lazar, Amber Halse, Todd Graham, Mike Stone, Kenton Gregory, and Ben Wilson. Deep Learning-Based Pneumothorax Detection in Ultrasound Videos. In *mart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis*, pp. 74–82, 2019. doi: 10.1007/978-3-030-32875-7{_}9.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, and James Y Zou. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

Prattay Guha Sarkar and Vishal Chandra. A Novel Approach for Detecting Abnormality in Ejection Fraction Using Transthoracic Echocardiography with Deep Learning. *International Journal of Online and Biomedical Engineering (iJOE)*, 16(13):99, 2020.

Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022.

Anup K Singh, Paul H Mayo, Seth Koenig, Aranabh Talwar, and Mangala Narasimhan. The Use of M-Mode Ultrasonography to Differentiate the Causes of B Lines. *Chest*, 153(3):689–696, 2018.

Thomas M. Sutter, Sebastian Balzer, Ece Ozkan, and Julia E. Vogt. M(otion)-mode Based Prediction of Cardiac Function on Echocardiograms. In *Workshop on Medical Imaging meets NeurIPS*. ETH Zurich, 2022.

Yinbing Tian, Shibiao Xu, Li Guo, and Fuze Cong. A Periodic Frame Learning Approach for Accurate Landmark Localization in M-Mode Echocardiography. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

WHO. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 2022.

Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*, pp. 11964–11974, 2021.

A Appendix

M(otion)-mode Generation Given an echocardiogram video consisting of T sequential images, each with a size of $H \times W$, we place a straight line from the top-middle to the bottom-middle of each image. By concatenating the lines throughout the temporal axis we get one M-mode image with size $H \times T$ (see Figure 4). Then we rotate the video sequences at different angles and repeat this procedure to get several M-mode images for each video (patient).

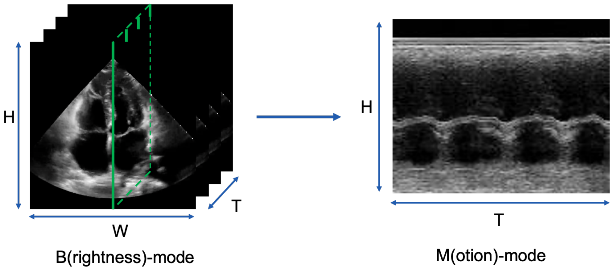

Figure 4: Converting B-mode video to M-mode images. The solid green vertical line is placed at the middle of each image. The dashed green box shows the extracted M-mode image with size $H \times T$.

Experimental Parameters Below we list the hyper-parameters used in the experiments.

Name	Value	Description
λ	50	neighborhood threshold
α	0.8	loss trade-off
τ	0.01	temperature scaling
lr	1.0	learning rate
bsz	256	training batch size
epoch	300	maximum training epochs
M	5	number of images per patient
D_e	2048	dimension of encoder output
D_p	128	dimension of projection output

Table 2: Hyper-parameters used in contrastive learning experiments.

Additional Experimental Results Below you can see the further experimental results.

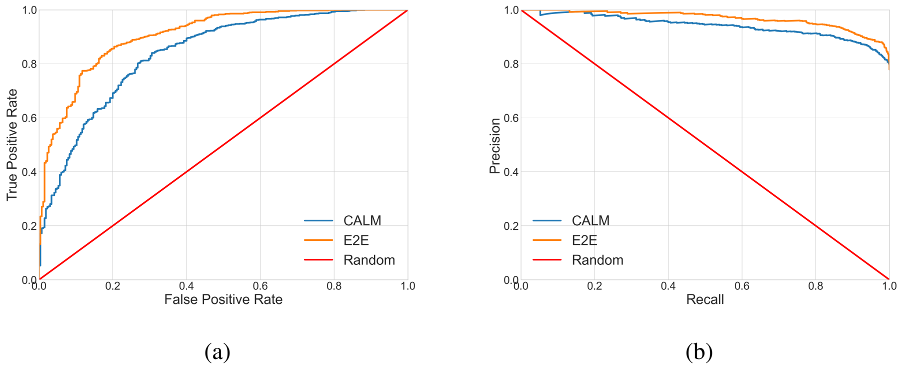

Figure 5: (a) Receiver operating characteristic and (b) precision-recall curves in the full labeled-data regime. The plots are generated from a run of seed 924.

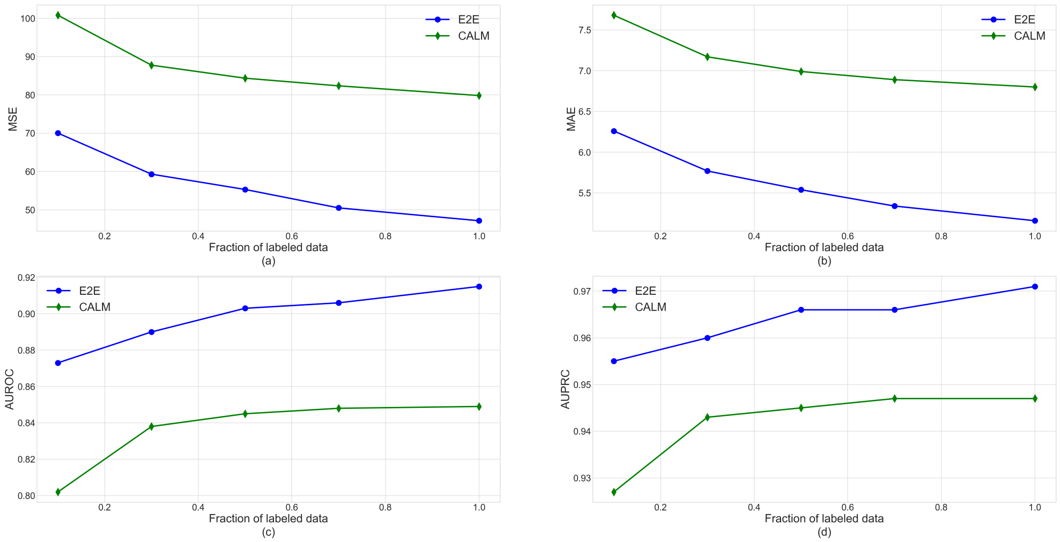

Figure 6: Influence of the percentage of labeled data on the model performance on the validation set.

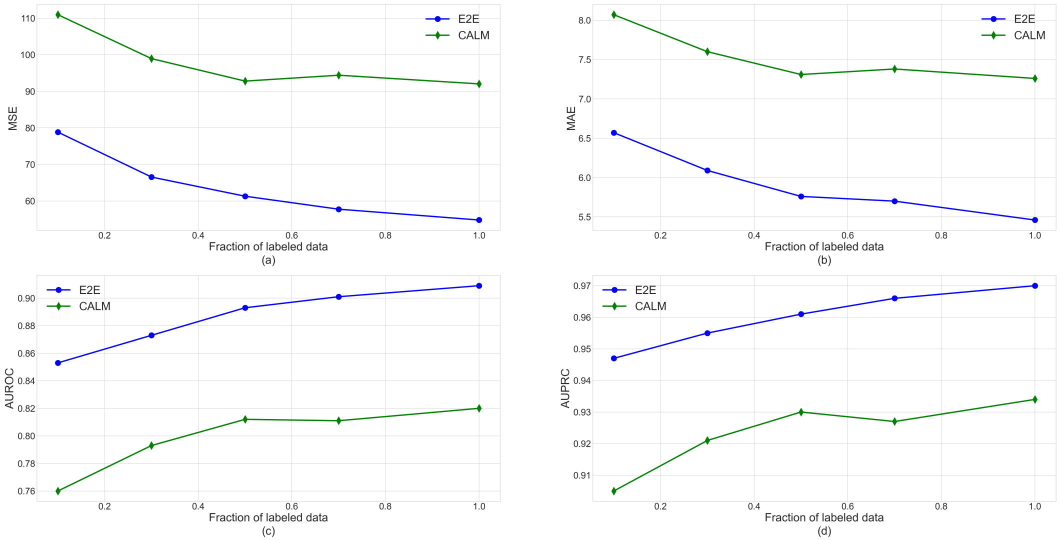

Figure 7: Influence of the percentage of labeled data on the model performance on the test set.