EchoGNN with Contrastive Learning

Nima Kondori
Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, BC V6T 1Z4
nimakondori96@gmail.com

Andrea Fung
School of Biomedical Engineering
University of British Columbia
Vancouver, BC V6T 1Z3
andrea.fung@alumni.ubc.ca

Jamie Alexis D. Goco
School of Biomedical Engineering
University of British Columbia
Vancouver, BC V6T 1Z3
jamiegoco@live.ca

Abstract

Ejection fraction (EF) is a key metric used to assess the systolic performance of the left ventricle of the heart. Within recent years, many deep learning models have been developed to help automate the estimation of EF from echocardiograms. Among these works includes EchoGNN, a model based on Graph Neural Networks (GNNs) from Mokhtari *et al.* that aimed to provide an explainable EF estimation. However, EchoGNN acts as a skeleton implementation of GNNs in EF estimation applications and can therefore be further developed and improved on. One such improvement involves further investigation into the randomness of the embeddings EchoGNN generates and whether the EF estimation accuracy would improve if EchoGNN is given better embeddings. Recent works using contrastive learning to produce more informed embeddings have shown a lot of success in many areas of deep learning including medical applications. Therefore, our project shows how the integration of contrastive learning in a semi-supervised manner increases EchoGNN's accuracy of EF estimation.

1 Introduction

Ejection fraction (EF) is defined as the percentage difference in left ventricular volume over one heartbeat. It is the most frequently used metric to assess the global systolic performance of the left ventricle (LV). EF is an important measurement that, when impaired, can help detect and monitor forward heart failure caused by a multitude of cardiac pathologies such as cardiomyopathy or ischemia. Typically in echocardiography (echo), EF is calculated by means of Modified Simpson's biplane, where clinicians trace the LV cavity in the apical four chamber (AP4) and apical two chamber (AP2) views at both end-diastole (ED) and end-systole (ES) of the cardiac cycle. Using these traces, LV volumes at both ED and ES are calculated using a method of disks and then a ratio is calculated to produce the EF.

As with most ultrasound applications, Modified Simpson's biplane is incredibly subjective and has high inter-observer variability in both the tracing of the LV cavity as well as selecting which frame ES and ED are located on [12]. Therefore, recent works in deep learning have developed models aimed to tackle this problem and help clinicians estimate EF from echo cine clips automatically. [4]

One such model is the EchoGNN [8]: a deep learning model that uses Graph Neural Networks (GNN) to estimate the EF of echo cine clips. The echo cine clip is first passed through a Video Encoder that

Submission for EECE 571F 2022 Winter Term 1.

converts each frame into an embedding. All embeddings are then combined into a complete graph structure called an echo-graph, to which an attention encoder then learns and assigns weights. A graph regressor network using GNN layers with the learned weights of the echo-graph is then used to produce an estimated EF. In addition to providing an estimated EF, EchoGNN has the added benefit of providing explainability through the learned weights of the echo-graph. EchoGNN seems to place larger weights on frames and edges between ED and ES, which corresponds to when the heart is contracting during the echo.

However, EchoGNN's weight distribution seems to contradict the clinical method of measuring ejection fraction. While clinicians estimate ejection fraction using volumes from the ED and ES frames in an echo video, EchoGNN routinely assigns smaller weights to these frames even though these frames are the most important when measuring EF. Our hypothesis is that this may occur due to the fact that the Video Encoder generates random embeddings without any clinical context or information about the cardiac cycle.

Adding contrastive loss to create more informed embeddings has been shown to be successful in many areas, including medical applications. Therefore, we are proposing to add a contrastive loss to EchoGNN's Video Encoder with the goal of preserving the temporal information associated with each frame's position in the cardiac cycle. By doing so, the model can create more informed embeddings based on the contextual similarity and dissimilarities of the input frames. The contrastive loss we implement aims to repel and distance dissimilar frames (ED and ES) in the resultant embedding space, while clustering together frames around the labeled ED and ES (neighbors). In addition to creating more informed embeddings, we hope that the addition of clustering neighboring frames together in our contrastive loss acts as a measure to address the aleatoric uncertainty associated with inter-observer variability of labeling ED and ES frames. Together, we hope that these benefits improve the EF estimation accuracy of EchoGNN.

2 Related Work

EchoGNN: Explainable Ejection Fraction Estimation with Graph Neural Networks [8] Using deep learning, EchoGNN produces an explainable EF estimation based on echocardiograms. The model encodes each frame of the echo cine clip into a node on a complete graph, then uses GNNs to learn and assign weights to each node and edge-based on its contribution towards predicting EF. A regressor network using GNN layers along with these weights then produces an estimation for EF.

Contrastive Learning In general, contrastive learning is used to form an embedding space in which similar sample pairs are close together and dissimilar sample pairs are far apart. Contrastive learning has been used specifically for clustering, a fundamental tool in unsupervised learning. For example, a paper by Li et al. [5] used a contrastive clustering network that constructed data pairs through data augmentations to learn more discriminative representations and better clustering assignments. However, unsupervised clustering algorithms often generalize poorly to highly complex real-world data, such as noisy ultrasound data. In addition, these unsupervised algorithms do not take advantage of relevant labels that are already routinely produced by experts with each ultrasound exam. In ultrasound studies, researchers have proposed other contrastive learning methods to create more meaningful image representations for ultrasound-related tasks [3], [2]. For example, since ED frames possess temporal and volume information that resembles other ED frames and differs from ES frames, Cheng et al.[2] proposed a volume contrastive network with a loss function that encouraged (1) attraction between ED frame embeddings across image views and patients, (2) attraction between ES frame embeddings across image views and patients, and (3) repulsion between ED and ES frame embeddings across views and patients. These changes facilitate feature fusion between AP2 and AP4 views, thereby enabling better feature representation for LV volume and EF estimation. However, Cheng et al.'s contrastive loss function only applies to two image frames from ultrasound videos, the ED and ES frames. Therefore, the network does not learn any temporal information on anatomical changes and movement from echocardiograms that may be important for LV volume or EF estimation. In addition, these networks with single image inputs do not account for labeling uncertainty caused by high intra- and inter-observer variability in echocardiogram interpretation. A paper by Basu et al.[1] applied contrastive learning to ultrasound videos, but the network was unsupervised and defined positive and negative data pairs solely based on temporal distance with respect to a set of randomly chosen image frames (anchors). To our knowledge, there is no semi-supervised contrastive learning method that attracts and repels embeddings based on anatomic and temporal information for entire ultrasound video sequences.

Our Contributions Our contributions are two-fold:

1. **Addition of contrastive loss to EchoGNN in a semi-supervised manner**. By doing so, the Video Encoder will generate more informed embeddings that take into account the temporal information associated with the frames' position in the cardiac cycle. We hope that by adding this clinical context to the model, the EF estimation accuracy of EchoGNN improves.
2. **Addressing the aleatoric uncertainty related to the high inter-observer variability** associated with labeling the ED and ES frames by clustering and attracting neighboring frames around ED and ES in the embedding space. We hope that by adding this method, we generate more robust embeddings that are more resistant to this uncertainty. To our knowledge, there are no other contrastive learning models that predict EF that take this aleatoric uncertainty into account and consider the inter-observer variability of determining the ES and ED frames. Therefore, our model will be the first to do so.

3 Model and method

3.1 Research idea

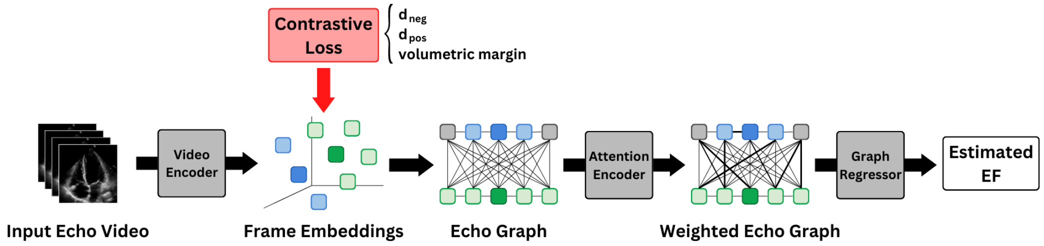

Figure 1: The architecture of the proposed model from [8] with contrastive loss addition.

As shown in Fig. 1, we are adding contrastive loss to the Video Encoder component of the EchoGNN model. Our goal is to make the initial embeddings more informed of the similarities between neighboring frames and the differences between ED and ES frames. An overview of the contrastive loss procedure is depicted in Fig. 2.

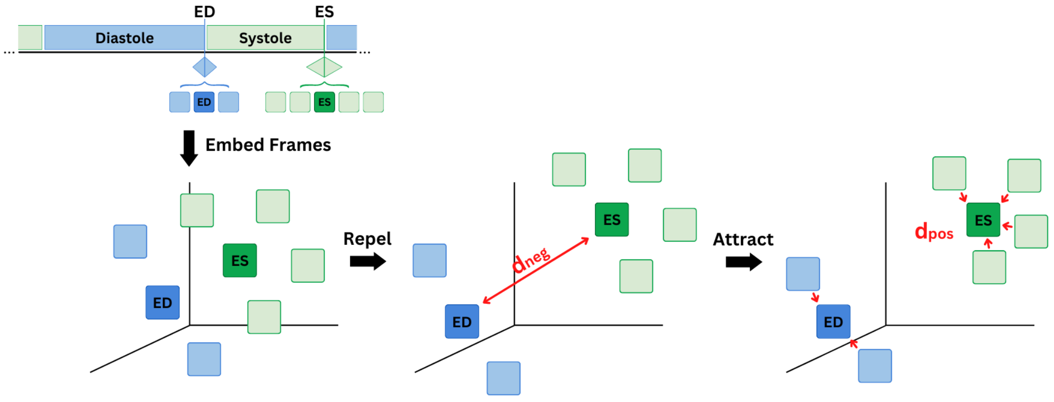

Figure 2: The suggested contrastive learning procedure attraction and repulsion of frame clusters.

Inspired by [2], we will be using the following loss function:

$$L = \sum_{i=1}^{N} d_{pos_i} - d_{neg_i} + \frac{EDV_i - ESV_i}{\beta} \tag{1}$$

where N is the number of samples in the batch, d_{pos} is the average distance between the anchors (labeled ED and ES frames) and positive samples (neighboring frames), and d_{neg} is the normalized distance between the anchor (labeled ED frame) and the negative sample (labeled ES frame). We also have the volumetric margin where EDV_i and ESV_i are the ED and ES volumes respectively, and β = 200 is the scaling hyperparameter.

To calculate the distances in our loss formula, we have:

$$d_{pos} = \frac{1}{2n} \sum_{j \in \{ED,ES\}} \sum_{i=1}^{n} ||\sigma_j - \sigma_i||_2 \tag{2}$$

where σ_i and σ_j are the normalized embeddings of the anchor and positive sample respectively and n is the number of neighboring frames. n corresponds either to the number of neighboring ED frames (δ) or the number of neighboring ES frames (ε), both of which are hyperparameters that are tested for optimization in the following experiments we conduct.

Furthermore, we have:

$$d_{neg} = ||\sigma_{ED} - \sigma_{ES}||_2 \tag{3}$$

where σ_{ED} and σ_{ES} are the normalized ED and ES embeddings.

3.2 Volumetric margin

To encourage the model to repel the ED and ES clusters to a certain threshold, we are using a volume-based loss factor that is defined as follows:

$$\textbf{Volumetric Margin} = \frac{EDV - ESV}{\beta} \tag{4}$$

where EDV and ESV are the labeled ED and ES volumes, and β is the scaling factor. This margin would result in a higher loss and therefore more repulsion if ED and ES volumes are very different. On the other hand, the loss would be lowered if the ED and ES volumes are closer together. This is done to distance the frames that are more anatomically dissimilar while keeping the frames with volumetric similarity closer.

$\beta = 200$ is the scaling factor that was used to keep the margin centered around 0.25[2]. Prior works [10; 11] have used a fixed margin to impose the same type of distance-based loss, but as can be seen in Table1, using the varying margin method outperforms the fixed margin in our experiments.

3.3 Visualization

To analyze the effect of our contrastive loss on the embeddings, we are using Uniform Manifold Approximation and Projection (UMAP)[7]. UMAP is a dimensionality reduction technique similar to T-distributed Stochastic Neighbor Embedding (t-SNE)[6] that compresses the entire graph instead of a point-by-point approach. This allows UMAPs to preserve Euclidean distance much better than t-SNE. As seen in Fig 3, training the EchoGNN with contrastive loss causes clusters of ED and ES to be pushed away from each other.

4 Experiments

4.1 Dataset

All experiments are run on the subset of the EchoNet-Dynamic dataset [9]. The full EchoNet-Dynamic dataset consists of 10,030 AP4 echo videos obtained from Stanford University Hospital between the years of 2016 to 2018. For validation purposes and due to lack of time to run multiple experiments, we randomly sampled 30% of the EchoNet-Dynamic dataset with the condition of keeping the distribution of the EF the same as the original dataset. As shown in Figure 4, the complete EchoNet-Dynamic dataset and the 30% dataset we used have a similar distribution of EF, EDV, ESV, and data split.

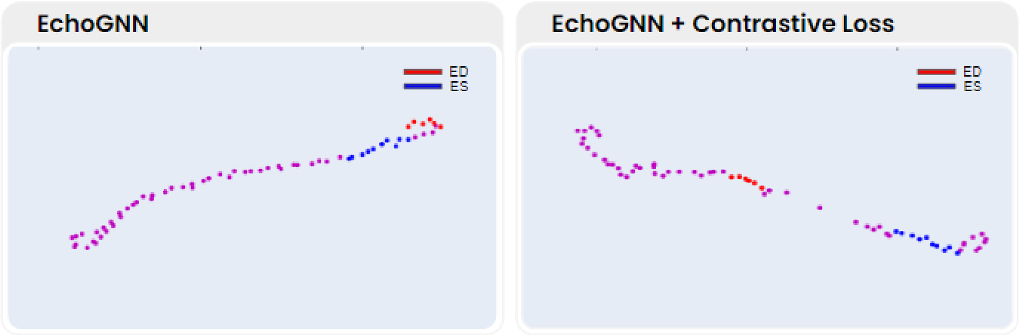

Figure 3: The UMAP visual comparison of the embeddings generated by baseline EchoGNN (left) and EchoGNN trained with contrastive loss (right). Frames associated to ED and its neighbors are in red, while frames associated to ES and its neighbors are in blue. In this two-dimensional space, EchoGNN seems to map ED and ES frame clusters close together, while EchoGNN with contrastive loss creates a distinct separation and distance between the two clusters.

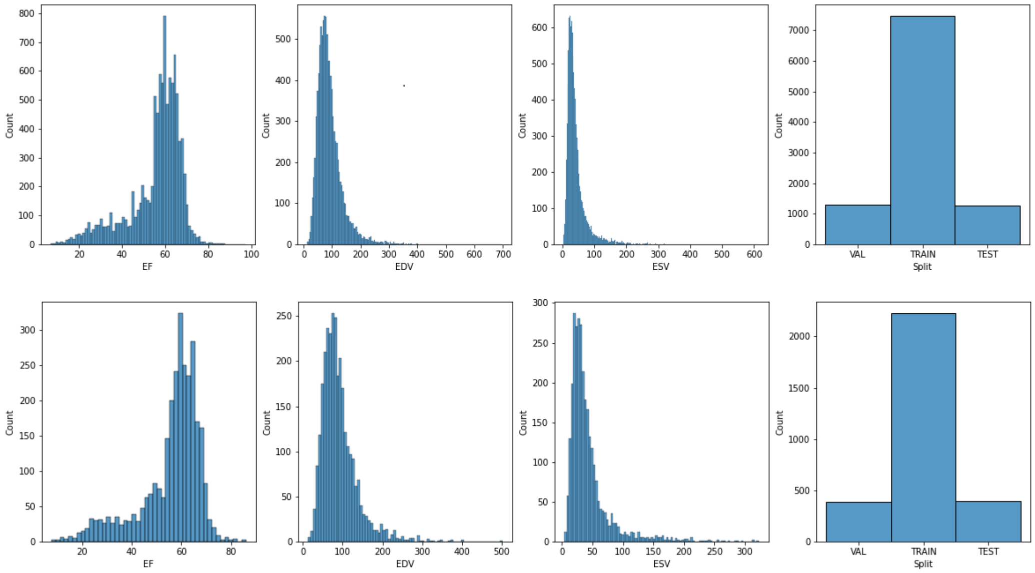

Figure 4: Comparison of the distribution of the EF, EDV, ESV, and split between the original EchoNet-Dynamic dataset (top) and the sampled 30% dataset (bottom).

4.2 Experiment 1: Predictive performance

The following metrics are used to evaluate the predictive performance of the proposed EchoGNN with contrastive loss as compared to the baseline EchoGNN: mean absolute error (MAE) between the predicted EF and the ground truth, coefficient of determination (R^2) which is used to indicate the amount of variation in EF data that can be explained by a model's predictions, and F1-score for classification of EF that falls below 40% (Table 1). This F1-score is used to assess model accuracy for cases where the ground truth EF is below 40% which indicates the patient is in moderate systolic heart failure.

We can see that EchoGNN using contrastive loss and varying margin outperforms the baseline in terms of R^2 and MAE. However, it performs slightly worse in terms of F1-score. This shows that our model, compared to the baseline EchoGNN, seems to perform worst in cases where a patient's EF is moderately reduced and impaired.

Table 1: Performance metrics of proposed EchoGNN and the volumetric margin contrastive loss variations.

	Metrics		
Model	MAE	R^2	F1 $<$ 40% EF
EchoGNN (baseline)	0.0519	0.65	**0.829**
EchoGNN Varying Margin	**0.0507**	**0.67**	0.816
EchoGNN Fixed Margin	0.0539	0.61	0.795

A possible explanation for this could reside with the volumetric margin we use to implement the contrastive loss. For patients with a moderately reduced EF, the difference between EDV and ESV is lessened as the heart is unable to contract as well. Due to a decrease in the difference between these volumes, the model may take these cases less into account as compared to normal EF cases. In addition, further analysis of the dataset shows that around 12.6% of the samples used had an EF $< 40\%$. It's possible that due to this imbalance of moderately reduced cases in the dataset in combination with our volumetric margin, our model focused more towards cases with normal EFs. We believe that with a more balanced dataset, it's likely the model's performance will improve.

4.3 Experiment 2: Neighboring frames hyperparameter

When deciding the number of neighboring frames to include in a cluster, we take a larger range of ES neighboring frames than ED neighboring frames as clinicians we interviewed tended to find the uncertainty associated with labeling the ES frame larger than that of labeling the ED frame. Therefore, $\varepsilon > \delta$ in all runs we use to run hyperparameter tests.

Further exploration into the statistical inter-observer variability between clinicians in frame labeling is an area in need of more research. However, even with the current research done in echocardiography reproducibility, it's shown that ESVs determined by different clinicians are more easily repeatable [12]. However, this only refers to the volumes obtained, not the indicated frame. Because ESV is more easily reproducible, this leads us to believe that there are potentially more neighboring end-systolic frames that look more similar to each other than neighboring end-diastolic frames that are similar, thus making it more difficult to identify the exact ES frame.

Table 2 shows the results of the hyperparameter values tested.

Table 2: Performance metrics when tuning the hyperparameters for the number of neighboring frames in a cluster.

Model		Metrics		
ED frames δ	ES frames ε	MAE	R^2	F1 $<$ 40% EF
1	1	0.0554	0.60	0.7477
3	5	0.0510	0.66	0.795
3	7	**0.0507**	**0.67**	**0.816**
7	11	0.0517	0.64	0.776

To analyze the effects clustering has on the model, we can compare our results to a model that essentially does not use clustering, i.e. $\delta = 1$ and $\varepsilon = 1$. We can see that clustering improves the performance of the model up until a certain point.

As seen in Table 2, a model that uses hyperparameters of $\delta = 3$ and $\varepsilon = 7$ performs the best. Increasing the value of δ and ε shows an increase in performance up until these values, but past these values performance starts to decrease. It's possible that by increasing the number of frames in the cluster past $\delta = 3$ and $\varepsilon = 7$, the frames being clustered no longer share enough similarity with the ED and ES frame. In other words, increasing the value would exceed the number of frames within the uncertainty associated with labeling, and the addition frames added most likely would not be frames labeled by clinician as either ES or ED. Therefore, increasing δ and ε past these points becomes detrimental to our model. However, confirmation of our hypothesis requires further addition experiments and analysis.

4.4 Experiment 3: Ablation study

In the ablation studies, we analyzed the effectiveness of our contrastive loss and how each component affects the accuracy of EF estimation. A quantitative summary of the results are shown in Table 3.

Table 3: Quantitative results of the contrastive loss analysis.

Model			Metrics		
d_{pos}	d_{neg}	volumetric margin	MAE	R^2	F1 $<$ 40% EF
✓	✓	✓	**0.0507**	**0.67**	0.816
✓	✓	✗	0.0544	0.62	0.739
✓	✗	✓	0.0544	0.61	0.733
✓	✗	✗	0.0559	0.63	0.768
✗	✗	✗	0.0519	0.65	**0.829**

We can see from the results that only the EchoGNN model using a contrastive loss with all three components outperforms the baseline EchoGNN model in terms of R^2 and MAE. Again, we see the trend that adding any component of contrastive loss causes the model to perform slightly worse in terms of F1-score.

With only a single component, d_{pos}, the model outputs suboptimal results. However, as we add d_{neg} and volumetric margin, the model's performance begins to improve. This suggests that distancing the embeddings based on temporal difference in the cardiac cycle and volumetric differences in the left ventricle benefits the models.

5 Conclusion

In this work, we applied contrastive learning in a semi-supervised manner to the EchoGNN model while tackling the inter-observer variability associated with labeling ED and ES frames. Our contrastive loss considers both the temporal and volumetric context associated with frames in the echo videos. We showed that by using a varying margin volume loss, we can successfully improve the EF estimation of the EchoGNN model.

5.1 Future work

There are multiple areas that were part of the original scope of the project, but due to the lack of time and the lengthy process of training EchoGNN (one week for approximately 700 epochs), we unfortunately were unable to explore them.

1. Expanding the clustering across multiple videos, and possibly even patients, given the similarity of the EDV and ESV.
2. Applying a spatial component to the contrastive loss.
3. Defining a pre-training regression task.

As we progressed through the project, there were other questions and areas that we discovered for further work.

1. The addition of the contrastive loss into EchoGNN increased the edge sparsity errors on the validation set compared to the baseline. The reason is unknown to us at the moment, but it is definitely worth investigating.
2. Although our model outperforms the baseline in terms of the MAE and R^2 score, we see a decrease in F1-score. As mentioned above, this could be due to the imbalance in the dataset or related to the decrease in the volumetric difference of EDV and ESV in low EF cases. While we have a hypothesis, there are further experiments needed in order to confirm these theories.
3. We hypothesize above that increasing δ and ε past 3 and 7 respectively exceeds the number of frames that share similarity with ED and ES. In order to confirm our hypothesis, more

experiments are required. This includes running more models with varying values of δ and ε, as well as analyzing visual results with clinical consultation.

References

[1] Basu, S., Singla, S., Gupta, M., Rana, P., Gupta, P., Arora, C.: Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining (2022). https://doi.org/10.48550/ARXIV.2207.13148, https://arxiv.org/abs/2207.13148

[2] Cheng, L.H., Sun, X., van der Geest, R.J.: Contrastive learning for echocardiographic view integration. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 340–349. Springer Nature Switzerland, Cham (2022)

[3] Fu, Z., Jiao, J., Yasrab, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Anatomy-aware contrastive representation learning for fetal ultrasound (2022). https://doi.org/10.48550/ARXIV.2208.10642

[4] Jafari, M.H., Girgis, H., Van Woudenberg, N., Liao, Z., Rohling, R., Gin, K., Abolmaesumi, P., Tsang, T.: Automatic biplane left ventricular ejection fraction estimation with mobile point-of-care ultrasound using multi-task learning and adversarial training. International Journal of Computer Assisted Radiology and Surgery **14**(6), 1027–1037 (2019)

[5] Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J.T., Peng, X.: Contrastive clustering (2020). https://doi.org/10.48550/ARXIV.2009.09687, https://arxiv.org/abs/2009.09687

[6] van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008), http://www.jmlr.org/papers/v9/vandermaaten08a.html

[7] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction (2018). https://doi.org/10.48550/ARXIV.1802.03426, https://arxiv.org/abs/1802.03426

[8] Mokhtari, M., Tsang, T., Abolmaesumi, P., Liao, R.: Echognn: Explainable ejection fraction estimation with graph neural networks. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 360–369. Springer Nature Switzerland, Cham (2022)

[9] Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C., Heidenreich, P., Harrington, R., Liang, D., Ashley, E., Zou, J.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580** (04 2020)

[10] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2015). https://doi.org/10.1109/cvpr.2015.7298682, https://doi.org/10.1109%2Fcvpr.2015.7298682

[11] Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S.: Time-contrastive networks: Self-supervised learning from video (2017). https://doi.org/10.48550/ARXIV.1704.06888, https://arxiv.org/abs/1704.06888

[12] Thorstensen, A., Dalen, H., Amundsen, B.H., Aase, S.A., Stoylen, A.: Reproducibility in echocardiographic assessment of the left ventricular global and regional function, the hunt study. European Journal of Echocardiography **11**(2), 149–156 (2010)