EchoFM: A View-Independent Echocardiogram Model for the Detection of Pulmonary Hypertension

Shreyas Fadnavis*[1], Chaitanya Parmar*[1], Nastaran Emaminejad[1], Alvaro Ulloa Cerna[1], Areez Malik[1], Mona Selej[1], Tommaso Mansi[1], Preston Dunnmon[1], Tarik Yardibi[1], Kristopher Standish†[1], and Pablo F. Damasceno†[1]

[1]Janssen R&D, LLC, a Johnson & Johnson Company
{sfadnavi, cparmar, nemamine, aulloace, amalik6, mselej, tmansi, pdunnmon, tyardibi, kstandis, pdamasc1}@its.jnj.com

Abstract. Transthoracic Echocardiography (TTE) is the most widely-used screening method for the detection of pulmonary hypertension (PH), a life-threatening cardiopulmonary disorder that requires accurate and timely detection for its effective management. Automated PH risk detection from TTE can flag subtle indicators of PH that might be easily missed, thereby decreasing variability between operators and enhancing the positive predictive value of the screening test. Previous algorithms for assessing PH risk still rely on pre-identified, single TTE views which might ignore useful information contained in other recordings. Additionally, these methods focus on discerning PH from healthy controls, limiting their utility as a tool to differentiate PH from conditions that mimic its cardiovascular or respiratory presentation. To address these issues, we propose EchoFM, an architecture that combines self-supervised learning (SSL) and a transformer model for view-independent detection of PH from TTE. EchoFM 1) incorporates a powerful encoder for feature extraction from frames, 2) overcomes the need for explicit TTE view classification by merging features from all available views, 3) uses a transformer to attend to frames of interest without discarding others, and 4) is trained on a realistic clinical dataset which includes mimicking conditions as controls. Extensive experimentation demonstrates that EchoFM significantly improves PH risk detection over state-of-the-art Convolutional Neural Networks (CNNs).

Keywords: Transthoracic Echocardiograms · Foundation Model

1 Introduction

Diagnosis of Pulmonary Hypertension (PH) is challenged by non-specific, overlapping symptoms that negatively impact patient outcomes [7]. Transthoracic Echocardiography (TTE) is a non-invasive technique that uses sound waves to

*represents co-first authors, †represents co-last authors

generate heart images. TTE is crucial for PH screening by enabling the measurement of cardiac velocities, sizes, and pressures that, in combination with international guidelines [10], can be used to compute PH risk. Despite its utility, TTE has shown a false negative rate of up to 36% [9, 20] and significant inter-reader variability in diagnosis [12] compared to right heart catheterization (RHC), an invasive gold standard for PH diagnosis. Recent studies have explored using deep learning techniques for automation of PH-related measurements [21] or direct estimation of PH risk [24]. While not addressing the prevalent issue of clinical discrimination of PH from other mimicking conditions, these works present an important step towards shortening the patient journey. However, state-of-the-art approaches still depend on either manual or automated pre-selection of specific views from numerous images that are acquired during a patient's clinical visit. This selection can introduce errors to downstream classification tasks and may overlook valuable information present in other views.

Here we investigate the use of a transformer-based model, EchoFM, for PH differential diagnosis from TTE videos without the need for explicit view classification. To maximize the benefits of multiple views, we combine robust frame-wise feature extraction using a foundational model (FM) with a transformer-based classifier, enabling simultaneous loading and dynamic analysis of all patient recordings without pre-selection of views. Experiments in two real-world clinical datasets demonstrate that EchoFM achieves state-of-the-art performance in differentiating PH from healthy controls as well as from other cardiovascular or respiratory conditions.

2 Related Work

Computer vision significantly aids heart measurement estimation and diagnosis from TTE waveforms. A typical TTE acquisition generates multiple still images, Doppler waves, and B-mode views, such as the parasternal long-axis (PLAX), parasternal short-axis (PSAX), apical four-chamber (A4c), and subcostal four-chamber (S4c) views, all essential for assessing cardiac structure and function (Fig. 1) [1]. Automated view classification has been explored using methods like spatio-temporal feature extraction, dictionary learning [13], and deep learning techniques such as convolutional neural networks (CNNs) and contrastive learning [16, 24, 15, 4]. Despite advancements, challenges persist due to the similarity among some views and intra-view variability, affecting classification performance and disease risk assessment. EchoNet, a 3D CNN, has been trained to predict cardiac volumes and ejection fraction from A4c views [6]. Other deep learning workflows have been developed for segmentation and annotation of cardiac measurements in TTE videos [21]. However, all these approaches often rely on a single TTE view and do not fully exploit the temporal dynamics and diverse imaging orientations, leaving potential cardiac dysfunction indicators unexplored. Additionally, these algorithms often provide diagnosis of PH with respect to healthy controls, a significantly less challenging task than diagnosis with respect to mimicking conditions or those with overlapping symptoms [2]. Our research addresses these issues by integrating a pre-training encoder with a transformer model,

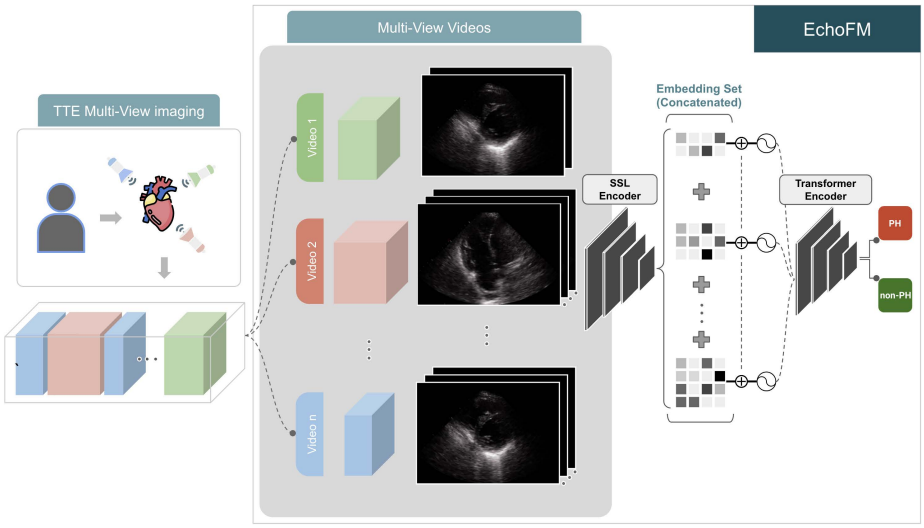

Fig. 1. EchoFM Architecture for View-Independent TTE-based PH Classification. TTE videos acquired during a patient visit (left) are inputted into a self-supervised model (center) that produces frame-wise embeddings vectors. Upon concatenation and location via positional tokens, they are fed into a transformer model (right) trained specifically to identify PH subjects from non-PH.

leveraging the strengths of Self-Supervised Learning (SSL), which have shown promise in medical imaging and video data analysis [5, 3, 18, 22].

3 Methods

Multi-view TTE image data can be represented as a set of matrices, each row corresponding to a specific view acquired during the patient visit. We consider a dataset comprising N patients. For each patient i, $\mathbf{V}$ distinct videos, each composed of m_v frames, are recorded. The video dataset can be expressed as:

$$\mathcal{X} = \left\{ X_{i,v} \in \mathbb{R}^{m_v \times D_1} \,\middle|\, i = 1, 2, \ldots, N;\ v = 1, 2, \ldots, \mathbf{V} \right\}$$

Here, $X_{i,v}$ denotes the matrix associated with the i-th patient in the v-th video. Each matrix has dimensions $m_v \times D_1$, where D_1 represents the feature dimensionality of individual frames in the video with m_v frames.

3.1 EchoFM Algorithm:

EchoFM is composed of two main components: a self-supervised encoder and a downstream task classifier, explained in detail below:

3.1.1 Self-supervised Pre-training for Robust Representations: We adopt a self-supervised pre-training strategy [8, 17, 19] to embed TTE video frames into dense and robust representations. EchoFM employs DINOv2 [18], using a loss function $\mathcal{L}_{\text{DINOv2}}$ defined as the mean negative dot product between teacher and

student representations across all augmentation training pairs, following temperature scaling and normalization:

$$\mathcal{L}_{\text{DINOv2}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{z_{m_{v_1},i}\cdot z_{m_{v_2},i}}{\left\|z_{m_{v_1},i}\right\|_2\cdot\left\|z_{m_{v_2},i}\right\|_2}\right)$$

Here, N signifies the total count of image augmentation pairs, $z_{m_{v_1},i}$ and $z_{m_{v_2},i}$ represent the temperature-scaled embeddings for the i-th image pair's augmentations m_{v_1} and m_{v_2}, produced by the student network. This loss function serves to harmonize the student network's representations of differing views of an identical image, as guided by the teacher network. Each frame of the TTE video, with dimensions $D_1 = 224 \times 224 \times 3$, was independently fed into the DINOv2 model. Subsequently, a Vision Transformer model (ViT B/16) [14] was pre-trained using the DINOv2 objective. This procedure employed a batch size of 256 and a learning rate, initialized at 0, warmed up to 8e-4 during the initial 10% of epochs, followed by a cosine decay to 1e-6 for the remaining iterations. The pre-training phase spanned 400k iterations on four NVIDIA A10G GPUs (two days). After DINOv2 pre-training, each frame was transformed into an embedding of $D_2 = 768$ dimensions. The resulting output for each video formed a matrix:

$$\mathbf{X}_{i,v} = \left[\mathbf{d}_1\ \mathbf{d}_2 \vdots \mathbf{d}_{m_v}\right]$$

Where $\mathbf{X}_{i,v}$ is the matrix for a video, $\mathbf{d}_i$ is the i-th frame's D_2-dimensional embedding, and m_v is the total number of frames for that video. These embeddings constitute the inputs for downstream analysis within the EchoFM model.

3.1.2 Transformer Network for View-Independent Tasks: We enhance TTE video analysis by incorporating positional encoding to capture the temporal sequence within each echocardiographic video. This approach enables dynamic function learning and frame relevance differentiation for disease risk assessment. The positional encoding formula for frame position *pos* and dimension D_2 (set to 768 for EchoFM) is:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/D_2}}\right),$$
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/D_2}}\right),$$

facilitating an understanding of temporal dynamics in a $n \times D_2$ matrix format. Our model employs a *multi-head attention mechanism* [23], crucial for analyzing temporal sequences in multi-view TTE data. By assigning weights to each frame, it identifies critical sections of the data through:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right),$$

where Q, K, and V are queries, keys, and values. This mechanism allows for dependency tracking across frames. The model also employs multi-head attention:

Dataset		Patients	Videos	Frames
Sheffield	Train	615	13,961	5,800,316
	Validate	205	4,731	2,004,004
	Test	204	4,449	1,664,886
CIPHER	Train	443	8,539	13,796,246
	Validate	147	3,257	5,964,317
	Test	149	3,189	5,219,280

Table 1. Data details for training, validation, and testing.

$$\text{MultiHead}(Q, K, V) = [\text{head}_1; \text{head}_2; ...; \text{head}_h]W_O,$$

to model inter-view correlations, enhancing classification accuracy. The term head_n represents an individual attention head capturing different aspects of the input, while W_O is a weight matrix that transforms the concatenated outputs of all attention heads into a vector.

3.2 Weakly Supervised Classification of TTE Videos: We explore an alternative to transformer-based spatio-temporal classification using weakly supervised learning (WSL) for state-of-the-art CNNs. WSL addresses the challenges of analyzing large video datasets by randomly sampling frames and learning from their features for classification. In this setup, features from frame-wise 2D CNNs are integrated using an attention-based aggregator, and the compiled feature vector is processed by a classifier network trained end-to-end. This approach can be used by both traditional single and multi-view CNNs. The impact of SSL pre-training was investigated as part of our ablation study.

4 Data and Experiments

Dataset: Our study utilized two distinct, private PH datasets: *Sheffield* [11] and *CIPHER* [9]. Both datasets include subjects initially suspected of PH and evaluated by both TTE and invasive RHC. The presence of this invasive procedure likely means that some cardiovascular or respiratory conditions was present for all subjects, making this a significantly more challenging dataset compared to those that include only PH and healthy controls. The *Sheffield* dataset consists of 1024 subjects, averaging 10 TTE views each, with a maximum of 20 views. The *CIPHER* dataset includes 739 subjects, each having an average of 10 views and up to 22 views (see Tab. 1).

Data pre-processing: An echocardiography study typically consists of 10+ videos, each containing a (sometimes repeated) view of the heart. For pre-processing, video pixel data found in the DICOM files were converted to RGB from either YCRCB or gray colormaps and the region where the beam-formed cone was located was found in the “Sequence of Ultrasound Regions" DICOM tag. Finally, the frames were saved into PNGs.

Evaluation Metrics: In diagnosing PH using TTE, we compute AUC, F1 Score, and Accuracy in a 4-fold cross-validation setting. These respectively evaluate the model's ability to distinguish between PH and non-PH cases, maintain

precision-recall balance, and correctly classify instances, with special consideration given to potential dataset imbalance. To binarize the continuous predictions the model output, a threshold of 0.5 was used on the sigmoid outputs.
View Classification: We employed off-the-shelf view classification algorithms [24] to identify, for each patient, the video with highest probability of being A4c – the *de facto* standard echocardiogram view, used by most traditional methods to acquire relevant anatomical and functional information for PH classification.

Method	Training	AUC Score	F1 Score	Accuracy
Baseline	CNN (single)	0.67 ± 0.05	0.56 ± 0.33	0.52 ± 0.20
EchoFM	**DINOv2**	**0.80 ± 0.01**	**0.87 ± 0.01**	**0.79 ± 0.02**

Table 2. Comparison of performance metrics between baseline method (CNN with single video) and EchoFM.

5 Results

Baseline Experiment: The goal of this experiment was to assess the efficacy of EchoFM at distinguishing patients with PH from non-PH patients. For benchmarking purposes, we used an off-the-shelf view-classification algorithm [24] to single out the one A4c view per patient, which was fed into a CNN for disease classification. When addressing the challenge posed by the varying lengths of videos, we utilized a WSL model that employs attention-based CNNs. The performance metrics for the baseline using the CNN method revealed an AUC Score of 0.67 ± 0.05, an F1 Score of 0.56 ± 0.33, and an Accuracy of 0.52 ± 0.20. In contrast, our proposed EchoFM, integrating DINOv2 and a custom transformer architecture, addresses the constraints of a variable number of videos and their lengths inherent in the baseline method. The performance enhancement is evident across all metrics with EchoFM recording an AUC Score of 0.80 ± 0.01, an F1 Score of 0.87 ± 0.01, and an Accuracy of 0.79 ± 0.02 (see Tab. 2). We attribute the relatively low performance of traditional methods to the difficulty in distinguishing different types of heart disease present in this dataset – a significantly more challenging problem than identifying PH from healthy controls.
EchoFM Experiments: EchoFM's success in complex diagnosis could be attributed to one or more of the following architectural advancements: 1) the incorporation of multiple views, 2) the deployment of a effective pre-training encoder, and 3) the application of a spatio-temporal network. To evaluate this hypothesis, we methodically implemented the following architectural adjustments: 1) integration of multiple concatenated views within the WSL framework used as baseline, 2) utilization of multiple pre-training encoders, and 3) adoption of varied architectures for the downstream classification. The outcomes of these experiments are detailed in Tab. 3. While SimCLR does not show significant improvement over traditional CNNs, DINO and DINOv2 demonstrate large gains in performance across all three metrics. Comparing WSL with transformers suggests that extracting spatio-temporal features through the transformer model

provides significant improvements for the PH classification task when SimCLR or DINO are used but less so when a stronger encoder is employed. These collective findings suggest that the coupling of DINOv2 with a downstream classification architecture, such as WSL or EchoFM, can considerably enhance the baseline performance of PH classification derived from TTE videos. With SimCLR as the training mechanism, the EchoFM method achieves an AUC score of 0.77 ± 0.01, F1 score of 0.80 ± 0.01, and accuracy of 0.71 ± 0.02. Despite these metrics being lower than those achieved with EchoFM using DINOv2, they still significantly surpass the baseline model. This comparison accentuates the efficacy of modern encoders for risk disease assessment from TTE.

Ablation Study: Table 4 details our ablation study, evaluating the impact of DINOv2 pre-training and EchoFM architecture on model performance. Replacing DINOv2 with ImageNet pre-trained ResNet(34) reduced AUC by 0.10, F1 score by 0.04, and accuracy by 0.05, highlighting DINOv2's critical role. Substituting EchoFM with a weakly-supervised Attention CNN decreased AUC by 0.01, and both F1 score and accuracy by 0.03, indicating the Transformer's slight but positive impact. Limiting DINOv2 to single-view (A4c) pre-training, performance dropped: AUC by 0.05, F1 score and accuracy by 0.01, underscoring the benefits of multi-view analysis for PH assessment.

Generalization Study: Our cross-validation analysis demonstrates our model's generalization across different datasets. Specifically, when trained on the Sheffield dataset and tested on CIPHER (see Tab. 1), the model achieved an AUC of 0.73 ± 0.02, F1 score of 0.88 ± 0.01, and accuracy of 0.81 ± 0.01. Conversely, training on CIPHER and testing on Sheffield yielded an AUC of 0.70 ± 0.02, F1 score of 0.85 ± 0.01, and accuracy of 0.75 ± 0.01 (see Tab. 5). These results validate the model's robustness and generalization capability, though performance variations suggest dataset-specific influences, warranting further investigation into the model's adaptability to different data distributions.

Intepretability via Self-Attention Maps: The EchoFM architecture features a three-tiered attention granularity beneficial for clinical interpretability: video-, frame-, and pixel-level attention, as shown in Fig. 2. Video-level attention (Fig. 2(a)) enables clinicians to identify significant views for PH risk assessment, aligning with standard clinical protocols for PH diagnostics [10]. Frame-level attention highlights frame-wise features critical for PH diagnosis, such as the

Method	Training Mechanism	AUC Score	F1 Score	Accuracy
Weakly Supervised (+Attention)	SimCLR	0.65 ± 0.04	0.70 ± 0.15	0.61 ± 0.12
	DINO	0.76 ± 0.01	0.78 ± 0.08	0.69 ± 0.08
	DINOv2	0.79 ± 0.01	0.84 ± 0.01	0.76 ± 0.02
EchoFM	SimCLR	0.77 ± 0.01	0.80 ± 0.01	0.71 ± 0.02
	DINO	0.79 ± 0.01	0.85 ± 0.02	0.77 ± 0.02
	DINOv2	**0.80 ± 0.01**	**0.87 ± 0.01**	**0.79 ± 0.02**

Table 3. Comparative evaluation of performance metrics for the WSL (+Attention) and EchoFM methods with different training mechanisms.

FM (view)	Classifier	AUC Score	F1 Score	Accuracy
ImageNet (multi)	EchoFM	0.70 ± 0.02 (↓ 0.10)	0.83 ± 0.03 (↓ 0.04)	0.74 ± 0.03 (↓ 0.05)
DINOv2 (multi)	WSL	0.79 ± 0.01 (↓ 0.01)	0.84 ± 0.01 (↓ 0.03)	0.76 ± 0.02 (↓ 0.03)
DINOv2 (single)	EchoFM	0.75 ± 0.03 (↓ 0.05)	0.86 ± 0.02 (↓ 0.01)	0.78 ± 0.03 (↓ 0.01)

Table 4. Ablation study assessing the impact of substituting components of our proposed DINOv2 + EchoFM model.

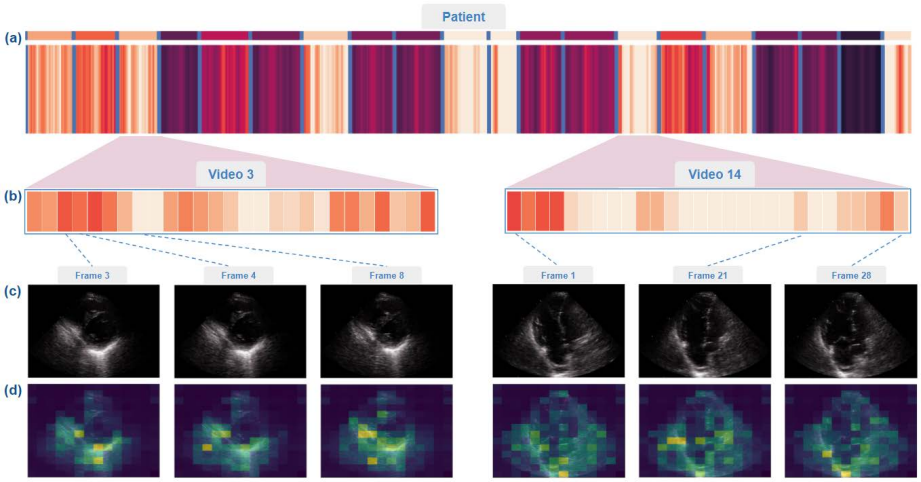

Fig. 2. Video, frame and pixel-wise attention vales for single patient. **(a)** Mean(top) and framewise attention for all **(a)** or selected subset **(b)** views and frames **(c)**. **(d)** DINOv2 attention maps overlaid on actual frames for frame-level interpretability.

flattened interventricular septum and right ventricle enlargement, observable in PSAX and A4c views (Fig. 2(b)). Pixel-level attention, depicted in Fig. 2(c), overlays attention maps on frames to pinpoint diagnostic indicators at a granular level. This structured attention mechanism aids in the clinical interpretation and validation of the model's analytical focus.

6 Conclusion

We introduce EchoFM, a novel Transformer-based method for differential diagnosis of pulmonary hypertension from TTE videos. Our method, leveraging Weakly Supervised Learning, outperforms conventional attention-based CNNs

Train	Test	AUC Score	F1 Score	Accuracy
Sheffield	CIPHER	0.73 ± 0.02	0.88 ± 0.01	0.81 ± 0.01
CIPHER	Sheffield	0.70 ± 0.02	0.85 ± 0.01	0.75 ± 0.01

Table 5. Model performance in cross-dataset validation. Metrics are reported for training on four folds of one dataset and testing on the other, and vice versa.

when pre-trained with DINOv2, offering superior classification without needing specific view detection. EchoFM provides permutation invariance, effectively analyzing frames regardless of sequence. Ablation studies highlight the importance of DINOv2 pre-training and the transformer structure, particularly for integrating multiple views for accurate heart disease analysis. The method's adaptability is proven in real-world datasets which include other, "mimicking" cardiovascular or respiratory conditions as PH-negative cases. Finally, EchoFM highlights critical attention maps within video frames, improving interpretability and aiding in diagnostic insights.

Disclosure of Interests. All authors were employees of Janssen R&D, LLC, and may own company stock/stock options.

References

1. Arnaout, R., Curran, L., Zhao, Y., Levine, J.C., Chinn, E., Moon-Grady, A.J.: An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. Nature medicine **27**(5), 882–891 (2021)
2. Barst, R.J., McGoon, M., Torbicki, A., Sitbon, O., Krowka, M.J., Olschewski, H., Gaine, S.: Diagnosis and differential assessment of pulmonary arterial hypertension. Journal of the American College of Cardiology **43**(12S), S40–S47 (2004)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Chartsias, A., Gao, S., Mumith, A., Oliveira, J., Bhatia, K., Kainz, B., Beqiri, A.: Contrastive learning for view classification of echocardiograms. In: Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 2. pp. 149–158. Springer (2021)
5. 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. PMLR (2020)
6. Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Deep learning interpretation of echocardiograms. NPJ digital medicine **3**(1), 10 (2020)
7. Hambly, N., Alawfi, F., Mehta, S.: Pulmonary hypertension: diagnostic approach and optimal management. CMAJ **188**(11), 804–812 (2016)
8. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. Advances in neural information processing systems **32** (2019)
9. Howard, L., Chin, K., Fong, Y.L., Gargano, C., Stamatiadis, D., Maron, B., Preston, I., Quinn, D., Rosenkranz, S., Toshner, M., et al.: Cipher: a prospective, multicentre study for the identification of biomarker signatures for early detection of pulmonary hypertension (2020)
10. Humbert, M., Kovacs, G., Hoeper, M.M., Badagliacca, R., Berger, R.M., Brida, M., Carlsen, J., Coats, A.J., Escribano-Subias, P., Ferrari, P., et al.: 2022 esc/ers guidelines for the diagnosis and treatment of pulmonary hypertension: Developed by the task force for the diagnosis and treatment of pulmonary hypertension of

the european society of cardiology (esc) and the european respiratory society (ers). endorsed by the international society for heart and lung transplantation (ishlt) and the european reference network on rare respiratory diseases (ern-lung). European heart journal **43**(38), 3618–3731 (2022)
11. Hurdman, J., Condliffe, R., Elliot, C.A., Swift, A., Rajaram, S., Davies, C., Hill, C., Hamilton, N., Armstrong, I.J., Billings, C., et al.: Pulmonary hypertension in copd: results from the aspire registry. European Respiratory Journal **41**(6), 1292–1301 (2013)
12. Janda, S., Shahidi, N., Gin, K., Swiston, J.: Diagnostic accuracy of echocardiography for pulmonary hypertension: a systematic review and meta-analysis. Heart **97**(8), 612–622 (2011)
13. Khamis, H., Zurakhov, G., Azar, V., Raz, A., Friedman, Z., Adam, D.: Automatic apical view classification of echocardiograms using a discriminative learning dictionary. Medical image analysis **36**, 15–21 (2017)
14. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022)
15. Kusunose, K., Haga, A., Inoue, M., Fukuda, D., Yamada, H., Sata, M.: Clinically feasible and accurate view classification of echocardiographic images using deep learning. Biomolecules **10**(5), 665 (2020)
16. Madani, A., Arnaout, R., Mofrad, M., Arnaout, R.: Fast and accurate view classification of echocardiograms using deep learning. NPJ digital medicine **1**(1), 6 (2018)
17. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations pp. 6707–6717 (2020)
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
19. Schiappa, M.C., Rawat, Y.S., Shah, M.: Self-supervised learning for videos: A survey. ACM Computing Surveys **55**(13s), 1–37 (2023)
20. Slegg, O.G., Willis, J.A., Wilkinson, F., Sparey, J., Wild, C.B., Rossdale, J., Ross, R.M., Pauling, J.D., Carson, K., Kandan, S.R., et al.: Improving pulmonary hypertension screening by echocardiography: Impulse. Echo Research & Practice **9**(1), 1–13 (2022)
21. Tromp, J., Seekings, P.J., Hung, C.L., Iversen, M.B., Frost, M.J., Ouwerkerk, W., Jiang, Z., Eisenhaber, F., Goh, R.S., Zhao, H., et al.: Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. The Lancet Digital Health **4**(1), e46–e54 (2022)
22. Truong, T., Mohammadi, S., Lenga, M.: How transferable are self-supervised features in medical image classification tasks? In: Machine Learning for Health. pp. 54–74. PMLR (2021)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
24. Zhang, J., Gajjala, S., Agrawal, P., Tison, G.H., Hallock, L.A., Beussink-Nelson, L., Lassen, M.H., Fan, E., Aras, M.A., Jordan, C., et al.: Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. Circulation **138**(16), 1623–1635 (2018)