ECHOINGECG: An Electrocardiogram Cross-Modal Model for Echocardiogram Tasks

Yuan Gao[1,2,3,7], Sangwook Kim[1,2,3], and Chris McIntosh[1,2,3,4,5,6,7]

[1] Peter Munk Cardiac Centre, University Health Network (UHN), Toronto, Canada
[2] Department of Medical Biophysics, UofT, Toronto, Canada
[3] Ted Rogers Centre for Heart Research, UHN, Toronto, Canada
[4] Department of Computer Science, University of Toronto (UofT), Toronto, Canada
[5] Toronto General Hospital Research Institute, UHN, Toronto, Canada
[6] Department of Medical Imaging, UofT, Toronto, Canada
[7] Vector Institute, Toronto, Canada
{yuan.gao, sangwook.kim, chris.mcintosh}@uhn.ca

Abstract. Electrocardiogram (ECG) is a widely used tool for assessing cardiac function due to its low cost and accessibility. Emergent research shows that ECGs can help make predictions on key outcomes traditionally derived from more complex modalities such as echocardiograms (ECHO), enabling the use of ECGs as a more accessible method to predict broader measurements of cardiac function. ECHO, in particular, are of great importance because they require considerable hospital resources while playing a key role in clinical cardiac assessment. To aid this use case, we introduce ECHOINGECG, a probabilistic student-teacher model that leverages uncertainty-aware ECG embeddings and ECHO supervision to improve ECG-based cardiac function prediction. Our approach integrates Probabilistic Cross-Modal Embeddings (PCME++), a probabilistic contrastive framework, with ECHO-CLIP, a vision-language pretrained model trained on ECHO-text pairs, to distill ECHO knowledge into ECG representations. Through experiments and external validation, we showed that ECHOINGECG outperforms state-of-the-art foundation ECG models in zero-shot, few-shot, and fine-tune settings for ECHO predictions based on ECG. We also highlighted that variance estimation (enabled through our method) enhanced our understanding of model performance by identifying underlying regions of uncertainty within ECGs. The code is available: `https://github.com/mcintoshML/EchoingECG`.

Keywords: Probability Learning · Knowledge Distillation · ECG

1 Introduction

Electrocardiograms (ECGs) are widely used for cardiac assessment due to their accessibility and cost-effectiveness [3]. A growing area of research leverages the use of ECGs to estimate echocardiogram (ECHO) variables to evaluate heart function [31,35], which could enhance early disease detection and expand healthcare accessibility, particularly in resource-limited settings. However, replicating
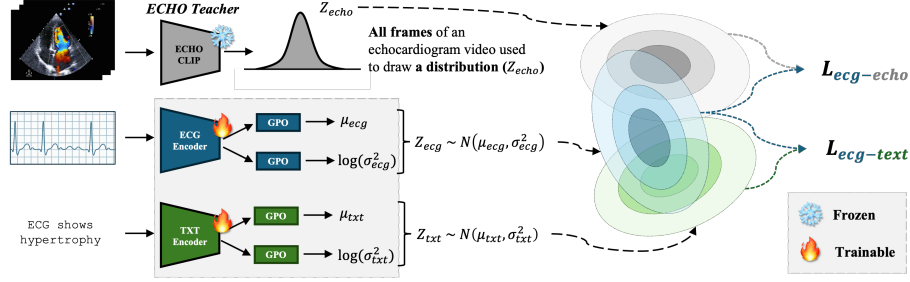
Fig. 1. The probabilistic approach employs PCME++ to train the VLPM within a probabilistic space. **ECHO Teacher (top)**: ECHO-CLIP extracts the distribution's mean μ and variance $\log(\sigma^2)$ by utilizing multiple frames from ECHO videos. **Student Model (bottom):** Using PCME++, each encoder generates an l_2-normalized vector μ and a vector $\log(\sigma^2)$ to represent a normally distributed random variable. EchoingECG is trained to bind both ECG-Text $L_{ecg-text}$ and ECG-ECHO $L_{ecg-echo}$.

ECHO-derived insights from ECGs remains challenging due to **(1)** complexity of ECG signals and **(2)** scarcity of ECG data labeled with ECHO outcomes.

For the **first challenge**, cross-modality contrastive learning offers a promising approach to learning semantically rich embeddings that capture clinically relevant signal patterns [10, 23, 26]. Emerging models such as ECG-CLIP, MED-Bind [10], and others adapt this paradigm, using text descriptions (ECG impressions) for training. These ECG models demonstrate broad applicability in various downstream tasks. However, traditional cross-modal contrastive learning to ECG-ECHO and/or ECG-text integration introduces additional complexities. ECG signals exhibit subtle variations, with distinct patterns recurring at non-uniform time intervals and manifesting differently across recordings [4, 11]. For example, a 120-second ECG is subdivided into many 10-second windows; some windows capture irregular beats, while others do not, yet all windows correspond to the paired report or ECHO. Consequently, injective embeddings in traditional contrastive learning may be limiting as a deterministic embedding space, which means there exists only one-to-one mappings.

To address this, we propose contrastive cross-modal learning using probabilistic vision-language models (VLPMs), introducing more flexible embeddings. Unlike traditional contrastive learning, probabilistic VLPMs enable many-to-many mappings, better capturing the inherent uncertainty in cross-modality binding. Although we are subject to natural physiological and measurement variations in our framework, different ECG recordings can converge on the same underlying cardiac condition observed in the text or ECHO. This many-to-many mapping reflects the reality that various representations (ECG signals) may all indicate the same cardiac state, thereby enhancing the model's robustness. We employ Probabilistic Cross-Modal Embeddings (PCME++) [7,8], which explicitly model uncertainty by learning both the mean and variance of latent representations. By applying the variational information bottleneck principle, PCME++ enables the

network to capture uncertainty inherent in paired modalities. This probabilistic structure allows all ECG windows to map to all ECHO frames and/or possible texts, with lower certainty assigned to mappings between less relevant segments.

To address the **second challenge**–the need for large-scale annotated ECG-ECHO datasets–recent research has explored using foundation models as teachers to enhance learning in domains with limited labeled data. The Unified Representation of Language, Images, and Point Clouds (ULIP) [34] exemplifies this, where it leveraged large-scale VLPMs (2D model) to assist training 3D point cloud understanding. ULIP demonstrated that models using a foundation teacher on paired unlabeled data could effectively transfer knowledge to different modalities. We extended this idea by leveraging ECHO-CLIP [6], a VLPM trained on ECHO-text pairs, as a teacher model to guide ECG representation learning, through **training on ECHO-ECG pairs**. Our approach distills echo-based cardiac information into ECG representations and, through this alignment, enables ECHO-based training while mitigating large-scale annotated datasets.

Contributions: We introduce ECHOINGECG, a probabilistic student-teacher framework that integrates PCME++ and ECHO-CLIP for enhanced ECG modeling. Our contributions are: **(1)** the development of a probabilistic multimodal ECG model that captures structural and functional insights from ECHO, **(2)** a knowledge distillation approach using a student-teacher method that reduces the dependency on large-scale annotated datasets, and **(3)** experimental evidence showing that probabilistic modeling in ECHOINGECG enables an uncertainty-aware evaluation of downstream tasks.

2 Methods and Materials

2.1 Probabilistic Contrastive Learning

Temporal variability is critical in ECGs, as disease manifestation can occur over time. Traditional deterministic contrastive learning methods, InfoNCE [28], enforce one-to-one mappings, which often struggle to capture the inherent many-to-many relationships in ECGs–where distinct signal patterns can encode similar clinical insights simultaneously. Using a probabilistic embedding space: our framework explicitly models this variance, allowing ECG representations to account for signal dynamics, noise, and artifacts. To achieve this, we utilized the PCME++ framework [7], described below.

Traditional VLPM binding of ECG embedding x_{ecg} and text embedding x_{text} involves deterministic contrastive learning [29]. Instead, following the probabilistic approach in [7], we defined a modality embedding x_m as a random variable with a multivariate normal distribution: $x_m \sim \mathcal{N}\left(\mu_m, \sigma_m^2\right)$, where $\mu_m \in \mathbb{R}^D$ is the mean vector and $\sigma_m^2 \in \mathbb{R}^D$ denotes the **diagonal covariance**, as proposed in [7]. In practice, we optimized $\log(\sigma^2)$, to reduce numerical instability. PCME++, Figure 1, optimizes embeddings to maximize the mutual likelihood of pairs (e.g., ECG signal, Z_{ecg}, and its corresponding text, Z_{text}), while separating unmatched pairs. This is achieved through PCME++'s Closed-Form Sampled

Distance (CSD) metric (Eq. 1), which measures the **probabilistic distance** between embeddings by combining their μ and σ^2.

$$d\left(Z_{m1},\ Z_{m2}\right)\ =\ \|\ \mu_{m1}\ -\ \mu_{m2}\ \|_2^2\ +\ \|\ \sigma_{m1}^2\ +\ \sigma_{m2}^2\|_1 \tag{1}$$

The properties in Eq. 1 are described in detail in [7]. In short, it involves the first term encouraging distributions between modality pairs to have a proper distance between pairs and the second term encouraging embeddings to reflect certainty through minimal σ^2 for aligned pairs and larger σ^2 for ambiguity. Notably, PCME++'s CSD is directly integrated into InfoNCE loss [28], enabling contrastive learning in a probabilistic embedding space.

2.2   Echo Teacher for ECG-to-ECHO Binding

While ECG-text binding using CSD is applied directly from [7], EchoingECG used a **ECHO Teacher** model that extends learning to strengthen semantic representation through cross-modality binding (ECG-ECHO). We focused on enriching ECG embeddings using cross-modality linking with the existing foundation model: ECHO-CLIP [6]. ECHO-CLIP produces frame-level embeddings for each ECHO video. We denote these embeddings as $e_1, e_2, \ldots, e_n$, where n is the number of frames in the ECHO and, for each $i = 1, \ldots, n$, $e_i \in \mathbb{R}^d$ is the embedding of the ith frame with d dimensions. To simplify probabilistic modeling herein, we assumed each frame, e_i, of the ECHO as independent. While this approach does not explicitly model temporal dependencies, it does capture inter-frame variation across the cardiac cycle (i.e., low-LVEF cases typically show reduced movement, leading to lower σ^2, and vice versa). In future work, we aim to explore modeling inter-frame dependencies.

Thus, we estimated the overall representation of the ECHO video using statistical measures considering the embeddings, e, from independent frames as distribution samples. ECHO-CLIP processed each frame; then, we calculated the sample μ and σ^2 of the frame embeddings to represent the ECHO video as a probabilistic embedding suitable for PCME++. The μ embedding is $\frac{1}{n}\sum_{i=1}^n e_i$, while σ^2 embedding: $\sigma^2 = \frac{1}{n}\sum_{i=1}^n (e_i - \mu) \odot (e_i - \mu)$, where $\odot$ denotes elementwise multiplication. By computing these: we modeled the ECHO video as a multivariate normal distribution: $Z_{echo} \sim \mathcal{N}\left(\mu, \sigma^2\right)$, where σ^2 is a diagonal σ^2 matrix formed from the σ^2 vector. Thus, (1) can be applied to ECG-to-ECHO binding, where we compare Z_{ecg} and Z_{echo} (i.e. Z_{m1} and Z_{m2}, respectively).

2.3   Loss Combination

While ECG-ECHO learning is the primary focus, text alignment plays a role in aligning ECG and ECHO through shared supervision, as ECHO-CLIP was trained in ECHO-text pairs. To leverage both ECG-text (Section 2.1) and ECG-ECHO learning (i.e., ECHO Teacher proposed in Section 2.2), we introduce a combined loss function that balances their respective contributions via a hyperparameter λ. The overall loss used in EchoingECG, $\mathcal{L}_{\text{total}}$, is defined as:

$$\mathcal{L}_{\text{total}} = \lambda \cdot \mathcal{L}_{\text{ecg-text}} + (1 - \lambda) \cdot \mathcal{L}_{\text{ecg-echo}} \tag{2}$$

where $\mathcal{L}_{modality-modality}$ represents the probabilistic contrastive loss (Section 2.1), the pairs for modality represent the corresponding ECG-text and ECG-ECHO (student with ECHO Teacher), and λ provides a balance for training.

2.4   Model Architecture

ECHOINGECG integrates trainable modality-specific encoders to process text and ECG signals. For the text encoder, we employed BioBERT [22], a biomedical variant of BERT [18], to capture nuanced medical language relevant to cardiac function. We utilized a 1D-ResNet as the ECG encoder, an efficient framework for extracting features from ECG [19,32]. The primary focus of this research was on the training method and highlighting the utility of probabilistic embeddings. To isolate our training strategy, we used standard architectures, reserving architecture changes for future work. Following the text and ECG encoders, two separate linear layers are used–one to produce μ and the other for $\log \sigma^2$. So, an input is mapped to a normal distribution parameterized by μ and $\log \sigma^2$. For ECHO-ECG learning, we used a frozen ECHO-CLIP [6], as a teacher network, to extract semantic representations of ECHO. ECHO-CLIP processed each frame of an ECHO video separately and μ and $\log \sigma^2$ for each ECHO (Section 2.2).

2.5   EchoingECG Evaluations

We tested ECHOINGECG against ECG foundation models, ECG-CLIP [10], ECG-FM (2 versions) [26], and MEDBind [10], in ECHO prediction. Leveraging ECG as an advanced screening tool of traditionally ECHO-derived outcomes could enhance accessibility and early detection of cardiovascular disease [2,31]. To this end, we tested the predictive capacity on ECHO-derived variables: left ventricular ejection fraction (LVEF), severe left ventricular hypertrophy (SLVH), dilated left ventricular hypertrophy (DVH) using zero-shot (ZS), few-shot (FS), and fine-tune (FT) [5,21,31]. We also assess model performance on traditional ECG-text retrieval. For evaluating probabilistic models, we followed [7], where μ embedding was used for prediction, while σ^2 was used as an uncertainty measure. **Implementation Details:** ECG signals were normalized and augmented using standard approaches from [30]. Text inputs consisted of full-length notes, padded to a maximum length of 244 tokens. ECHO videos were processed, following ECHO-CLIP [6]. All modalities' final μ and $\log(\sigma^2)$ embedding dimension were 256. The λ for our loss (Section 2.3) was set to 0.9. Models were trained for 150 epochs, batch size 128, AdamW [24], weight decay of 1×10^{-1}, and learning rate of 4×10^{-4}. Experiments were conducted using a 48GB L40S GPU.

3   Experiments and Results

3.1   Datasets

The datasets used are presented in Table 1. We pretrained our models using MIMIC-ECG [15] and MIMIC-ECHO [14]. Patient level splits were maintained

Table 1. Overview of datasets used, tasks, and splits. Modalities column reflects the modality pairs obtained from each dataset. Tasks for each dataset are highlighted, where either pretrain, retrieval, ZS, FS, or FT was conducted with each dataset. [1]Subset of MIMIC-ECHO with paired ECGs based on `hadm_id`. [2]Only key variables of the ECHO report were provided. Demographics of these datasets can be found in [14,15,25]

Dataset	Modalities	Task	Train	Valid	Test
MIMIC-ECG	ECG/TXT	Pretrain/Retrieval/ZS/FS	88,291	12,065	24,644
MIMIC-ECHO[1]	ECHO/TXT	Pretrain	814	67	194
MUSIC ECG[2]	ECG	ZS/FT	512	-	125

Table 2. Zero-shot (ZS) and few-shot (FS), 10% and 100%, on LVEF<40% in MIMIC using ECGs. ECHO labels were derived from ECHO reports. [*]Uncertainty splits (low/high) based on entropy > 0.5. [†]Uncertainty splits in EchoingECG can be calculated directly on median-split of σ^2 vector, as in [7]. Balanced accuracy is reported.

Model	Prob?	All			Low Uncertainty			High Uncertainty		
		ZS	10%	100%	ZS	10%	100%	ZS	10%	100%
ECG-CLIP [10]	✗	56.3	64.5	<u>77.1</u>	-	66.2[*]	77.4[*]	-	62.8[*]	**76.8**[*]
MEDBind [10]	✗	<u>57.5</u>	<u>66.9</u>	76.2	-	61.0[*]	75.7[*]	-	**72.8**[*]	<u>76.7</u>[*]
ECG-FM$_M$ [26]	✗	-	55.0	66.8	-	58.4[*]	69.1[*]	-	51.6[*]	64.5[*]
ECG-FM$_P$ [26]	✗	-	61.5	70.7	-	64.2[*]	72.4[*]	-	58.8[*]	69.0[*]
EchoingECG (Ours)	✓	**63.2**	**69.3**	**78.9**	68.3[†]	**74.3**[†]	**84.4**[†]	58.1[†]	<u>64.2</u>[†]	73.4[†]

for all MIMIC datasets. **MIMIC-ECG** dataset contains 10-second recordings of 12-lead ECG signals sampled at 500Hz, which were downsampled to 100 Hz using a low-pass filter [20]. We used machine-generated and free-form clinical notes. When free-form text was accessible, it was prioritized; in its absence, machine-generated reports were subsequently utilized for text pairing. **MIMIC-ECHO** dataset contains ECHOs from MIMIC patients. For our experiments, we used the subset of MIMIC-ECHO, which is comprised of patients who had both an ECG and ECHO during their visit. We derived this subset based on `subject_id` and `admittime` that connects to MIMIC-ECG. **MIMIC-NOTE** was used to derive ECHO labels [17]. We linked patients who had both ECG and ECHOs with their corresponding discharge note using `hadm_id` and extracted ECHO findings following [13] using Llama3.1-instruct 8B [9]. The MUerte Subita en Insuficiencia Cardiaca **(MUSIC) ECG** dataset [25] comprises vectorcardiograms (VCGs) at 1000Hz. Each recording is associated with ECHO labels of SLVH, DLV, and LVEF. We used the Kors regression transformation [16] to convert the VCGs into 12-lead ECGs. The derived ECGs were also downsampled to 100 Hz [20].

3.2 Predicting ECHO-derived Labels using ECGs

MIMIC Results: We benchmarked EchoingECG on ECHO-related tasks within MIMIC, focusing on ZS and FS for predicting LVEF<40% using only ECG in Table 2. In ZS, we computed cosine distance between text prompts and

Table 3. Comparison with state-of-the-art for zero-shot (ZS) and fine-tune (FT) (100%) on ECHO-related classification in MUSIC dataset. MUSIC ECGs have longer time windows (20-minutes) vs. traditional ECGs (10-seconds). We report the start result, as it performed best on baseline models. Balanced accuracy is reported.

| Model	LVEF<40%		SLVH		DLV		SLVH+DLV		Ovr. Rank	
	ZS	FT	ZS	FT	ZS	FT	ZS	FT	ZS	FT
ECG-CLIP (start) [10]	64.1	58.8	58.9	59.4	60.4	58.7	61.3	84.9	5	7
MEDBind (start) [10]	64.2	60.2	56.3	62.0	61.1	61.5	62.6	85.5	4	5
ECG-FM$_M$ (start) [26]	-	63.2	-	76.0	-	**69.6**	-	87.2	-	3
ECG-FM$_P$ (start) [26]	-	64.8	-	60.8	-	62.4	-	74.4	-	5
EchoingECG (start)	66.4	**66.0**	76.0	77.5	64.0	57.6	62.4	89.0	2	2
EchoingECG (High σ^2)	65.0	60.0	77.4	76.8	65.6	51.2	60.6	89.0	3	4
EchoingECG (Low σ^2)	**68.0**	65.6	**79.2**	**77.6**	**70.4**	63.2	**66.4**	**91.4**	1	1
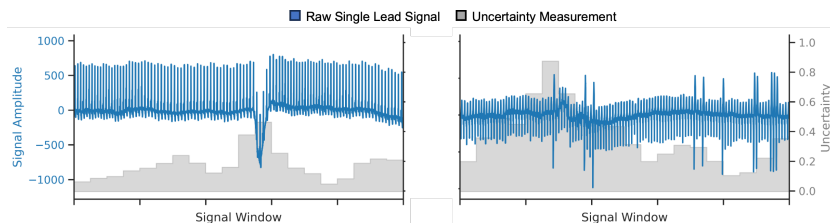
Fig. 2. Uncertainty in MUSIC dataset using EchoingECG. 10-second sliding window of ECG Lead I in uncertainty in the ECGs are visualized. Areas of signal instability are associated with **higher** σ^2, suggesting that σ^2 captures the uncertainty.

ECG [33], while FS used embeddings from frozen ECG encoders with linear probing [12]. Note, **ECG-FM** is unimodal. Thus, all **ECG-FM** ZS results are omitted. Additionally, Table 2 also highlights the impact of data uncertainty on predicting LVEF<40%, quantified via baseline methods (deterministic) and our method (probabilistic). **For deterministic:** we can quantify uncertainty in FS results by entropy $H[x] = p(x)\log_2 p(x) + (1 - p(x))\log_2 (1 - p(x))$ [1], where x represents the input and $p(x)$ is the softmax output from linear probing. For simplicity, we considered an entropy above 0.5 as high, while below was low uncertainty. **For probabilistic (ours):** we can split data with the σ^2 vector [7] that EchoingECG calculates for each embedding, where, a median split of the test set into high- and low-σ^2. EchoingECG predicts $Z_m \sim \mathcal{N}\left(\mu_m, \sigma_m^2\right)$; thus, the predicted σ^2 reflects uncertainty as it measures the spread of Z_m.

In both ZS and FS, EchoingECG outperformed foundation ECG models, highlighting the value of our method in improving ECHO-related task performance. Additionally, no clear relationship can be observed when splitting deterministic methods using entropy. In contrast, splitting via σ^2 highlights that EchoingECG can appropriately split these subsets. From Table 2, we observed that *lower* uncertainty (via σ^2 embedding) corresponded to better performance and vice versa–showing evidence of correct uncertainty estimation behavior.

Table 4. Top-k retrieval (100%) on **text-to-ECG retrieval**. We report the top-1 and top-10 recall (R@k). *Low vs. high uncertainty based on median-split of σ^2 embedding.

Model	Trained Modalities	All		Low Uncertainty*		High Uncertainty*	
		R@1	R@10	R@1	R@10	R@1	R@10
ECG-CLIP [10]	ECG+TXT	50.2	93.9	-	-	-	-
MEDBind [10]	ECG+TXT+CXR	53.6	**94.5**	-	-	-	-
EchoingECG (Ours)	ECG+TXT+ECHO	**67.0**	**94.5**	76.8	98.5	45.9	72.7

Table 5. Ablation on various tasks. For classification (ACC), we highlighted **ZS** results. **For all tasks:** all samples were considered, not just low-uncertainty, for fair comparisons. For MUSIC, start window was used for all models.

Methods	Prob?	MIMIC-ECHO LVEF (ACC)	MUSIC SLVH+DLV (ACC)	MIMIC-ECG Retrieval (R@1)
InfoNCE Loss	✗	56.8	61.0	53.2
+ ECHO Teacher	✗	61.2	62.1	52.9
PCME++ Loss	✓	62.0	61.7	**67.9**
+ ECHO Teacher (Ours)	✓	**63.2**	**62.4**	67.0

External MUSIC Results: We evaluated EchoingECG performance on external MUSIC dataset for both ZS and FT tasks–focusing on a use-case in handling more general ECGs (i.e. ECGs longer than the 10-second durations seen in training), in Table 3. Traditional methods typically process short windows to tackle longer ECGs—often extracted from the beginning or a random offset—to handle longer recordings [27]. We observed that using the start windows worked best for the baseline models reported herein. In contrast, EchoingECG leverages uncertainty estimation: higher values of σ^2 indicate greater dispersion, while lower uncertainty corresponds to more consistent signal characteristics. This uncertainty metric guides the selection of the most informative window, enabling our approach to outperform baseline models across multiple tasks consistently. Figure 2 visualizes this relationship, showing that noisy ECG segments exhibit higher uncertainty, whereas stable signals correspond to lower uncertainty.

3.3   Traditional Top-K Text-ECG Retrieval

We further tested Text-ECG retrieval using recall at top-K [10, 33] to measure alignment between ECGs and text (Table 4), EchoingECG remained competitive with other models–showing that adding our combined loss did not compromise ECG-text alignment. We also observed that ECGs with low uncertainty had better retrieval (Table 4), reinforcing the learned uncertainty in our approach.

3.4   Ablation

We conducted an ablation study comparing the performance of InfoNCE [29], PCME++ [7], and the proposed additional ECHO Teacher loss in EchoingECG.

Empirically, we observed minor differences when varying λ; notably, lower values preserved ECHO performance but decreased ECG-text retrieval accuracy. As shown in Table 5, InfoNCE loss performed poorly in all metrics, highlighting the limitation of deterministic methods in capturing the temporal and structural complexity of multimodal ECG models, where many-to-many relationships may exist. This observation is supported by the PCME++ loss results, which marked improvements and demonstrated a potential improvement in training due to the probabilistic nature of the loss. Finally, our proposed addition achieved the highest performance across all ECHO-based classification tasks. This underscores the guiding model with the ECHO Teacher and how our method could support using ECGs to emulate the structural and functional insights derived from ECHO.

4   Conclusion

Our proposed EchoingECG model demonstrated improved performance across ECHO-related tasks by leveraging probabilistic modeling and a foundational model as a teacher (i.e., ECHO-CLIP). Although training involves multiple stages, inference remains modular, allowing EchoingECG's ECG encoder to be deployed independently for efficient clinical use and light-weight applications.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information fusion **76**, 243–297 (2021)
2. Al Younis, S.M., Hadjileontiadis, L.J., Khandoker, A.H., Stefanini, C., Soulaidopoulos, S., Arsenos, P., Doundoulakis, I., Gatzoulis, K.A., Tsioufis, K.: Prediction of hf patients with distinct left ventricular ejection fraction levels using circadian ecg features and machine learning. Plos one **19**(5), e0302639 (2024)
3. Attia, Z.I., Kapa, S., et al.: Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature medicine **25**(1), 70–74 (2019)
4. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Proceedings of the 3rd international conference on knowledge discovery and data mining. pp. 359–370 (1994)
5. Bhave, S., Rodriguez, V., Poterucha, T., Mutasa, S., Aberle, D., Capaccione, K.M., Chen, Y., Dsouza, B., Dumeer, S., Goldstein, J., et al.: Deep learning to detect left ventricular structural abnormalities in chest x-rays. European Heart Journal **45**(22), 2002–2012 (2024)
6. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine pp. 1–8 (2024)

7. Chun, S.: Improved probabilistic image-text representations. In: The Twelfth International Conference on Learning Representations (2024)
8. Chun, S., Oh, S.J., De Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8415–8424 (2021)
9. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
10. Gao, Y., Kim, S., Austin, D.E., McIntosh, C.: Medbind: Unifying language and multimodal medical data embeddings. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 218–228. Springer (2024)
11. Gatzoulis, K.A., Arsenos, P., Trachanas, K., Dilaveris, P., Antoniou, C., Tsiachris, D., Sideris, S., Kolettis, T.M., Tousoulis, D.: Signal-averaged electrocardiography: Past, present, and future. Journal of Arrhythmia **34**(3), 222–229 (2018)
12. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
13. Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L.H., Hao, X., Jaber, B., Reddy, S., Kartha, R., et al.: Llms accelerate annotation for medical information extraction. In: Machine Learning for Health (ML4H). pp. 82–100. PMLR (2023)
14. Gow, B., Pollard, T., Greenbaum, N., Moody, B., Johnson, A., Herbst, E., Waks, J.W., Eslami, P., Chaudhari, A., Carbonati, T., Berkowitz, S., Mark, R., Horng, S.: Mimic-iv-echo: Echocardiogram matched subset. Physionet (2024)
15. Gow, B., Pollard, T., Nathanson, L.A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Berkowitz, S., Moukheiber, D., Eslami, P., et al.: Mimic-iv-ecg-diagnostic electrocardiogram matched subset. Physionet (2023)
16. Jaros, R., Martinek, R., Danys, L.: Comparison of different electrocardiography with vectorcardiography transformations. Sensors **19**(14), 3072 (2019)
17. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023)
18. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. vol. 1, p. 2 (2019)
19. Khan, F., Yu, X., Yuan, Z., Rehman, A.U.: Ecg classification using 1-d convolutional deep residual neural network. Plos one **18**(4), e0284791 (2023)
20. Kher, R., et al.: Signal processing techniques for removing noise from ecg signals. J. Biomed. Eng. Res **3**(101), 1–9 (2019)
21. Kusunose, K.: Transforming echocardiography: The role of artificial intelligence in enhancing diagnostic accuracy and accessibility. Internal Medicine (2025)
22. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
23. Li, J., Liu, C., Cheng, S., Arcucci, R., Hong, S.: Frozen language model helps ecg zero-shot learning. In: Medical Imaging with Deep Learning. PMLR (2024)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018)
25. Martın-Yebra, A., de Luna, A.B., Vázquez, R., Caminal, P., Laguna, P., Martınez, J.P.: The music database: Sudden cardiac death in heart failure patients. Physionet (2024)
26. McKeen, K., Oliva, L., Masood, S., Toma, A., Rubin, B., Wang, B.: Ecg-fm: An open electrocardiogram foundation model. arXiv preprint arXiv:2408.05178 (2024)

27. Mehari, T., Strodthoff, N.: Advancing the state-of-the-art for ecg analysis through structured state space models. arXiv preprint arXiv:2211.07579 (2022)
28. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
30. Raghu, A., Shanmugam, D., Pomerantsev, E., Guttag, J., Stultz, C.M.: Data augmentation for electrocardiograms. In: Conference on Health, Inference, and Learning. pp. 282–310. PMLR (2022)
31. Sangha, V., Nargesi, A.A., Dhingra, L.S., Khunte, A., Mortazavi, B.J., Ribeiro, A.H., Banina, E., Adeola, O., Garg, N., Brandt, C.A., et al.: Detection of lvef from electrocardiographic images. Circulation **148**(9), 765–777 (2023)
32. Strodthoff, N., Wagner, P., Schaeffter, T., Samek, W.: Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. IEEE journal of biomedical and health informatics **25**(5), 1519–1528 (2020)
33. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887 (2022)
34. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1179–1189 (2023)
35. Yao, X., Rushlow, D.R., Inselman, J.W., McCoy, R.G., et al.: Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nature Medicine **27**(5), 815–819 (2021)