Multi-Modal Contrastive Learning across Cardiac Diagnostics

Bryan He[1], Milos Vukadinovic[2,3], Grant Duffy[3], James Zou[1,4,5], David Ouyang[3,6]

1. Department of Computer Science, Stanford University
2. Department of Bioengineering, University of California Los Angeles
3. Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center
4. Department of Biomedical Data Science, Stanford University
5. Department of Electrical Engineering, Stanford University
6. Division of Artificial Intelligence in Medicine, Department of Medicine, Cedars-Sinai Medical Center

Introduction

Cardiovascular diseases range from electrical conduction abnormalities to myocardial dysfunction and structural abnormalities leading to abnormal blood flow. Due to the diversity of cardiovascular conditions, many different diagnostic tools and imaging modalities are needed to gather comprehensive information about an individual patient's cardiac physiology. To assess the heart, echocardiograms assess heart motion, electrocardiograms (ECGs) evaluate electrical conduction, angiograms evaluate coronary artery blockages, and chest X-rays (CXRs) can identify fluid buildup (examples in Figure 1a). While the data differ greatly in appearance (ranging from waveforms to images and videos), these modalities share complementary information that can uncover information not apparent in another modality. In parallel, diseases frequently result in findings across diagnostic tests (a coronary artery blockage results in electrical changes seen on ECG as well as myocardial dysfunction on echocardiogram videos).

While all these cardiovascular diagnostic modalities encode information about the heart, prior work in machine learning predominantly focuses on one modality at a time. In applying deep learning to cardiology, models have been developed to predict left ventricular ejection fraction from echocardiograms[1], identify rhythm disorders from ECGs[2], interpret angiographic coronary artery stenosis from angiograms[3], and estimate cardiac size and pulmonary edema from CXRs[4].

In clinical practice, physicians use orthogonal diagnostic tests because findings can be more apparent in a particular modality than another (a subtle ECG change can be visualized as an obvious occlusion on a coronary angiogram). However, modalities vary in cost, availability, and invasiveness. Given the close relationship of findings for the same disease between these diagnostic modalities, it is important to have a shared multimodal representation between them.

A recent development in machine learning is the use of contrastive learning to link images and corresponding text captions[5,6]. These contrastive methods have also been used to link medical images and text reports in several different areas, including radiology[7–9], pathology[10,11], and cardiac ultrasound[12].

In this work, we introduce Contrasting Learning Embedding Representation of Cardiology (CLERC), a multimodal model linking diagnostic modalities across cardiovascular testing. CLERC builds on the key insight that information from the same patient, even of different

modalities, are closely related and, as a result, can be treated as positive pairs in contrastive learning. CLERC learns encoders for each of the diagnostic modalities, as well as for the corresponding text reports. We show that the representations learned by CLERC can perform retrieval across modalities and that the representations can be used to perform important clinical diagnostic tasks.

Results

Curating a multimodal cardiology dataset

We collected a dataset of 382,803 patients from Cedars-Sinai Medical Center from 2005 to 2022. The dataset consisted of four diagnostic modalities (echocardiograms, EKGs, angiograms, and CXRs), along with the text reports of clinician interpretations for each of the diagnostic modalities. While sharing some similar vocabulary, text reports vary tremendously across modalities, so each text report language is treated as an additional modality. In the dataset, 102,480 patients had echocardiograms, 307,559 patients had EKGs, 17,338 patients had angiograms, and 226,835 patients had CXRs. The number of patients with each pair of modalities is given in Table 1, and the demographics of the patients in the dataset are given in Table 2.

Learning multimodal representations with contrastive learning

CLERC is a medical model linking echocardiogram videos, ECG waveforms, angiogram videos, and CXR images, along with the corresponding clinician text reports into a shared latent space. CLERC trains separate video convolutional neural networks as encoders for echocardiogram and angiogram videos, a 1-D convolutional neural network as the encoder for the ECG waveforms, a vision transformer as the encoder for CXR images, and a shared transformer for all of the text report modalities. CLERC is trained using a contrastive learning loss between all pairs of distinct modalities. In each batch, a set of inputs from each modality from distinct patients is given to CLERC, and the contrastive loss trains the encoders to move embeddings from the same patient closer together while moving embeddings from distinct patients further apart.

Cross-modal retrieval

The latent space learned by CLERC is shared across modalities, allowing us to identify related concepts or similar findings across modalities. To assess this ability, we use CLERC to match samples across modalities measured from the same patient (Figure 1b).

For each distinct pair of modalities, we select all patients that have both modalities available and select one sample from each modality. If multiple samples are available, the pair of samples that are closest temporally are selected. Next, CLERC is provided one sample from the first modality, and ranks all samples from the second modality by cosine similarity to attempt to identify the matched sample from the same patient. This retrieval process is repeated for all samples from the first modality, and the median percentile of the retrievals is used to measure CLERC's performance. This task is then repeated for all distinct pairs of modalities (Table 3).

The cross-modal retrievals fall into several groups: (1) diagnostic modality and corresponding text report (ex. echocardiogram video used to retrieve echocardiogram text report), (2) diagnostic modality to retrieve distinct diagnostic modality (ex. echocardiogram video to retrieve EKG waveform), (3) diagnostic modality to retrieve mismatched text report (ex. echocardiogram video to retrieve EKG text report), and (4) text report to retrieve distinct text report (ex. echocardiogram text report to retrieve EKG text report).

First, we find that diagnostic modality and corresponding text reports can be retrieved with high accuracy, with all pairs resulting in a median percentile of at most 13.7. Other than the angiogram modalities, which had the smallest amount of training data and the least variation between patients, all other diagnostic-to-text and text-to-diagnostic retrievals resulted in a median percentile of at most 4.1.

Next, we find that CLERC is able to accurately retrieve across the diagnostic modalities, with EKG waveform-to-angiogram video retrieval resulting in the worst performance with a median percentile of 8.6. Similarly, if the angiogram modalities are excluded, the worst performance remaining is 3.1.

We additionally find that diagnostic modality to mismatched text reports and text-to-text retrieval perform well above random. However, retrieval using the raw diagnostic data performs better in all cases, suggesting that the raw diagnostic data contains a substantial amount of information not present in the text reports.

Predictive tasks

Next, we assess the ability of CLERC's embeddings to predict various clinical measurements from each modality (Table 4). We assess this predictive ability in two settings: zero-shot and linear probing. In the zero-shot setting (Figure 1c), there is no explicit training; instead, the text encoder is used to generate an embedding for a prompt corresponding to each predictive task, and its cosine similarity with the embedding of the corresponding diagnostic modality is used as the prediction (full list of prompts in Supplementary Table 1). In the linear probing setting, the training set is used to train a linear regression or logistic regression model over the embeddings of the diagnostic modality.

In the zero-shot setting, CLERC's embeddings predict all tasks well above the random baseline. In the linear probing setting, the performance improves on all tasks. The echocardiogram and EKG prediction tasks perform comparably to prior fully supervised models. The angiogram predictive tasks are the most challenging: the training set is the smallest, the views are less standardized than those of echocardiograms, and occlusions are not visible in all videos.

Cross-modality predictions

The shared representation of CLERC allows predictions of measurements using other modalities, which can allow cheaper and faster modalities to estimate measurements from the more difficult-to-obtain modalities. We also find that the embeddings are closely linked to common demographic attributes.

First, we find that the CLERC's representations for all modalities are closely linked to the age (R^2 of at least 0.529 across all modalities) and gender (AUROC of at least 0.905 across all modalities) of patients. Several potential cross-modal predictions are also of potential interest. First, we find EKGs are able to estimate LVEF with an AUROC of 0.415. While the performance is lower than that of using echocardiograms, EKGs are much faster to obtain, potentially allowing EKGs to be used as a screening tool. Similarly, echocardiograms perform relatively well in predicting occlusion, consolidation, and edemas, potentially enabling their use as a screening or triage tool to reduce the invasiveness of angiograms or the radiation of CXRs.

Discussion

Leveraging contrastive learning to inform the relationship between medical diagnostics for the same patient, multimodal models can recapitulate the relationships between complementary medical tests. Different diagnostic modalities provide distinct insights into the heart, and a joint embedding from a foundation model can highlight important clinical findings of cardiac diseases. In this study, we utilize the comprehensive cardiac testing of over 300,000 patients from an academic medical center to train CLERC, a multimodal cardiovascular foundation model.

Methods

Data curation

We collected a dataset consisting of 382,803 distinct patients from Cedars-Cinai Medical Center between 2005 and 2022. The dataset was split into training, validation, and test sets by patient. The training set consisted of 306,242 patients, the validation set consisted of 38,281 patients, and the test set consisted of 38,201 patients.

For each patient, we collected all available echocardiogram videos, echocardiogram text reports, EKG waveforms, EKG text reports, angiogram videos, angiogram text reports, chest X-ray images, and chest X-ray text reports. The echocardiogram videos were cropped to a tight square around the scanning sector and scaled to 112 x 112 pixels. The EKG we processed as 12-channel waveforms at 500 Hz. The angiogram videos were scaled to 112 x 112 pixels. The chest X-ray images were cropped to 256 x 256 pixels.

This research was approved by the Cedars-Sinai Medical Center Institutional Review Boards.

Encoders

For each modality, we use a deep learning model to encode the samples into 512-dimensional embeddings. We use separate weights for the echocardiogram, EKG, angiogram, and CXR modalities, and a shared set of weights for all text modalities.

We used the R(2+1)D-18 architecture, a convolutional neural network with decomposed spatial and temporal convolutions, as echocardiogram and angiogram encoders[13]. The models were initialized with Kinetics-400 weights[14], and trained separately after initialization. For the echocardiogram videos, clips of 16 frames were generated by sampling every other frame (videos were natively 30 frames per second). For the angiogram videos, clips of 16 frames were generated by sampling every frame (videos were natively 15 frames per second). For data augmentation, both echocardiogram and angiogram videos were padded with 12 pixels per side and cropped back to 112 x 112 pixels.

We used a 1D convolutional neural network as the EKG encoder[15]. The model was initialized with random weights. Clips of 2500 samples were sampled as input (waveform is natively 5000 samples at 500 Hz).

We used the ViT-B/32 architecture, a vision transformer, as the CXR encoder[16]. The model was initialized with weights trained by CLIP[6]. For data augmentation, random 224 x 224-pixel crops were used as input.

We used a masked self-attention transformer as the encoder for all text modalities[17]. The model was initialized with weights trained by CLIP[6]. No data augmentation was applied. We used the CLIP byte pair encoding as the tokenizer[18]**.**

Model training

We train CLERC using the sum of the CLIP losses between all pairs of modalities present in a batch. We use a stochastic gradient descent optimizer with an initial learning rate of 1e-4, a momentum of 0.9, and batch size of 32 for 60 epochs. The learning rate is decayed by a factor of 10 every 20 epochs. The epoch with the lowest validation loss is selected as the final model.

During training, each patient appears once per epoch. If a patient has multiple samples from a modality, a random sample is selected for that batch. We group the patients into batches based on the modalities available for that patient. The remaining unmatched patients are randomly grouped into batches.

Retrieval

We performed pairwise retrieval for each distinct pair of modalities. For each pair, we sampled all patients with samples from both modalities. If a patient had multiple samples from either modality, the pair of samples that were the closest temporally was selected. For all samples, the corresponding encoder is used to generate an embedding. Then, for each sample of the first modality, the cosine similarity is used to rank all samples of the second modality to retrieve the sample from the matching patient. This process is repeated in the opposite direction and then for all pairs of modalities.

Zero-shot predictions

We perform zero-shot prediction tasks for both regression and binary classification tasks. For both tasks, we use the encoder to compute the embeddings for all test samples of the selected modality.

For binary classification tasks, we used a single text prompt and calculated an embedding for it using the text encoder. The cosine similarity between the embedding of the sample and the embedding of the text prompt was used as the prediction for the label.

For regression tasks, we used a list of text prompts sweeping the normal range of values for the parameter. To predict the parameter, we use the embedding of the sample to calculate a probability distribution over the values, and use the expected value of this distribution as the estimate. To do this, we generate an embedding for each text prompt using the text encoder. We then calculate the cosine similarity between the sample embedding and the embedding of each prompt. We scale the cosine similarities by the temperature learned during training and take the softmax of the resulting values to generate the distribution.

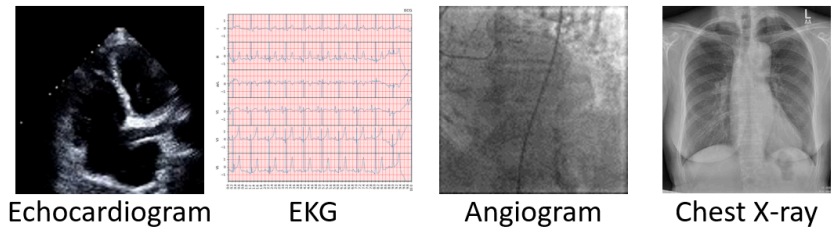
For both binary classification and regression tasks, we average the predictions over all samples in a study.

Linear Probing

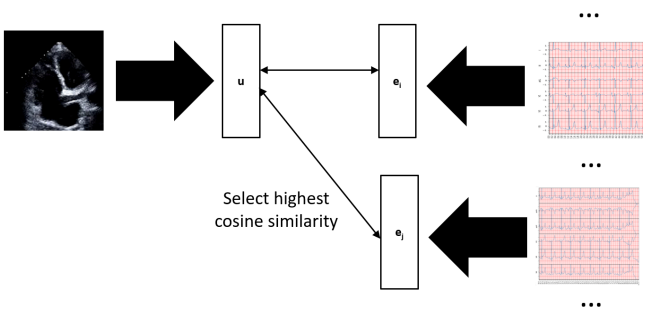
For both regression and binary classification tasks, we calculate embeddings for all samples in the training, validation, and test sets. The training embeddings are used to train a linear regression model for regression tasks and used to train a logistic regression model for binary classification tasks. The validation set is used to select a regularization value, and the test set is used to report final performance.

For both binary classification and regression tasks, we average the predictions over all samples in a study.

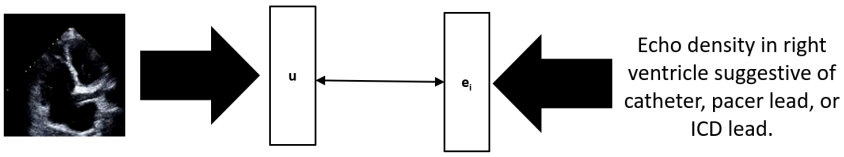
For the cross-modal predictions, we select patients with the modality used for making the prediction, along with the modality used to determine the ground truth label. If multiple labels are available, the closest label temporally is used.



(a)



(b)



(c)

Figure 1: (a) Example diagnostic modalities used in cardiology. (b) Cross-modal retrieval using CLERC. (c) Zero-shot predictions using CLERC.

Table 1: Number of patients with each pair of modalities.

	Echo	EKG	Angio	CXR
Echo	102,480			
EKG	89,873	307,559		
Angio	14,646	16,698	17,338	
CXR	76,803	161,122	13,802	226,835

Table 2: Patient demographics in the dataset.

	Total	Train	Validation	Test
Patients	382,803	306,242	38,281	38,280
Age (mean ± std)	64.6 ± 17.9	64.5 ± 17.9	64.8 ± 17.9	65.0 ± 17.7
Female (%)	192,699 (50.3%)	154,168 (50.3%)	19,159 (50.0%)	19,372 (50.6%)
Race (%)				
Non-Hispanic White	214,733 (56.1%)	171,688 (56.1%)	21,562 (56.3%)	21,483 (56.1%)
Black	55,599 (14.5%)	44,402 (14.5%)	5,539 (14.5%)	5,658 (14.8%)
Hispanic	53,526 (14.0%)	42,957 (14.0%)	5,310 (13.9%)	5,259 (13.7%)
Asian	26,714 (7.0%)	21,410 (7.0%)	2,646 (6.9%)	2,658 (6.9%)
Other	16,476 (4.3%)	13,198 (4.3%)	1,644 (4.3%)	1,634 (4.3%)
Unknown	14,094 (3.7%)	11,265 (3.7%)	1,406 (3.7%)	1,423 (3.7%)
Pacific Islander	934 (0.2%)	735 (0.2%)	103 (0.3%)	96 (0.3%)
Native American	727 (0.2%)	587 (0.2%)	71 (0.2%)	69 (0.2%)
Patients with modality				
Echo	102,480	81,800	10,344	10,336
EKG	307,559	245,920	30,925	30,714
Angiogram	17,338	13,850	1,750	1,738
CXR	226,835	181,400	22,648	22,787
Number of studies				
Echo	242,567	193,630	24,653	24,284
EKG	1,262,750	1,009,295	126,142	127,313
Angiogram	22,836	18,255	2,305	2,277
CXR	792,144	633,165	79,693	79,286
Total samples				
Echo	985,009	786,534	99,573	98,902
EKG	1,262,750	1,009,295	126,142	127,313
Angiogram	255,914	205,063	25,820	25,031
CXR	2,050,776	1,640,002	206,061	204,713

Table 3: Median percentile for retrieval across modalities in the test set.

			Retrieved Modality							
			Echo		EKG		Angio		CXR	
			Video	Text	Waveform	Text	Video	Text	Image	Text
Query Modality	Echo	Video		2.8	1.3	10.1	3.2	11.6	1.7	7.0
		Text	2.8		13.4	22.3	12.9	16.9	15.3	14.0
	EKG	Waveform	1.4	13.8		2.6	8.4	22.7	3.2	14.7
		Text	9.2	20.3	2.2		17.7	28.9	19.0	21.9
	Angio	Video	3.0	11.5	7.8	20.1		12.2	5.1	9.8
		Text	13.4	14.5	20.8	30.6	13.7		21.2	19.2
	CXR	Image	1.8	15.0	3.0	25.6	5.0	20.3		4.1
		Text	7.1	14.0	14.4	29.6	10.1	20.3	4.1	

Table 4: Zero-shot and linear probing of CLERC embeddings using standard modalities for predicting various measurements.

Modality	Task	Metric	Zero-Shot	Linear Probing
Echo	LVEF	R^2	0.744	0.794
	Pacemaker	AUROC	0.885	0.951
EKG	RBBB	AUROC	0.958	0.977
	LBBB	AUROC	0.968	0.982
	Pacemaker	AUROC	0.947	0.974
Angiogram	LAD Occlusion	AUROC	0.678	0.703
	RCA Occlusion	AUROC	0.642	0.756
	LCX Occlusion	AUROC	0.673	0.719
CXR	Consolidation	AUROC	0.729	0.791
	Edema	AUROC	0.762	0.864

Table 5: Cross modal predictions with linear probing.

Task	Metric	Modality			
		Echo	EKG	Angiogram	CXR
Age	R^2	0.709	0.599	0.529	0.717
Gender	AUROC	0.966	0.905	0.963	0.984
LVEF	R^2	0.794	0.415	0.380	0.284
Pacemaker	AUROC	0.951	0.974	0.933	0.924
RBBB	AUROC	0.838	0.977	0.659	0.684
LBBB	AUROC	0.918	0.982	0.752	0.728
LAD Occlusion	AUROC	0.738	0.600	0.703	0.595
RCA Occlusion	AUROC	0.778	0.696	0.756	0.615
LCX Occlusion	AUROC	0.572	0.654	0.719	0.594
Consolidation	AUROC	0.711	0.678	0.610	0.791
Edema	AUROC	0.740	0.743	0.753	0.864

1. Ouyang, D. *et al.* Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
2. Hughes, J. W. *et al.* A deep learning-based electrocardiogram risk score for long term cardiovascular death and disease. *Npj Digit. Med.* **6**, 169 (2023).
3. Avram, R. *et al.* CathAI: fully automated coronary angiography interpretation and stenosis estimation. *Npj Digit. Med.* **6**, 142 (2023).
4. Ueda, D. *et al.* Artificial intelligence-based model to classify cardiac functions from chest radiographs: a multi-institutional, retrospective model development and validation study. *Lancet Digit. Health* **5**, e525–e533 (2023).
5. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive Learning of Medical Visual Representations from Paired Images and Text.
6. Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision.
7. Chambon, P. *et al.* RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. Preprint at http://arxiv.org/abs/2211.12737 (2022).
8. Zhang, X., Wu, C., Zhang, Y., Xie, W. & Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat. Commun.* **14**, 4542 (2023).
9. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).
10. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).
11. Lu, M. Y. *et al.* Towards a Visual-Language Foundation Model for Computational Pathology. Preprint at http://arxiv.org/abs/2307.12914 (2023).
12. Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Multimodal Foundation Models For Echocardiogram Interpretation. Preprint at https://doi.org/10.48550/arXiv.2308.15670 (2023).

13. Tran, D. *et al.* A Closer Look at Spatiotemporal Convolutions for Action Recognition. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 6450–6459 (IEEE, Salt Lake City, UT, 2018). doi:10.1109/CVPR.2018.00675.

14. Kay, W. *et al.* The Kinetics Human Action Video Dataset. Preprint at http://arxiv.org/abs/1705.06950 (2017).

15. Ouyang, D. *et al.* Electrocardiographic deep learning for predicting post-procedural mortality: a model development and validation study. *Lancet Digit. Health* **6**, e70–e78 (2024).

16. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at http://arxiv.org/abs/2010.11929 (2021).

17. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners.

18. Sennrich, R., Haddow, B. & Birch, A. Neural Machine Translation of Rare Words with Subword Units. in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1715–1725 (Association for Computational Linguistics, Berlin, Germany, 2016). doi:10.18653/v1/P16-1162.

Supplementary Table 1: Text prompts used for zero-shot predictions.

Task	Prompt
LVEF	The left ventricular ejection fraction is estimated to be [EF]%. LV ejection fraction is [EF]%. EF $= 20, 25, \ldots, 80$
Pacemaker (Echo)	Echo density in right ventricle suggestive of catheter, pacer lead, or ICD lead.
RBBB	Right bundle branch block.
LBBB	Left bundle branch block.
Pacemaker (EKG)	Pacemaker.
LAD Occlusion	Conclusion: 1. LAD had an occlusion. 100% occluded left ascending artery.
RCA Occlusion	Conclusion: 1. RCA had an occlusion. 100% occluded right coronary artery.
LCX Occlusion	Conclusion: 1. LCX had an occlusion. 100% occluded left circumflex artery.
Consolidation	There is consolidation in the lung. Consolidation is unchanged.
Edema	There is a pulmonary edema in the lung. Edema is unchanged.