**A Study on a Multimodal Fusion-Based Cardiac Disease Diagnosis System: Leveraging ECG and ECHO Data from the MIMIC Databases**

Zhaoyang YE

**Abstract**

This study develops a multimodal disease diagnosis system using ECG and ECHO data from the MIMIC-III and MIMIC-IV databases, which include detailed clinical and physiological records from ICU patients. Identifying the root cause of left ventricular hypertrophy (LVH) can be challenging due to the overlap in clinical symptoms and cardiac morphology across various conditions. Distinguishing individuals with hypertrophic cardiomyopathy (HCM) is of major importance for family screening and the prevention of sudden death. The research involves preprocessing and feature extraction to ensure data quality. Convolutional neural networks (CNN) are employed to analyze ECG signals, while recurrent neural networks (RNN) with attention mechanisms process ECHO data. These data are then integrated using advanced multimodal fusion techniques. The system is validated across multiple cohorts, including internal and prospective validation sets. The results demonstrate that the multimodal approach significantly improves the accuracy of diagnosing cardiac diseases, such as arrhythmias and heart failure, compared to traditional single-modality methods. This highlights the potential of multimodal data fusion in enhancing clinical diagnostic accuracy and advancing medical AI applications.

**Keywords: ECG, ECHO, Deep Learning, Multimodal Machine Learning, Disease Diagnosis**

**Introduction**

Deep learning models specifically focused on single modalities in cardiology have shown impressive results for arrhythmia detection, age, and other clinical actionable insights [1]. Previously Ko *et al*., focused on using convolutional neural networks (CNN) for ECG interpretation with respect to HCM [2]. They showed high discriminatory power in classifying HCM against a background population of left ventricular hypertrophy by ECG alone. However, approximately 28-30% of HCM cases had concurrent hypertension, inhibiting a direct comparison of possible distinction between HCM and hypertension. Zhang *et al* [3] focuses exclusively on echocardiograms in a fully automated approach to disease detection. Our method differs in three important ways, first we consider both ECG and echocardiogram jointly to make a classification prediction in differentiating between HCM and hypertension. Secondly, LVH-fusion model architecture differs significantly from the aforementioned study. We explored model architectures with variable integration of temporal convolutions instead of an image-based 2D CNN which operates on individual frames of the video. Empirical studies have shown the benefits of different spatiotemporal convolutions for video-based classification over 2D CNNs which are unable to model temporal information and motion patterns, which one would deem to be critical aspects for correct video analysis [4]. Additionally, two different video views were necessary for detection of HCM, our method holds high discriminatory power using only one video view. To date, deep learning research addressing non-pulmonary hypertension detection using both electrocardiogram and echocardiogram was unknown.

One previous approach successfully used both ECG and echocardiogram data individually with a stepwise approach to diagnosis of cardiac amyloidosis, [5] whereas here we focus on fusion method applications of multimodal deep learning of electrocardiograms and echocardiograms together.

Multimodal data analysis has gained increasing importance in medical research for improving the precision and reliability of disease diagnosis. With the rapid expansion of medical data and the advancement of artificial intelligence, approaches based on multimodal data are receiving growing attention. Traditional single-modality diagnostic methods often fail to capture the complexity of certain medical conditions, prompting a shift towards combining multiple data types, such as ECG and ECHO, for more accurate cardiac disease diagnosis. [6][7]

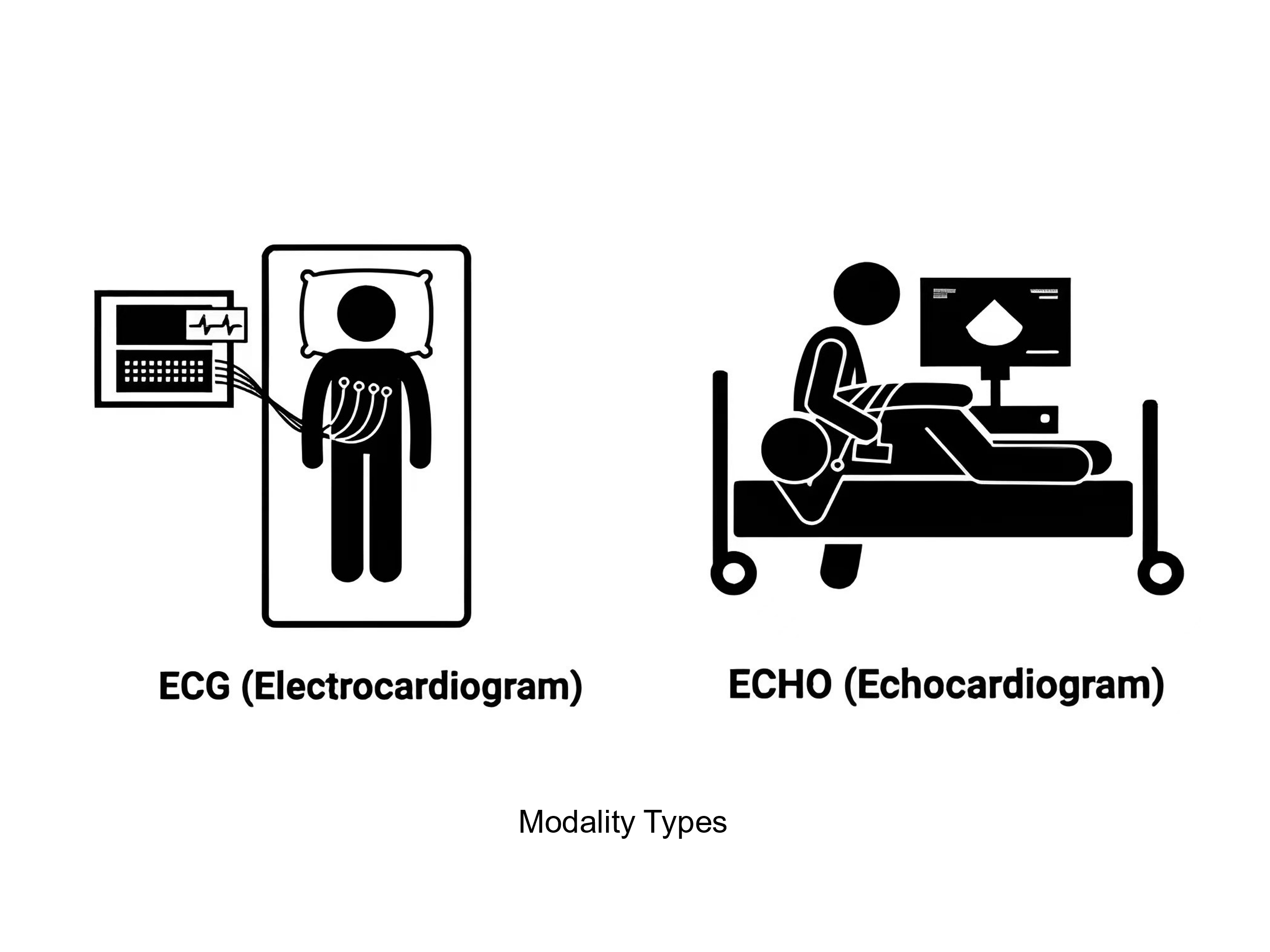


Figure 1 ECG and ECHO

However, effectively integrating diverse medical data types remains a significant challenge. This study aims to address this by developing a robust multimodal disease diagnosis system utilizing ECG and ECHO data from the MIMIC-III and MIMIC-IV databases. [8] Unlike previous research, which often focuses on single-modality analysis, this study explores the deep integration of multimodal data using convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with attention mechanisms. [9] The model's generalizability and robustness are ensured through comprehensive validation across multiple cohorts.

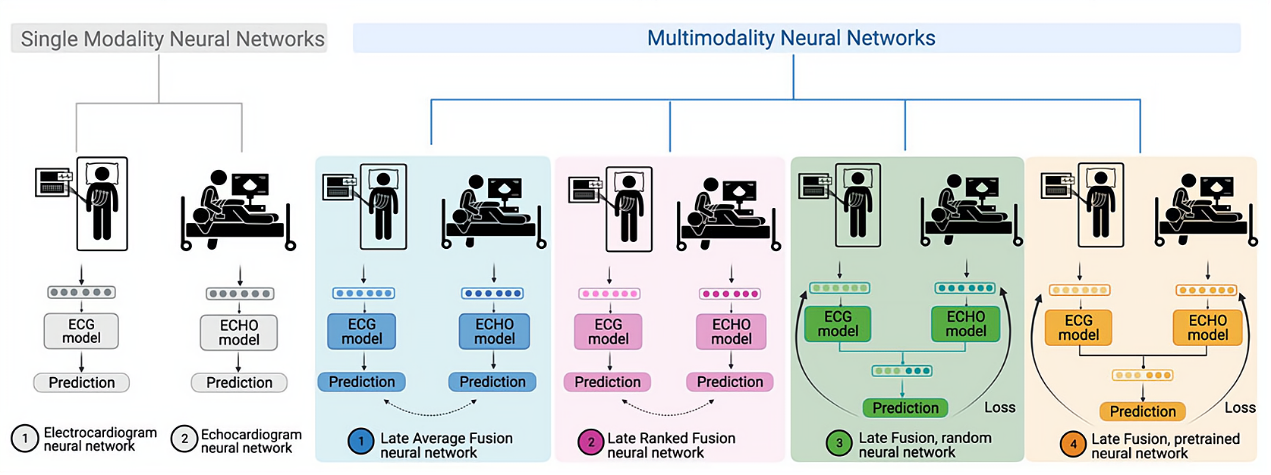


Figure 2 An Overview of the Model

Recent advancements in multimodal data analysis have demonstrated its potential, particularly in fields like medical imaging and natural language processing. Multimodal approaches have shown superior performance compared to single-modality methods, especially in cardiac disease diagnosis, where integrating various physiological signals offers a more complete view of a patient's health. Despite the progress, studies that deeply integrate ECG and ECHO data using large-scale datasets are still scarce. [10][11] This research aims to fill this gap by developing a comprehensive multimodal diagnostic system.

**Method**

**1. Data Collection and Preprocessing**

The study leverages the MIMIC-III and MIMIC-IV databases, which are rich in clinical and physiological data from ICU patients at Beth Israel Deaconess Medical Center. The focus is on extracting high-quality electrocardiogram (ECG) and echocardiogram (ECHO) data. Specific inclusion criteria are applied to ensure that only data meeting certain clinical relevance and quality thresholds are utilized. This selection process includes filtering out incomplete records, ensuring temporal alignment between ECG and ECHO data, and normalizing the data to account for differences in patient demographics, such as age and gender.

Once the relevant data is extracted, rigorous preprocessing steps are applied. For the ECG data, signal denoising is performed using methods like wavelet transform or median filtering to remove noise and artifacts, ensuring a clean signal for analysis. The signals are then segmented into time windows of interest, and features such as QRS complex, heart rate variability (HRV), and ST-segment deviation are extracted. These features are essential for identifying various cardiac conditions.

For the ECHO data, frames are extracted from the video recordings at specific cardiac cycle phases, such as end-diastole and end-systole. Image preprocessing includes resizing, normalization, and data augmentation techniques (e.g., rotation, flipping) to increase the robustness of the model. Feature extraction involves calculating parameters such as ejection fraction, wall motion abnormalities, and chamber dimensions, which are critical indicators of cardiac function.

**2. Model Architecture**

The study employs two distinct neural network architectures tailored to the different data modalities:

ECG Analysis with Convolutional Neural Networks (CNNs):

The ECG data is fed into a deep convolutional neural network. The CNN is designed with multiple convolutional layers followed by pooling layers to capture the spatial and temporal features of the ECG signals. The mathematical formulation for the convolutional operation is as follows:



Here,  is the output of the -th layer, is the input feature map from the previous layer,  are the convolutional weights, is the bias term, and is the activation function (e.g., ReLU). The CNN captures complex patterns in the ECG data, such as arrhythmias and ischemic changes.

**ECHO Analysis with Recurrent Neural Networks (RNNs) and Attention Mechanisms:** The ECHO data, represented as sequences of frames, is analyzed using a recurrent neural network equipped with attention mechanisms. The RNN captures the temporal dependencies within the sequences, while the attention mechanism allows the model to focus on specific frames or phases that are most relevant to the diagnosis. The RNN's operation is governed by:



where  ​is the hidden state at time step ,  and  are weight matrices,  is the input at time , and  is the bias term. The attention mechanism computes a context vector  as:



where  are attention weights, learned during training, determining the importance of each hidden state ​.

**3. Multimodal Fusion**

The integration of ECG and ECHO data is achieved through multimodal fusion, which combines the outputs from the CNN and RNN models into a unified framework. This fusion occurs at the decision level, where the attention mechanisms from both models generate weighted context vectors, which are then concatenated and passed through fully connected layers to produce the final diagnostic output. The fusion process can be mathematically described as:



whereand are the context vectors from the ECG and ECHO models, respectively, and ​ represents the fusion function, typically a series of fully connected layers followed by a softmax activation function to yield the final class probabilities.

**4. Experimental Design and Validation**

The experiments are structured across three distinct stages to ensure robust model performance:

Development Cohort: This cohort is used for model training. The data is split into training and validation sets, and the model is trained using techniques such as cross-entropy loss minimization and Adam optimization. The model's performance is monitored using metrics like accuracy, precision, recall, and F1-score.

Internal Validation Cohort: This cohort consists of data from the same database but is not used during training. The internal validation assesses how well the model performs on unseen data from the same distribution, helping to tune hyperparameters and prevent overfitting.

Prospective Validation Cohort: To evaluate the model's generalization capability, a prospective validation is performed on a new set of data, possibly from a different time period or a different subset of patients. This stage is crucial for determining the real-world applicability of the model.

The proposed multimodal fusion system's performance is benchmarked against traditional single-modality methods, demonstrating superior accuracy and robustness in diagnosing complex cardiac conditions such as arrhythmias, cardiomyopathies, and heart failure. The results highlight the efficacy of multimodal data integration in clinical diagnostic tools, paving the way for future advancements in medical AI.

**Results**

The performance of the multimodal fusion model is evaluated across the various validation cohorts, with a particular focus on its generalizability and robustness, especially in the external validation cohort. The model's results are compared with traditional single-modality diagnostic methods, demonstrating the clear advantages of multimodal fusion in improving diagnostic accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **auROC** | **auPRC** | **F1-score** | **Precision** |
| **Multimodal**  **(ECG+ECHO)** | **0.92** | **0.80** | **0.73** | **0.73** |
| Single-model  (ECG) | 0.87 | 0.62 | 0.51 | 0.41 |
| Single-model  (ECHO) | 0.88 | 0.70 | 0.63 | 0.55 |

Table 1 The Results of Models

**Discussion**

In this study, we report the first multimodal (ECG and echocardiogram based) deep learning model in clinical cardiology and use it to predict the etiology of left ventricular hypertrophy. Combining complementary knowledge from multiple modalities can improve diagnostic performance in clinical practice. The trained model demonstrates high discriminatory ability in distinguishing HCM. Furthermore, ablation studies provided independent support from unsupervised analysis for clinicians’ focus on ECG lateral repolarization and echocardiographic proximal septal hypertrophy for the diagnosis of HCM. Combining complementary information from multiple modalities is intuitively appealing for improving the performance of learning-based approaches.

**Conclusion**

In summary, this study demonstrates a deep learning model incorporating ECG and echocardiogram time series data and apply it to help identify HCM patients. We use well known fusion methods of combining data streams from multiple modalities and compare these comprehensively to single-modal models. Future research should focus on optimizing the model further, expanding the scope of multimodal data integration, and exploring its application to other diseases.

**Reference**

[1] Attia ZI , Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, Pellikka PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Kapa S Age and sex estimation using artificial intelligence from standard 12-Lead ECGs. Circ Arrhythm Electrophysiol2019;12:e007284.

[2] Ko W-Y , Siontis KC, Attia ZI, Carter RE, Kapa S, Ommen SR, Demuth SJ, Ackerman MJ, Gersh BJ, Arruda-Olson AM, Geske JB, Asirvatham SJ, Lopez-Jimenez F, Nishimura RA, Friedman PA, Noseworthy PA Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. J Am Coll Cardiol2020;75:722–733.

[3] Zhang J , Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan C, Fleischmann KE, Melisko M, Qasim A, Shah SJ, Bajcsy R, Deo RC Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation2018;138:1623–1635.

[4] Tran D , Wang H, Torresani L, Ray J, LeCun Y, Paluri M A closer look at spatiotemporal convolutions for action recognition. Proc IEEE Conf Comput Vision Pattern Recognition2018:6450–6459.

[5] Goto S , Mahara K, Beussink-Nelson L, Ikura H, Katsumata Y, Endo J, Gaggin HK, Shah SJ, Itabashi Y, MacRae CA, Deo RC Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. Nat Commun2021;12:2726.

[6] Yen-hsiu Chou, Shenda Hong, Yuxi Zhou, Junyuan Shang, Moxian Song, and Hongyan Li. Knowledge-shot learning: An interpretable deep model for classifying imbalanced electrocardiography data. Neurocomputing, 417:64–73, 2020.

[7] Xiang Lan, Dianwen Ng, Shenda Hong, and Mengling Feng. Intra-inter subject self-supervised learning for multivariate cardiac signals. In Proceedings of the AAAI Con-ference on Artificial Intelligence, volume 36, pages 4532–4540, 2022.

[8] Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. Computers in Biology and Medicine, 141:105114, 2022.

[9] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learn-ing—a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence, 41(9):2251–2265, 2018.

[10] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, and Xi-Zhao Wang. A review of generalized zero-shot learning methods. arXiv preprint arXiv:2011.08641, 2020.

[11] Gokul S Krishnan and S Sowmya Kamath. A supervised learning approach for icu mortality prediction based on unstructured electrocardiogram text reports. In International Conference on Applications of Natural Language to Information Systems, pages 126–134. Springer, 2018.