**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]

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.

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.