**Introduction**

Multimodal data analysis has gained increasing importance in medical research, particularly for improving the precision and reliability of disease diagnosis. With the rapid expansion of medical data and advancements in artificial intelligence, combining multiple data types, such as electrocardiograms (ECG) and echocardiograms (ECHO), has become a promising approach for more accurate cardiac disease diagnosis [6][7]. Traditional single-modality diagnostic methods often fail to capture the complexity of certain medical conditions, prompting a shift towards these multimodal approaches.

Recent studies in deep learning, specifically in cardiology, have demonstrated impressive results for arrhythmia detection, age estimation, and other clinically actionable insights using single modalities. For example, Ko et al. used convolutional neural networks (CNNs) to interpret ECGs, showing high discriminatory power in classifying hypertrophic cardiomyopathy (HCM) against a background population of left ventricular hypertrophy [2]. However, their study faced limitations due to the presence of concurrent hypertension in approximately 28-30% of HCM cases, which hindered direct comparisons between HCM and hypertension.

Zhang et al. took a different approach by focusing exclusively on echocardiograms with a fully automated method for disease detection [3]. Although this method advanced the field, it did not address the potential benefits of integrating multiple modalities for a more comprehensive diagnostic tool. Additionally, prior work has utilized ECG and ECHO data individually in a stepwise manner for diagnosing conditions like cardiac amyloidosis [5], but the fusion of these modalities in a deep learning context remains underexplored.

In contrast to these previous efforts, our study introduces a novel method that jointly considers both ECG and ECHO data to differentiate between HCM and hypertension. Our LVH-fusion model architecture significantly differs from earlier approaches, as it integrates temporal convolutions rather than relying on 2D CNNs that operate on individual video frames. Empirical evidence suggests that spatiotemporal convolutions are superior for video-based classification tasks, as they better capture temporal information and motion patterns—critical aspects for accurate video analysis [4]. Notably, our method achieves high discriminatory power using only one video view, whereas other methods required two different video views for effective HCM detection.

Despite the progress in multimodal data analysis, studies deeply integrating ECG and ECHO data using large-scale datasets are still scarce [10][11]. This research aims to fill this gap by developing a robust multimodal disease diagnosis system utilizing data from the MIMIC-III and MIMIC-IV databases [8]. Unlike previous research, which often focuses on single-modality analysis, our approach leverages convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with attention mechanisms to enhance the model's generalizability and robustness across multiple cohorts.