2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) | 979-8-3503-4501-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/CHASE60773.2024.00020

A Deep-Learning-Based Multi-modal ECG and PCG Processing Framework for Label Efficient Heart Sound Segmentation

1st Qijia Huang
Department of Electrical and Computer Engineering
Duke University
Durham, USA
qijia.huang@duke.edu

2nd Huanrui Yang
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, USA
huanrui@berkeley.edu

3rd Eric Zeng
eric.zeng@gmail.com

4th Yiran Chen
Department of Electrical and Computer Engineering
Duke University
Durham, USA
yiran.chen@duke.edu

***Abstract*—The COVID-19 pandemic has intensified the need for home-based cardiac health monitoring systems. Despite advancements in electrocardiograph (ECG) and phonocardiogram (PCG) wearable sensors, accurate heart sound segmentation algorithms remain understudied. Existing deep learning models, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), struggle to segment noisy signals using only PCG data. We propose a two-step heart sound segmentation algorithm that analyzes synchronized ECG and PCG signals. The first step involves heartbeat detection using a CNN-LSTM-based model on ECG data, and the second step focuses on beat-wise heart sound segmentation with a 1D U-Net that incorporates multi-modal inputs. Our method leverages temporal correlation between ECG and PCG signals to enhance segmentation performance. To tackle the label-hungry issue in AI-supported biomedical studies, we introduce a segment-wise contrastive learning technique for signal segmentation, overcoming the limitations of traditional contrastive learning methods designed for classification tasks. We evaluated our two-step algorithm using the PhysioNet 2016 dataset and a private dataset from Bayland Scientific, obtaining a 96.43 F1 score on the former. Notably, our segment-wise contrastive learning technique demonstrated effective performance with limited labeled data. When trained on just 1% of labeled PhysioNet data, the model pre-trained on the full unlabeled dataset only dropped 2.88 in the F1 score, outperforming the SimCLR method. Overall, our proposed algorithm and learning technique present promise for improving heart sound segmentation and reducing the need for labeled data.**

***Index Terms*—heart sound (PCG) segmentation, self-supervised learning, multi-modal signal processing**

I. Introduction

Cardiovascular diseases (CVDs) are a leading cause of death worldwide. Heart diseases pose a significant threat as they often go unnoticed until they reach a severe and potentially fatal stage. With hospitals overwhelmed by COVID-19 patients, the demand for at-home self-monitoring systems has grown. These systems can detect potential CVDs by monitoring patients' physiological signals, alerting them to abnormal data, and transmitting information to doctors for diagnosis and treatment planning.

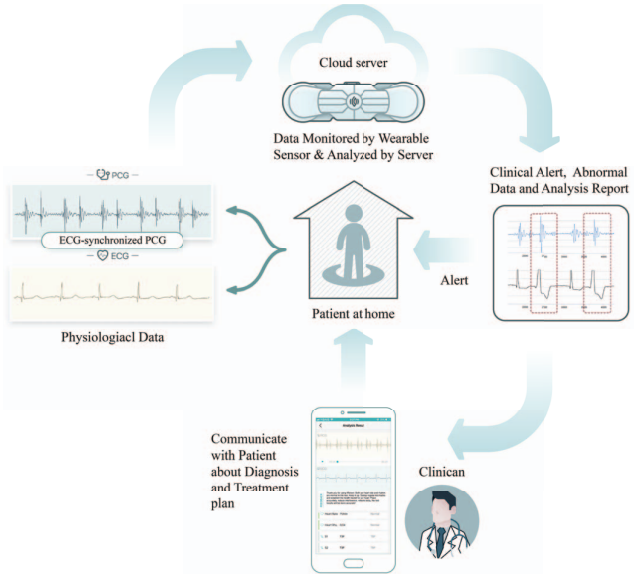

Fig. 1. Illustration of our self-monitoring system workflow. First, the user's physiological data is collected by the wearable device and subsequently uploaded to the server. The server-side algorithm then performs data analysis on the collected information. If any abnormalities are detected, a clinical alert will be sent to the user, and the filtered data, along with an analysis report, will be forwarded to the clinician. The clinical can diagnose based on the report and data, and develop a further treatment plan.

We have developed a self-health monitoring system, as

This work was supported in part by the Directorate for Computer and Information Science and Engineering (CISE) under award number 1822085; and in part by Bayland Scientific.

2832-2975/24/$31.00 ©2024 IEEE
DOI 10.1109/CHASE60773.2024.00020

Fig. 2. Illustration of the Bayland Scientific wearable device used for data collection.

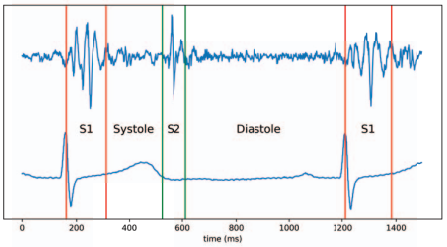

Fig. 3. Illustration of the four states of the heart cycle, along with the corresponding ECG and PCG signals.

shown in Fig. 1. Patients can attach wearable electrocardiograph (ECG) and phonocardiograph (PCG) sensors to their chest, collect ECG and PCG signals at home, and upload them to a server for analysis. Any detected abnormalities are then sent to patients and clinicians for further diagnosis. The collection of ECG and PCG signals is enabled by a novel technology developed by WENXIN and Bayland Scientific Technology: a band-aid-like wearable ECG and PCG device, illustrated in Fig. 2. The device has received Chinese National Medical Products Administration (NMPA) approval and has been used in a heart failure study [1] for data collection purposes. Patients can attach the device to their chest and easily perform ECG and PCG tests at home. The sensed data can then be recorded and transmitted to the server in real time. Being a wearable device, the chest sticker enables continuous signal monitoring without interfering with daily activities. Furthermore, it significantly diminishes the impact of environmental noise on the PCG signal, unlike other types of wearable devices, such as wristbands and life vests, which have a looser fit to the body. This study concentrates on advancing the system by developing the algorithm for analyzing the ECG and PCG data gathered by the wearable sensors.

While both ECG and PCG are crucial for diagnosing and treating CVD, PCG is particularly useful in detecting abnormalities in heart valve function. Recent publications have explored the direct use of machine-learning-aided techniques to diagnose CVDs from PCG signals [2], [3]. These expert algorithms focusing on particular cardiac tasks normally could achieve high accuracy on their specific tasks and datasets. However, the usage of expert diagnosing algorithms in real health-monitoring scenarios might be limited due to the variety of diseases and the need to provide enough interpretability to clinicians for further diagnosis. Therefore, we focus on the heart sound segmentation task, which allows us to accurately locate key components of heart sounds. These locations can be used to detect the presence of extra sound components or measure the left ventricular ejection time (LVET), which is often associated with heart failure [4], [5]. The segmentation results provide users and clinicians with interpretable measures of heart activities for diagnosis and reference.

During a cardiac cycle, the heart generates two fundamental sounds that are related to different phases of the cycle. The first sound, S1, is produced by the closure of the atrioventricular valves when the heart's ventricles begin to contract. The second sound, S2, is produced by the closure of the aortic and pulmonary valves and their vibrations. Occasionally, additional sounds may be heard. The systole interval spans from S1 to S2, while the diastole interval covers the period from S2 to the beginning of the subsequent S1. Fig. 3 visually illustrates these states and intervals. Our task involves accurately segmenting these four states from the PCG signal, which poses challenges due to the sequential nature of the data. Some recent works on heart sound segmentation used Convolutional Neural Network (CNN) [6] and Recurrent Neural Network (RNN) [7], [8]. Their methods may struggle to sequentially segment PCG signals with interference caused by significant noise, murmurs, or extra heart sound components. Since their algorithms only take PCG signals as input, this problem might be inevitable. With synchronized ECG and PCG signals, we can utilize the periodic nature of heart sounds to aid segmentation. Knowing that the first and second heart sounds appear only once in one heartbeat cycle, the segmentation task for a single heartbeat becomes much easier. Motivated by the recent success of two-step detection algorithms, such as Mask RCNN in semantic segmentation tasks [9], we developed a two-step heart sound segmentation algorithm: the first step involves finding a bounding box for each heartbeat based on ECG R-peak detection, and the second step segments the heart sounds within the bounding box. This two-step algorithm divides the challenging heart sound segmentation task into two sub-tasks. For the first step, building upon previous research [10] for R-peak prediction, we developed a variant model with the same stacked 1D-CNN and LSTM structure for R-peak prediction. Since the number of R-peaks is significantly lower than non-R points, we developed a regression objective for algorithm optimization to replace the commonly used classification objective for segmentation. For the second task of segmentation within a heartbeat cycle, we developed a 1-D CNN-based U-Net variant model to perform fine-grained segmentation on the PCG signal with a length of one heartbeat. ECG information is also taken as multi-modal input to improve model performance.

Deep learning algorithms for biomedical tasks often suffer from insufficient labeled data due to a lack of experts. Prior deep learning methods, such as LSTM [7], have not achieved the same level of accuracy on partially synchronized ECG and PCG datasets from Physionet 2016 as on full databases. Furthermore, using multi-modal input requires more labeled

data to prevent overfitting caused by the curse of dimensionality. To reduce reliance on a large amount of data, researchers developed semi-supervised learning methods that pre-train the input encoder on a pretext task with data whose labels are pseudo-labels created by itself. Among all pretext tasks, the contrastive learning method, such as SimCLR [11], achieved high accuracy on multiple downstream image classification tasks using limited labeled data. However, transferring the contrastive learning method to our fine-grained signal segmentation task resulted in sub-optimal outcomes. Thus, we developed a contrastive learning method for downstream signal segmentation that contrasts between embeddings of each signal segment. The trained encoder and decoder are then fine-tuned on the labeled dataset.

The main contributions of this work can be summarized as follows:

- We propose a novel two-step heart sound segmentation algorithm that utilizes the temporal correlation between two modalities of heart activity, and evaluate its advantages on both public and private PCG and ECG datasets.
- We propose a novel semi-supervised learning method for the signal segmentation task that overcomes the shortcomings of using contrastive learning methods designed for classification.
- We evaluate the semi-supervised learning algorithm with respect to the required labeled data size and for transfer learning, and demonstrate its effectiveness in reducing the need for labels.

II. Related Work

A. Heart Sound Segmentation

Early heart sound segmentation approaches relied on traditional signal processing techniques, such as envelope-based methods or wavelet transforms, for feature extraction, followed by threshold-based peak-finding algorithms to identify the boundaries of S1s and S2s [12]–[15]. However, these traditional methods rely on threshold-based peak-finding algorithms and cannot be generalized to signals from different sources. Additionally, they are not robust against the significant noise typically associated with PCG signals.

As machine learning techniques have advanced, researchers have explored combining signal processing techniques for feature extraction with machine learning algorithms for classification [16], [17]. Deep learning frameworks have also been employed for heart sound segmentation, with a significant focus on using temporal models such as Hidden Markov Models (HMM) [18]–[20] and Deep Recurrent Neural Networks (DRNN) [7], [8]. The Logistic Regression Hidden Semi-Markov Model (LR-HSMM) [18] was considered highly accurate and used in the 2016 PhysioNet Challenge for generating heart sound segmentation labels. In another study, RNN [7] demonstrated better performance than CNN in analyzing the sequential states of PCG signals. However, these temporal methods lack the ability to process raw signals, so feature extraction algorithms must be applied first. Typically, frequency-domain features are extracted and have proven effective, such as wavelet transform in [18] and Mel-frequency spectral coefficients (MFCC) in [7], [8]. As an alternative or complement to signal processing techniques, CNNs have also been employed to extract features from raw or processed signals [6], [21], and CNN modules can be effectively combined with temporal models for heart sound segmentation tasks [22]–[24].

The joint processing of ECG and PCG signals has been applied in heart sound classification and the detection of heart diseases, where deep learning techniques are employed to extract and fuse features from both modalities [25]–[27]. Despite this, the synchronized analysis of ECG and PCG signals remains relatively unexplored in the context of heart sound segmentation. Some studies have attempted to enhance heart sound segmentation by incorporating information from the ECG signal [28], [29]. Their approaches, based on HMMs, leverage events (e.g. R-peaks and T waves) detected in ECG signals to inform more accurate segmentation predictions on PCG signals, based on their temporal relationship. Utilizing ECG events is very inspiring; however, this integration occurs only at the decision-making level. The actual detection within ECG and PCG signals relies on conventional feature extraction methods, without a true coupling of the ECG and PCG data streams.

In our proposed two-step approach, we first employ our R-peak detection algorithm to separate data into single heartbeats, and then apply a 1D variant of the U-Net model [30] for fine-grain segmentation. In this case, the heartbeat signal feeding the U-Net is treated as a static object rather than having temporal dependency, allowing us to leverage the strengths of CNN-based models in recognizing spatial patterns on raw signals and achieve higher accuracy.

B. Semi-Supervised Learning

Semi-supervised learning (SSL) effectively utilizes large unlabeled datasets to learn data representations for supervised downstream tasks, reducing the reliance on labeled data. It achieves this by designing a self-supervised learning method that transforms unsupervised learning problems into supervised ones through ”pretext tasks”. One powerful pretext task is contrastive learning, which conforms similar (positive) and contrasts dissimilar (negative) pairs of examples. Several contrastive learning methods, such as SimCLR [11] and MoCo [31], have established benchmarks in computer vision (CV), particularly in image classification tasks. Contrastive learning has also been employed to learn representations for medical images [32]. However, transferring the data representations learned from contrastive learning to downstream segmentation tasks, which involve pixel-level predictions, is challenging. Building on the intuition of extracting local features and contrasting local regions or pixels, region-level contrastive learning has shown promising results in image segmentation tasks [33]–[36]. The methods for generating pseudo-labels in these approaches can be classified into two categories: label-based [33], [35], [36] and indices-based [34].

There have been numerous efforts to apply SSL methods to enhance the performance of biosignal processing. Some of these efforts involve defining new pretext tasks, such as SSL-ECG [37], which learns ECG representations by separating augmented signals based on their augmentation types. However, typical image data augmentation methods, like clipping or rotation used by SimCLR [11], are not suitable for time-series data, making it difficult to apply contrastive learning to biosignal data. In [38], the authors utilized domain-specific transformations to augment EEG signals; in [39], they used additional temporal, source of collected position, and source patient labels to generate pseudo-labels; and in [40], they applied spectrogram augmentation, widely used in audio-related tasks [41], to create augmentations for heart and lung sound signals. Despite these attempts, there is still a lack of research on applying contrastive learning to biosignal segmentation, particularly for PCG signal or multi-modal signal segmentation.

In this work, we propose a novel contrastive learning method to obtain general representations for synchronized ECG and PCG signals. The learned representations contain distinctive local representations that are beneficial for downstream segmentation tasks.

III. Proposed Methods

In this section, we present the details of our two-step heart sound segmentation algorithm, which includes the first step of heartbeat detection, the second step of heartbeat-level heart sound segmentation, and the representation learning method for the downstream segmentation task. The overall algorithm is described in Algorithm 1.

A. ECG R-peak Detection

Accurately identifying the R-peaks in an ECG signal is the initial step toward localizing the complete heartbeat, thereby enabling the subsequent identification of detailed cardiac activities. Deep learning-based R-peak detection algorithms have been well-developed in previous research. CNN-based methods extract local features from waveforms, exhibiting noise robustness, while RNN-based methods effectively utilize temporal information for sequential detection. We employ a CNN-LSTM structure that combines the strengths of CNN and LSTM architectures, which has been shown to perform effectively under noisy conditions [10]. Given that the original model structure is designed for a sampling rate lower than that of our dataset and to reduce complexity while maintaining accuracy, we have slightly modified the model structure. This modified model accepts the ECG signal as input and comprises two 1D-CNN layers, each with a kernel size of 101 and 8 channels, followed by an average pooling layer. These convolutional layers are followed by an LSTM layer with a dimension of 8. The output of the LSTM layer is then passed to a fully connected layer to produce the final output. The CNN employs the ReLU activation function, and the LSTM layer employs the Tanh activation function.

Algorithm 1 Training and inference of two-step heart sound segmentation algorithm

Input: training set $(\mathbf{X}, \mathbf{y})$, ECG R-peak detection algorithm $f_R(\cdot)$, our U-Net model $f_{dec}(f_{enc}(\cdot))$, and a projection head $f_{proj}(\cdot)$

Output: The final U-Net model and the segmentation results

$E \leftarrow X_{ecg}, P \leftarrow X_{pcg}$
Optimize f_R using the $l_{MSE}(f_R(E), y_{r-peaks})$
R-peak positions $R_{pos} \leftarrow f_R(E)$
$\{x_i, y_i\} \leftarrow (\mathbf{X}, \mathbf{y})$ based on R_{pos} {heartbeats detection}
for $x_i \in \{x_i\}$ **do**
    $e_i \leftarrow x_i$ by masking the pcg channel
    $p_i \leftarrow x_i$ by masking the ecg channel
    $z_{i,e} \leftarrow f_{proj}(f_{enc}(e_i))$
    $z_{i,p} \leftarrow f_{proj}(f_{enc}(p_i))$
    Optimize $f_{proj}(f_{enc}(\cdot)))$ by gradient descent using lose function $l_{segcon}(z_{i,e}, z_{i,p})$ {pre-training}
end for
for $x_i, y_i \in \{x_i, y_i\}$ **do**
    $\hat{y}_i \leftarrow f_{dec}(f_{enc}(x_i))$
    Optimize f_{enc} and f_{enc} together by gradient descent using lose function $l_{CE}(\hat{y}_i, y_i)$ {finetuning}
end for
$\hat{Y} \leftarrow f_{dec}(f_{enc}(X))$ {inference}

Due to the relatively small number of R-peaks compared to other points, the classes are imbalanced. For a T ms heartbeat, the proportion of R-peaks to the entire heartbeat is only $\frac{1}{T}$. To address this issue, a weighted classification loss function is necessary for detecting R-peaks through a classification task that works through each point in the sequence [10]. However, determining the appropriate weight would be challenging since the length of heartbeats varies among individuals. To overcome this challenge, we convert the task from classification to regression by converting the label to a Gaussian-shaped target. Specifically, if we consider x on the time axis centered at the R-peak position (i.e., $x = 0$ at the R-peak), then the label can be expressed as follows:

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}}. \tag{1}$$

Set it in a standard form by letting $a = 1$, $b = 0$ and $c = 1$, then:

$$f(x) = e^{-\frac{x^2}{2}}. \tag{2}$$

In our private dataset, we conduct an ablation study to compare the continuous Gaussian-shaped labeling and categorical labeling approaches. In this experiment, the categorical label for R-peaks is set to 1, while the rest of the sequence is labeled as 0. We utilize a weighted Cross-Entropy loss function with a weight ratio of $1 : 550$ for class 0 versus class 1. For R-peak detection using regression with Gaussian-shaped labeling, we apply the Mean Squared Error (MSE) as the loss function. The model's predictions are depicted in Fig. 4. Within a tolerance of 50 ms from the true R-peak locations in this dataset,

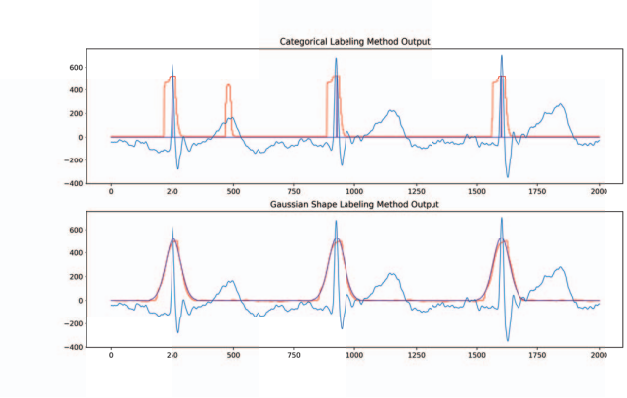

Fig. 4. The illustration depicts typical CNN-LSTM model predictions for R-peak, where the red line represents the prediction and the purple line represents the true label. The upper figure displays the output using a categorical labeling strategy, while the lower figure displays the output with Gaussian-shaped labeling.

the categorical labeling strategy yields 99.96% precision and 99.11% recall, whereas the Gaussian-shaped labeling strategy achieves 99.89% precision and 99.84% recall. The lower recall observed with the categorical labeling strategy suggests a higher probability of the model misclassifying normal ECG points as R-peaks. This issue could be due to the sub-optimal class weight setting for Cross-Entropy Loss, whereas the Gaussian-shaped labeling does not exhibit such a problem. Furthermore, when evaluated on the MIT-BIH dataset [42], the algorithm achieves an F1 score of 99.68% which demonstrates its adequate accuracy in R-peak detection.

B. Heartbeat Level Segmentation Task

Since we have already developed an algorithm that can identify heart sound cycles, our next task is to segment the individual heart states within a specific heartbeat extracted by our R-peak identification algorithm.

1) Data Preprocessing: Let s represent a normalized 2-channel signal from ECG and PCG, containing N heartbeats as segmented by the R-peak identification algorithm. Each heartbeat b is defined as starting from 100ms before one R-peak and ending at 100ms before the subsequent R-peak, so $s = [b_1, ..., b_n]$. In this configuration, the heartbeat period encompasses all four states of the cardiac cycle.

Since CNN processing requires inputs to have the same shape, we resize the heartbeat sequences to have equal lengths. For heartbeats with a length less than 1536, we pad them with zeros after the original sequence to reach 1536; for those longer ones, we only retain the first 1536 samples. After resizing the signals, we obtain a set of equal-length heartbeats from the original sequence to use in our training set: $X = [b'_1, ..., b'_n]$ and $Y = [y_1, ..., y_n]$, where each y_i is a 1D array filled with class index (0,1,2,3) corresponding to systole, S1 period, diastole, and S2 period respectively.

2) U-Net Based CNN model: Inspired by the widespread use of 2D U-Net in biomedical image segmentation tasks, we designed a 1D variant of U-Net for our heart sound segmentation framework. Similar to the original U-Net, we retain the Encoder-Decoder structure, the two convolutional layers per block architecture with ReLU activation function, and the skip connections. However, we modify all the convolution layers, max-pooling layers, and up-convolution layers to be one-dimensional. Additionally, we adjust the number of filters in each convolution layer to better extract spatial features from 1D signals. The kernel size is set to 7 for smooth feature extraction on signals, with padding applied to maintain the sizes of input and segmented output. We use average pooling layers instead of max-pooling layers, and the step of the pooling and up-convolution layers is set to 4. The structure of the model is shown in Fig. 5.

3) Loss Function: For this segmentation task, we try to minimize the categorical difference on each pixel between our model output and the true segmentation. In the experiment, we use Cross-Entropy Loss function for our optimization problem. The optimization object for a segmentation object in the batch is shown as:

$$argmin_\theta \frac{1}{L} \sum_{i=1}^{L} \sum_{j \in (0,1,2,3)} p_j(y_i) \log f_j(b'_i; \theta), \qquad (3)$$

where (b'_i, y_i) is a pair of input point on heartbeat and label at index i; L is the heartbeat's length which in our case is 1536; $p_j(y_i) = 1$ when j is the same as the class in y_i, 0 otherwise; and the function $f_j(b'_i; \theta)$ denotes the output probability on class j from the U-Net model. Thus our optimization aims to minimize the pixel-wise difference between the model prediction and the true labeling of original signals.

C. Self-supervised Training

We propose a self-supervised learning strategy that encourages the encoder of a U-Net to extract distinctive local representations suitable for segmentation by the decoder. The first step is to perform a temporal invariant data augmentation on all the heartbeat signals, which will not change the location of each sampling point on the time-axis. For each input heartbeat, b', we generate two augmentations, $Aug_1(b')$ and $Aug_2(b')$. In our case, since we use synchronized ECG and PCG signals, we mask one input channel respectively as two augmentations, which will not perturb the temporal information of the original signals. Then we pass the signals to the encoder of our U-Net, Enc, which maps $Aug_1(b')$ and $Aug_2(b')$ to representation vectors, $r_1 = Enc(Aug_1(b'))$ and $r_2 = Enc(Aug_2(b'))$. Then the representations are passed through a projection network, $Proj$, to obtain the feature maps $z_1 = Proj(r_1)$ and $z_2 = Proj(r_2)$, each of dimensions $L' \times C$, with $L' < L$. We instantiate $Proj$ as a 1D Convolutional layer with a kernel size of 1 and an output dimension of 128, so the z has the same dimension as the output of U-Net encoder. Then the feature maps z with length L' can be considered as having L' pixels, where each pixel corresponds to a segment from the input signal with overlapping. Based on the understanding that two segments of signals from different views of cardiac activities

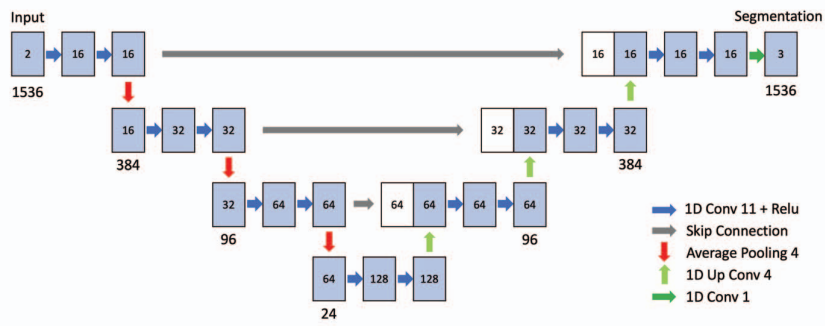

Fig. 5. Illustration of Proposed U-Net Structure. In this example, the input length is set to 1536 with 2 channels for ECG and PCG. The numbers inside the boxes represent the number of signal channels, while the numbers below the boxes indicate the signal length.

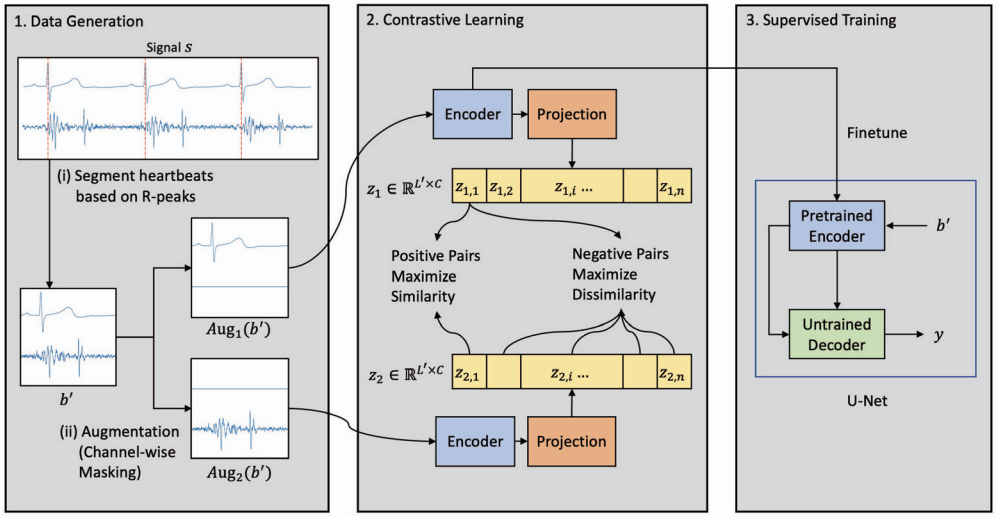

Fig. 6. Illustration of contrastive learning Framework. 1. Prepare data by applying R-peak detection on signal S, segmenting heartbeats, and reshaping them to the same dimensions to obtain b'. Then, apply two temporal invariant augmentations. 2. Generate feature maps for two augmented heartbeat signals for contrastive learning. Each feature map is divided into n segments. Pairs of segments from the two feature maps are labeled as positive or negative pairs based on their indices, which are used for optimizing the encoder and projection layer using segment-wise contrastive loss. 3. Remove the projection layer from the encoder and perform supervised training jointly with U-Net's decoder.

should reflect similar cardiac activity if they are collected at the same time, contrasting the representations of segments with different indices from the feature maps of two views will make the representation have distinctive features for heart activity detection. Then our segment-based contrastive loss for a given input signal can be defined as:

$$l_{segcon} = -\frac{1}{L'}\sum_{i\in L'}\log\frac{\exp\left(sim(z_{1,i}, z_{2,i})/\tau\right)}{\sum_{j\in L'}\exp\left(sim(z_{1,i}, z_{2,j})/\tau\right)}, \quad (4)$$

where the $sim(\cdot,\cdot)$ computes the cosine similarity between two vectors $sim(a,b) = \frac{a^T b}{\|a\|\|b\|}$. In this optimization problem, segments from different views that share the same indices are considered as positive pairs, while those with different indices are considered as negative pairs. The framework of proposed contrastive learning is shown in Fig. 6.

IV. Experiment

A. Heart Sound Segmentation Performance Evaluation

1) Datasets: We use both the 2016 Physionet Challenge database and the private Bayland Scientific dataset to evaluate our method.

The Physionet 2016 database [43] is the most widely used database for heart sound research. Although the objective of this challenge is the heart sound normal/abnormal classification, this database is also the primary benchmark for research on the heart sound segmentation task. This database includes 3,126 heart sound recordings. Each recording lasts from 5 seconds to 120 seconds with a sampling frequency of 2000 Hz. Since signals in this database are collected from different locations of the body, from both adults and children, under clinical and non-clinical settings, with or without diseases, the scales and the patterns vary among different signals. Also due to the uncontrolled environment, significant noise including

TABLE I
RESULT ON PHYSIONET 2016 DATASET

Method	PPV	Se	Spe	Acc	F1
U-Net [6]	93.2	92.3	98.2	95.0	92.7
BiLSTM Attention [7]	94.2	95.0	95.1	93.5	94.75
GRNN [8]	92.41	93.16	97.92	94.47	92.77
U-Time [44]	83.58	82.93	95.66	93.75	83.24
HRNet [45]	94.05	92.37	98.26	97.4	93.16
Proposed U-Net	**96.3**	**96.55**	**99.12**	**97.53**	**96.42**

TABLE II
PRE-TRAINING RESULT ON PHYSIONET 2016 DATASET

Method	PPV	Se	Spe	Acc	F1
Train from scratch	**96.3**	96.55	**99.12**	97.53	96.42
SimCLR [11]	96.12	96.51	98.69	97.24	96.40
Proposed U-Net+Pre-train	96.13	**96.76**	99.11	**97.55**	**96.43**

talking, breathing and etc. will be captured by the sensors. For the classification task, this dataset contains annotations of normal/abnormal. For the segmentation task, this challenge provides annotations for fundamental heart sound, S1, systole, S2, and diastole on signals. These annotations are generated by the LR-HSMM algorithm [18] and manually decide their correctness. The challenge provides five training sets, but only the 'training-a' subset contains 2-channel ECG and PCG signals. Therefore, we only use data from 'training-a' with accurate labels. In total, 288 recordings are used for evaluation, with 186 recordings designated for training, 43 recordings for validation, and 59 recordings for testing purposes.

The second dataset we use is collected by Bayland Scientific using the wearable device stuck on the chest. The private dataset is collected from 2,072 adult patients in a clinical setting. This dataset contains 2,076 dual-channel synchronized ECG and PCG signals with an average length of 50 seconds. The recordings have a sampling rate of 1,000 Hz. This ECG signal is a single lead signal of Lead II, which is the same one used in the MIT-BIH dataset. Professionals from Bayland Scientific have fully labeled all the R-peak positions. For the PCG signal, professionals identify and label the positions of S1 start, S1 end, S2 start, and S2 end. We split the dataset into training, validation, and testing sets with sizes of 1272, 300, and 500, respectively.

2) Evaluation Metrics: We assess the performance of our heart sound segmentation algorithm using five metrics: positive predictive value (PPV), sensitivity (Se), specificity (Spe), F1 score, and accuracy (Acc). Heart sound segmentation performance is evaluated for each of the four states of the heart cycle, and to evaluate the overall performance of the segmentation algorithm, we compute the final metrics by globally averaging across the four classes. Accuracy is calculated globally as the ratio of correctly classified states to the total number of pixels.

3) Implementation: The initial learning rate is set to 0.001 for contrastive learning, training the U-Net segmentation model from scratch, and finetuning the U-Net segmentation model with a pre-trained encoder. We use the Adam optimizer to train the proposed model. Training is terminated if the validation loss does not decrease within 20 epochs, and we set the maximum training epoch at 200 for both encoder training and U-Net training.

4) PhysioNet Results: Table I presents the evaluation results of our proposed segmentation algorithm on the PhysioNet dataset, compared to other state-of-the-art algorithms. All values in the table are expressed as percentages. For our proposed method, we included and applied two-step segmentation and multi-modal inputs as previously discussed. The baseline models we compared include a U-Net model with the same structure as in [6] with a moving window for segmentation, a GRNN model [8], and a Bi-LSTM with attention mechanism [7]. Due to the unavailability of public implementations and differences in evaluation metrics, we implemented the algorithms based on the descriptions in the original papers and reported their performance using our metrics. Another two baselines are widely used benchmarks for segmentation from different tasks. The first is U-time [44] for sleep stage segmentation, we made adaptations by setting the input length as 2560 ms and segmenting resolution as 20 ms, removing dilation for the encoder, setting all kernel sizes for pooling and up-sampling as 4. For the HRNet [45], we adopted the 4-stage HRNet-18. The adaptations include converting the model to 1D CNN, setting the input length as 2560 ms, and setting the kernel sizes by branch as 16, 32, 64, and 128.

Observations indicate that RNNs achieve overall better results than the regular 1D U-Net, suggesting that temporal models and their frequency-domain feature extraction methods are effective in processing cardiac sequential data. Our two-step multi-modal techniques make the U-Net competitive. By using an R-peak detection algorithm to select an appropriate heartbeat-long window size, the CNN-based method can now perform fine-grained segmentation. Since recurrent networks can only classify a selected small window, this approximation is likely to result in less accurate predictions at the boundaries of fundamental heart sounds. Since the convolutional structures from U-time and HRNet, as well as the RNN structures from BiLSTM and GRNN, are not less complex than those from the two-step detection structure, the improvements in accuracy and F1 scores are likely due to the adoption of multi-modal inputs and the effective leverage of combined information.

We also implemented SimCLR [11] as a benchmark self-supervised training method to compare with our segment-wise contrastive learning. The same U-Net segmentation architecture is used for both the benchmark and proposed methods. Since augmentations such as clipping or rotation from the original paper are difficult to apply to signal data, we used two types of augmentation: the frequency masking method from SpecAugment [46] and masking the ECG channel. The results from Table II show that SimCLR does not perform well on the downstream segmentation task, and the pre-trained weights can even compromise the effect of finetuning. This demonstrates the necessity of an alternative pretext task. Our proposed segment-wise contrastive training either matches or

slightly exceeds the performance of the model trained from scratch, indicating that the proposed pretext training task does not adversely affect the downstream fine-grained task. The benefits of label-efficient training will be analyzed in subsequent experiments.

B. Pre-training Effect on Reduced Training Data Size

Besides improving the performance of the U-Net model, a more important motivation to utilize contrastive learning pre-training is to reduce the need for labeled training data. We designed an experiment to assess its efficacy by using the full training dataset without labels as the self-supervised set and partially revealing labels for the labeled set. We compared our segment-wise contrastive learning approach with the baseline and SimCLR [11], using the same U-Net architecture. The baseline involved training U-Net from scratch on the labeled set. We trained two U-Net encoders using SimCLR and our method on the self-supervised set. The SimCLR-trained encoder was finetuned with the decoder on the labeled set. For the encoder trained with our method, we conducted experiments with gradients frozen and finetuned when jointly trained with the decoder on the labeled set. In both Bayland and Physionet datasets, we progressively reduced the proportion of revealed labels to investigate the relationship between labeled data size and self-supervised training methods. For the Bayland dataset, we conducted experiments with label proportions ranging from 100% to 30%, 10%, 5%, and 1%, corresponding to 9000, 2700, 900, 450, and 90 labeled heartbeats. For the Physionet dataset, we performed experiments with label proportions ranging from 100% to 30%, 10%, 5%, 1%, and 0.1%, corresponding to 7807, 2342, 781, 390, 78, and 8 labeled heartbeats. The number of training epochs will be increased corresponding to the decrease in data size of supervised learning. We used the F1 score to evaluate the results, as shown in Fig. 7.

As observed, with a 100% labeled dataset, finetuning the encoder pre-trained with the proposed segment-wise contrastive learning improves accuracy on the Bayland dataset. On the Physionet dataset, the scratch, SimCLR, and proposed methods show no significant difference, but the F1 scores are consistently lower for the proposed method with frozen gradient finetuning on both datasets. The advantage of using a pre-trained encoder with finetuning over training U-Net from scratch becomes more evident as the training set size reduces. As the labeled portion of the Physionet dataset decreases from 100% to 30%, 10%, 5%, 1%, and 0.1%, the differences in F1 scores between the proposed method with finetuning and training from scratch increase from 0.01 to 0.18, 0.49, 0.56, 1.95, and 10.60. This indicates that the pre-trained encoder has better generalization ability than random weight initialization. However, the improvement is not substantial with SimCLR when the labeled data size is small. Using only 5%, 1%, and 0.1% of labeled data, our proposed method results in F1 score drops of only 0.94, 2.88, and 11.59 compared to training with the complete labeled data.

TABLE III
COMPARISON OF TRANSFER LEARNING PERFORMANCE

Method	B2P10%	B2P1%	P2B10%	P2B1%
Frozen Encoder				
Our Semi-supervised	**94.94**	**93.46**	88.31	86.08
Supervised	**94.94**	92.24	**88.89**	**86.7**
Finetuned				
Our Semi-supervised	95.53	**93.43**	89.65	87.09
Supervised	**95.63**	93.37	**90.11**	**87.27**
Scratch	95.33	92.08	89.65	86.14

Interestingly, the model with a gradient-frozen pre-trained encoder achieves higher F1 scores when the data size shrinks, even outperforming the finetuned encoder in extreme cases with only 8 heartbeats. This suggests that our pre-trained encoder generates better-informed and more-generalized representations of the input. Sample outcomes are shown in Fig. 8. The U-Net with a gradient-frozen encoder makes accurate predictions with only 0.1% labeled data, although it has less smooth and accurate heart sound boundaries and may misclassify extra heart sounds not covered in the limited labeled set.

C. Transfer Learning Performance

Transferring knowledge from a more accessible dataset without the desired labels to a private labeled dataset offers a way to leverage the need for labels. The efficacy of transfer learning is also a measure of the quality of the learned representations. We evaluate the performance of representations learned through our segment-wise contrastive learning for transfer learning across the Physionet and Bayland datasets, both with a frozen encoder and in finetuning settings. The dataset configuration involves pre-training the model on one complete but unlabeled dataset and then further training it on 10% and 1% of another labeled dataset, in settings where the encoder gradients are either frozen or finetuned. The baselines include supervised training on a complete and labeled dataset, followed by finetuning with encoder frozen or finetuned, and a model that is randomly initialized and trained from scratch in a finetuned setting.

The results from Table III demonstrate that the proposed semi-supervised learning approach with finetuning attains higher F1 scores compared to supervised training from scratch across all dataset combinations. Without finetuning, neither the semi-supervised nor the supervised method successfully transfers to the new dataset. When finetuned, supervised training on complete labeled data does not reach the same F1 score levels as the model pretrained and finetuned on the same set, as shown in Fig 7; however, it does recover some of the performance gap resulting from the limitation in labeled data size. Our proposed semi-supervised learning method achieves performance close to the supervised method, indicating that its learned representation is not significantly worse than that learned from labeled data.

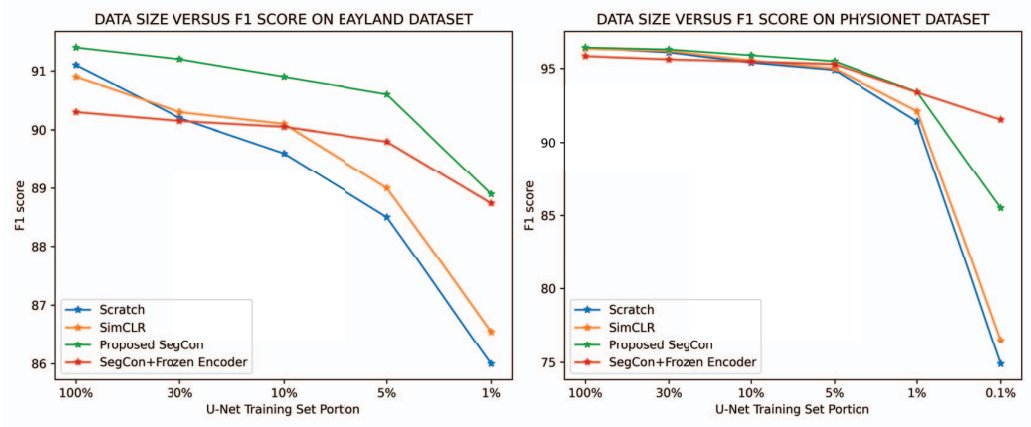

Fig. 7. Illustration of labeled data size influence on the performance of U-Net on Bayland and Physionet dataset. The U-Net is evaluated in 4 cases: i) trained from scratch; ii) encoder pre-trained with SimCLR and finetuned on partially labeled data; iii) encoder pre-trained using proposed segment-wise contrastive learning and finetuned on partially labeled data; iv) encoder pre-trained with segment-wise contrastive learning, weights frozen during finetuning.

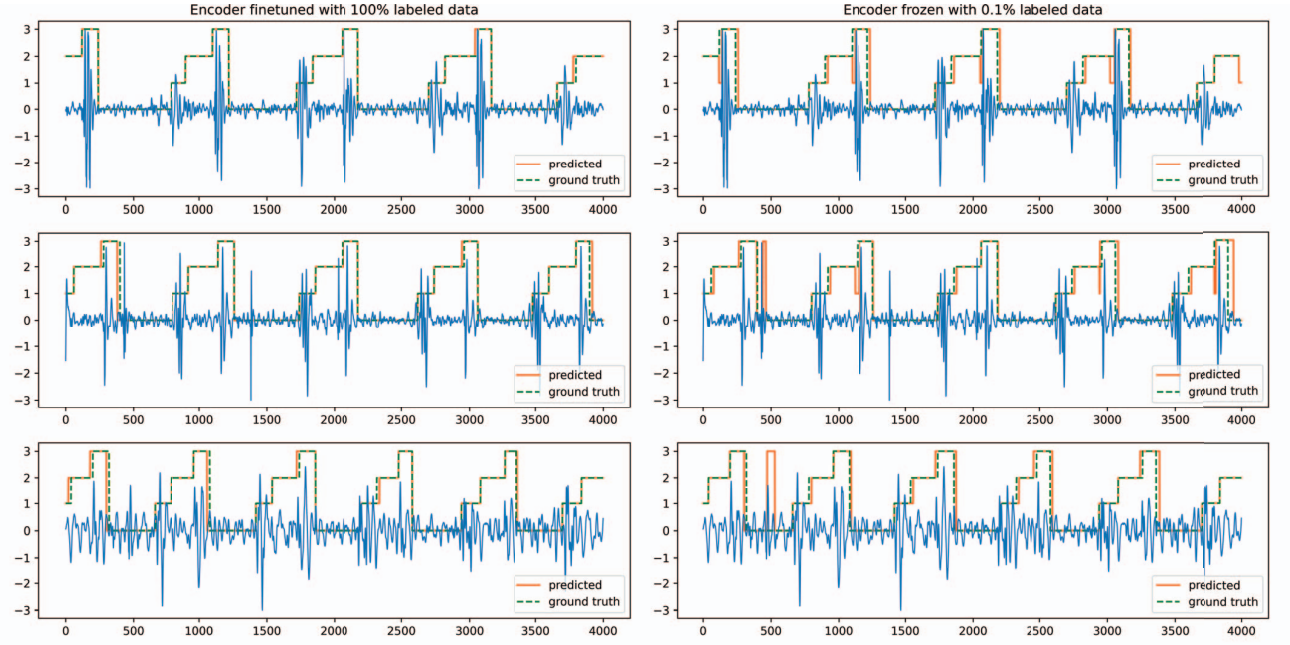

Fig. 8. The segmented results of the proposed two-step heart sound segmentation algorithm on the Physionet training-a subset. Left plots use a U-Net with a pre-trained encoder finetuned on 100% labeled data; right plots use a U-Net with a frozen encoder and a decoder trained on 0.1% labeled data. The corrected annotations and our algorithm predictions are shown in each sub-figure as green and orange staircase plots respectively. The level of the staircase plot corresponds to the heart states of diastole, S1, systole, and S2 in ascending order.

V. Conclusions and Future Work

In this paper, we developed a deep learning framework for automatic heart sound segmentation in self-monitoring systems using multi-modal signals. By leveraging the periodic nature of heart activity, we designed a two-step heart sound segmentation algorithm that first detects heartbeats based on R-peaks from ECG signals and then segments heart sounds within heartbeat durations. The modified R-peak detection achieves accurate results, and we analyze the advantages of regression on Gaussian-shaped labels over classification on categorical labels. Our beat-wise heart sound segmentation method attains state-of-the-art results on the PhysioNet 2016 dataset, and we investigate the benefits of using synchronized multi-modal ECG and PCG signals for segmentation instead of a single channel.

To reduce the reliance on large annotated training sets, we proposed an extension to contrastive loss-based pre-training. Since traditional contrastive learning methods are designed for classification tasks, we developed a method suitable for signal

segmentation tasks. We introduced a contrastive loss for learning local segment representations of signals based on temporal information, useful for dense prediction tasks like segmentation. Evaluating contrastive learning as self-supervised, semi-supervised, and transfer learning demonstrates performance improvement in segmentation tasks and less dependence on costly labeled data. The developed method can benefit few-shot learning, enabling the transfer to different yet related tasks with minimal additional labeling. It can also support domain adaptation, where a model trained on data from one hospital or medical agency can be transferred to another.

However, our proposed method has some limitations. Although our two-step algorithm is more accurate than a CNN model with moving window processing, it has more parameters to train (U-Net has 176K Params and Conv-LSTM has 2K Params), resulting in more FLOPs during the inference stage. However, to achieve higher accuracy with the moving window method, a temporal model like the Viterbi algorithm or HMM may be needed for post-processing. In comparison to RNN-based methods, the U-Net model's parameters are considered extra, as our algorithm takes raw signals as inputs, while RNN-based methods typically employ signal processing methods like MFCC to extract features. Another trade-off is the need for two-step training and inference. Although some methods can stack Conv-LSTM and U-Net into a single model using a soft-argmax layer [47], making the heartbeat detection process differentiable, outputting R-peak prediction results can be beneficial for other medical diagnostic tasks and filtering to improve heartbeat detection accuracy.

Building upon our proposed method, one important future aspect of this work is to improve the efficiency of the utilized deep learning models. As we would like to deploy the proposed method into the mobile devices or smart sensors of end-users, the memory usage and computation cost are typically constrained by the limited resources available on the target hardware. To enable wider and more efficient deployment of deep-learning-based methods, model compression and efficient computation techniques like pruning [48], [49], quantization [50], parallel computation [51], and neural architecture search [52] on vital signal processing tasks are worth investigating.

REFERENCES

[1] X.-C. Li, X.-H. Liu, L.-B. Liu, S.-M. Li, Y.-Q. Wang, and R. H. Mead, "Evaluation of left ventricular systolic function using synchronized analysis of heart sounds and the electrocardiogram," *Heart Rhythm*, vol. 17, no. 5, pp. 876–880, 2020.

[2] F. A. Khan, A. Abid, and M. S. Khan, "Automatic heart sound classification from segmented/unsegmented phonocardiogram signals using time and frequency features," *Physiological measurement*, vol. 41, no. 5, p. 055006, 2020.

[3] Q. Abbas, A. Hussain, and A. R. Baig, "Automatic detection and classification of cardiovascular disorders using phonocardiogram and convolutional vision transformers," *Diagnostics*, vol. 12, no. 12, p. 3109, 2022.

[4] S. P. Collins, C. J. Lindsell, W. F. Peacock, V. D. Hedger, J. Askew, D. C. Eckert, and A. B. Storrow, "The combined utility of an s3 heart sound and b-type natriuretic peptide levels in emergency department patients with dyspnea," *Journal of cardiac failure*, vol. 12, no. 4, pp. 286–292, 2006.

[5] T. Biering-Sørensen, G. Querejeta Roca, S. M. Hegde, A. M. Shah, B. Claggett, T. H. Mosley Jr, K. R. Butler Jr, and S. D. Solomon, "Left ventricular ejection time is an independent predictor of incident heart failure in a community-based cohort," *European journal of heart failure*, vol. 20, no. 7, pp. 1106–1114, 2018.

[6] F. Renna, J. Oliveira, and M. T. Coimbra, "Deep convolutional neural networks for heart sound segmentation," *IEEE journal of biomedical and health informatics*, vol. 23, no. 6, pp. 2435–2445, 2019.

[7] T. Fernando, H. Ghaemmaghami, S. Denman, S. Sridharan, N. Hussain, and C. Fookes, "Heart sound segmentation using bidirectional lstms with attention," *IEEE journal of biomedical and health informatics*, vol. 24, no. 6, pp. 1601–1609, 2019.

[8] E. Messner, M. Zöhrer, and F. Pernkopf, "Heart sound segmentation—an event detection approach using deep recurrent neural networks," *IEEE transactions on biomedical engineering*, vol. 65, no. 9, pp. 1964–1974, 2018.

[9] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[10] B. Yuen, X. Dong, and T. Lu, "Inter-patient cnn-lstm for qrs complex detection in noisy ecg signals," *IEEE Access*, vol. 7, pp. 169 359–169 370, 2019.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[12] L. Huiying, L. Sakari, and H. Iiro, "A heart sound segmentation algorithm using wavelet decomposition and reconstruction," in *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.'Magnificent Milestones and Emerging Opportunities in Medical Engineering'(Cat. No. 97CH36136)*, vol. 4. IEEE, 1997, pp. 1630–1633.

[13] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in *Computers in Cardiology 1997*. IEEE, 1997, pp. 105–108.

[14] A. Moukadem, A. Dieterlen, N. Hueber, and C. Brandt, "A robust heart sounds segmentation module based on s-transform," *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 273–281, 2013.

[15] S. Sun, Z. Jiang, H. Wang, and Y. Fang, "Automatic moment segmentation and peak detection analysis of heart sound pattern via short-time modified hilbert transform," *Computer methods and programs in biomedicine*, vol. 114, no. 3, pp. 219–230, 2014.

[16] I. Maglogiannis, E. Loukis, E. Zafiropoulos, and A. Stasis, "Support vectors machine-based identification of heart valve diseases using heart sounds," *Computer methods and programs in biomedicine*, vol. 95, no. 1, pp. 47–61, 2009.

[17] P. Narváez, S. Gutierrez, and W. S. Percybrooks, "Automatic segmentation and classification of heart sounds using modified empirical wavelet transform and power features," *Applied Sciences*, vol. 10, no. 14, p. 4791, 2020.

[18] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-hsmm-based heart sound segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 822–832, 2015.

[19] P. Sedighian, A. W. Subudhi, F. Scalzo, and S. Asgari, "Pediatric heart sound segmentation using hidden markov model," in *2014 36th annual international conference of the ieee engineering in medicine and biology society*. IEEE, 2014, pp. 5490–5493.

[20] S. E. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent hidden markov model," *Physiological measurement*, vol. 31, no. 4, p. 513, 2010.

[21] Y. He, W. Li, W. Zhang, S. Zhang, X. Pi, and H. Liu, "Research on segmentation and classification of heart sound signals based on deep learning," *Applied Sciences*, vol. 11, no. 2, p. 651, 2021.

[22] Y. Chen, Y. Sun, J. Lv, B. Jia, and X. Huang, "End-to-end heart sound segmentation using deep convolutional recurrent network," *Complex & Intelligent Systems*, vol. 7, pp. 2103–2117, 2021.

[23] F. Renna, J. Oliveira, and M. T. Coimbra, "Deep convolutional neural networks for heart sound segmentation," *IEEE journal of biomedical and health informatics*, vol. 23, no. 6, pp. 2435–2445, 2019.

[24] Y. Yin, K. Ma, and M. Liu, "Temporal convolutional network connected with an anti-arrhythmia hidden semi-markov model for heart sound segmentation," *Applied Sciences*, vol. 10, no. 20, p. 7049, 2020.

[25] M. Morshed and S. A. Fattah, "A deep neural network for heart valve defect classification from synchronously recorded ecg and pcg," *IEEE Sensors Letters*, 2023.

[26] H. Li, X. Wang, C. Liu, P. Li, and Y. Jiao, "Integrating multi-domain deep features of electrocardiogram and phonocardiogram for coronary artery disease detection," *Computers in Biology and Medicine*, vol. 138, p. 104914, 2021.
[27] J. Li, L. Ke, Q. Du, X. Ding, and X. Chen, "Research on the classification of ecg and pcg signals based on bilstm-googlenet-ds," *Applied Sciences*, vol. 12, no. 22, p. 11762, 2022.
[28] J. Oliveira, C. Sousa, and M. T. Coimbra, "Coupled hidden markov model for automatic ecg and pcg segmentation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 1023–1027.
[29] M. M. Movahedi, M. Shakerpour, S. Mousavi, A. Nori, S. H. M. Dehkordi, and H. Parsaei, "A hardware-software system for accurate segmentation of phonocardiogram signal," *Journal of Biomedical Physics & Engineering*, vol. 13, no. 3, p. 261, 2023.
[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
[31] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
[32] D. Zeng, Y. Wu, X. Hu, X. Xu, H. Yuan, M. Huang, J. Zhuang, J. Hu, and Y. Shi, "Positional contrastive learning for volumetric medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*. Springer, 2021, pp. 221–230.
[33] H. Hu, J. Cui, and L. Wang, "Region-aware contrastive learning for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 291–16 301.
[34] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 546–12 558, 2020.
[35] X. Zhao, R. Vemulapalli, P. A. Mansfield, B. Gong, B. Green, L. Shapira, and Y. Wu, "Contrastive learning for label efficient semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 623–10 633.
[36] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8219–8228.
[37] P. Sarkar and A. Etemad, "Self-supervised ecg representation learning for emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1541–1554, 2020.
[38] J. Y. Cheng, H. Goh, K. Dogrusoz, O. Tuzel, and E. Azemi, "Subject-aware contrastive learning for biosignals," *arXiv preprint arXiv:2007.04871*, 2020.
[39] D. Kiyasseh, T. Zhu, and D. A. Clifton, "Clocs: Contrastive learning of cardiac signals across space, time, and patients," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5606–5615.
[40] P. N. Soni, S. Shi, P. R. Sriram, A. Y. Ng, and P. Rajpurkar, "Contrastive learning of heart and lung sounds for label-efficient diagnosis," *Patterns*, vol. 3, no. 1, p. 100400, 2022.
[41] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3875–3879.
[42] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.
[43] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. Johnson *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, p. 2181, 2016.
[44] M. Perslev, M. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-time: A fully convolutional network for time series segmentation applied to sleep staging," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[45] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
[46] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.
[47] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 529–545.
[48] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," *Advances in neural information processing systems*, vol. 29, 2016.
[49] H. Yang, W. Wen, and H. Li, "Deephoyer: Learning sparser neural network with differentiable scale-invariant sparsity measures," *arXiv preprint arXiv:1908.09979*, 2019.
[50] H. Yang, L. Duan, Y. Chen, and H. Li, "Bsq: Exploring bit-level sparsity for mixed-precision neural network quantization," *arXiv preprint arXiv:2102.10462*, 2021.
[51] J. Mao, X. Chen, K. W. Nixon, C. Krieger, and Y. Chen, "Modnn: Local distributed mobile computing system for deep neural network," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*, 2017, pp. 1396–1401.
[52] T. Zhang, H.-P. Cheng, Z. Li, F. Yan, C. Huang, H. Li, and Y. Chen, "Autoshrink: A topology-aware nas for discovering efficient neural architecture," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6829–6836.