A Deep-Learning-Based Multi-modal ECG and PCG Processing Framework for Cardiac Analysis

Qijia Huang, Huanrui Yang, *Stufdent Member, IEEE*, Eric Zeng, and Yiran Chen, *Fellow, IEEE*

***Abstract*— The need for telehealth and home-based monitoring surges during the COVID-19 pandemic. Based on the recent advancement of concurrent electrocardiograph (ECG) and phonocardiogram (PCG) wearable sensors, this paper proposes a novel framework for synchronized ECG and PCG signal analysis for cardiac function monitoring. Our system jointly performs R-peak detection on ECG, fundamental heart sounds segmentation of PCG, and cardiac condition classification. First, we propose the use of recurrent neural networks and developed a new type of labeling method for R-peak detection algorithm. The new labeling strategy utilizes a regression objective to resolve the previous imbalanced classification problem. Second, we propose a 1D U-Net structure for PCG segmentation within a single heartbeat length. We further utilize the multi-modality of signals and contrastive learning to enhance model performance. Finally, we extract 20 features from our signal labeling algorithms to apply to two real-world problems: snore detection during sleep and COVID-19 detection. The proposed method achieves state-of-the-art performance on multiple benchmarks using two public datasets: MIT-BIH and PhysioNet 2016. The proposed method provides a cost-effective alternative to labor-intensive manual segmentation, with more accurate segmentation than existing methods. On the dataset collected by Bayland Scientific which includes synchronized ECG and PCG signals, the proposed system achieves an end-to-end R-peak detection with F1 score of 99.84%, heart sound segmentation with F1 score of 91.25%, and snore and COVID-19 detection with accuracy of 96.30% and 95.06% respectively.**

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

I. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death worldwide. According to the report by Centers for Disease Control and Prevention (CDC) [1], Coronary Artery Disease (CAD), the most common type of CVDs, killed 360,900 people in 2019. The characteristics of heart diseases that make them especially dangerous are that these diseases are salient and unaware by patients, and when the severity increase, the diseases become deadly in a short time. For a potential patient, physiological signals monitoring may uncover important information for doctors to detect potential CVDs and control the patient's condition. However, during the ongoing pandemic, hospitals are overwhelmed by COVID-19 patients, forcing millions of patients to stay home. Although the communication between patients and doctors can still be carried out via telehealth methods through platforms like Zoom, simple video and audio communication without real-time physiological signal monitoring and analysis made it difficult for doctors to accurately assess the status of their patients. Therefore, it becomes an increasing need for a stay-home self-monitoring system, which can enable the collection of patients' physiological data at home, provide analysis for gathered data, raise alerts for abnormal data and transfer it to doctors for diagnosis, and deliver the clinical treatment plan back to patients. Such a system can be beneficial via the following three advantages: (1) Patients can conduct self-monitoring by having real-time physiological signals analysis; (2) Abnormal signs in the signals can be detected and delivered as automated clinical alert, which could alarm the patient to get treatment, and assist doctors to diagnose the situation; (3) This system can help to distribute the medical resource more effectively and prevent an overwhelming burden for patients with no or mild illness seeking for in-hospital treatment.

This work was supported in part by Bayland Scientific.

Q. Huang, H. Yang, and Y. Chen are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708 USA (e-mail: {qijia.huang, huanrui.yang, yiran.chen}@duke.edu).

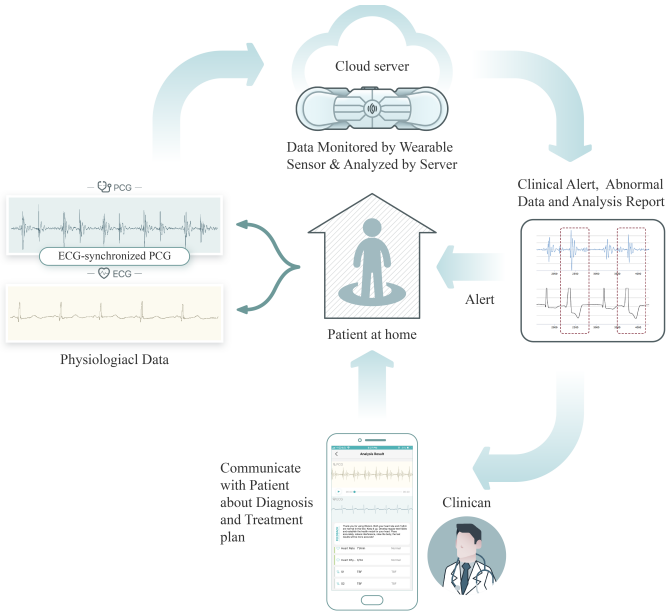

Fig. 1. Illustration of our proposed self-monitoring system workflow. The user's physiological data will first be collected by the wearable device and uploaded to the server. The algorithm implemented on the server will perform data analysis. If abnormal data is detected, a clinical alert will be sent to the user, and the filtered-out data will be sent to the clinician with an analysis report. The clinical can diagnose based on the report and data, and develop a further treatment plan.

In this work, we propose a novel self-monitoring system framework via multi-modal electrocardiograph (ECG) and phonocardiograph (PCG) signal processing, as illustrated in Fig. 1. Specifically, we propose to use a wearable sensor to gather simultaneous ECG and PCG signal. We then develop a deep-learning-based multi-modal signal processing framework to extract key features from ECG and PCG signals, which are useful for providing abnormal alert in downstream clinical tasks. Both ECG and PCG have been respectively identified by previous research as vital signal for monitoring cardiovascular diseases, which are the leading cause of death globally [2], and other diseases. Moreover, the multi-modal processing of ECG and PCG brings further information for monitoring the entire cardiac cycle. For example, the study of Shapiro et al. shows that the prolong of electromechanical activation time (EMAT), as measured by the duration of Q-peak on ECG and peak of first heart sound on PCG, is an indication of left ventricular dysfunction. To the best of our knowledge, this work presents the first deep-learning-based system for multi-model ECG and PCG processing, and for the first time proves that analyzing the joint information of synchronized ECG and

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

PCG has the potential to improve diagnosis accuracy.

The collection of ECG and PCG signal in our work is empowered by a novel piece of technology developed by WENXIN and Bayland Scientific Technology: a band-aid-like wearable ECG and PCG device, as illustrated in Fig. 2. The device has obtained the Chinese National Medical Products Administration (NMPA) approval and has been used in a heart failure study by Li et al. [3] for data collecting purpose. Patients can simply 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 doctor's office in real time. Being a wearable device, the chest sticker enables easy continuous signal monitoring without interfering with normal human activities. Moreover, unlike previous wearable devices such as wristband and life shirt [4] etc. that are loosely connected to the human body, Bayland Scientific's chest sticker significantly reduce the influence of outside noises, especially on the PCG signal. With this device in place, this paper focus on the development of deep-learning-based algorithms to automatically analyze the gathered ECG and PCG data, perform screening and raise alerts for potential issues, and extract critical features to aid the physicians in diagnosing the patient. We focus on the following two problems: (1) How to locate the key points on ECG and PCG accurately; (2) How to transform these identified key points on ECG and PCG into effective features that can be used for solving health problems e.g., detecting diseases in the early stage.

The ECG, which monitors cardiac conditions by detecting the muscular electrical signals, is typically cost-effective in collection and is suitable for processing. Given their wide availability, ECG signals were thoroughly studied [5], [6] by previous researchers, and has enabled downstream tasks like arrhythmia detection. Building upon previous research, in this work we propose the use of LSTM to identify the R-peaks on ECG signals. We further convert the commonly used classification objective to a regression objective, which resolves the problem that number of R-peaks is significantly less than non-R points faced by previous methods.

The PCG is a recording of sound and murmur of heart, and contains more information than ECG signals that can enable further diagnosis. For example, major cardiac disorders like heart failure and valve disease caused by heart structure pathological change are hard to be detected by the electrical signals of ECG; while the cardiac structure change is likely to generate abnormal vibrations that can de captured by the PCG signal. It has been demonstrated that vital biomedical information can be identified by the location of key points and the strength of heart sound, which are all captured by the PCG signal. Collins et al. [7] shows the presence of S3 heart sound has highly positive-correlation with primary heart failure; and the study of Roos et al. [8] shows the left ventricular ejection time (LVET) measured on PCG is negatively correlated with the probability of incident heart failure. Therefore, we find identifying the key points on PCG signals as a vital component of our self-monitoring pipeline for downstream clinical alerts. However, the studies on PCG are lacking compared to the ones on ECG. With only a few successful studies based on the PhysioNet 2016 challenge of normal and abnormal heart sound classification [9], [10]. The reason for the lack of attempts on PCG is that its signal has more complex waveform than ECG, and the collection process usually leads to significant noises. We develop a 1-D CNN-based U-Net model to perform fine-grained segmentation on the PCG signal. The ECG information are also taken as multi-modal input to increase model performance. Also, we applied Contrastive Learning method to pretrain the model to achieve higher accuracy.

Based on the temporal relationship and the amplitude of the key points identified on the ECG and PCG signal, we extract 20 features to perform further analyze on downstream tasks. ANN models are trained with our extracted features to perform critical medical tasks like classifying COVID-19 positive samples and snore detection during sleeping.

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

- We propose a deep-learning-based self-monitoring system which can help allocate medical resources more effectively.
- We propose a novel multi-modal processing framework for the identification of key points on ECG and PCG waveform, which attain state-of-the-art results on multiple benchmarks including both public and private PCG and ECG dataset.
- We extract useful features from ECG and PCG to achieve high performance on downstream medical tasks like COVID-19 detection and snoring detection.
- We for the first time shows the effectiveness of multi-modal ECG and PCG processing in self-monitoring systems.

The content of the paper is organized as follows: Section II introduces the previous research and dataset on ECG, PCG, and downstream tasks. Our proposed R-peak detection algorithm is presented in Section III, along with its evaluation results on the public MIT-BIH and our private Bayland Scientific dataset. In Section IV, the PCG segmentation algorithm is proposed, which we compare with other state-of-the-art methods on the PhysioNet dataset, and analyzes on our private dataset. The downstream tasks of snoring and COVID-19 detection based on features from ECG and PCG are presented in Section V. Finally, Section VI summarizes our work and discusses the advantages of our proposed methods and our future work.

II. Related Work

A. ECG R-peak Detection

Among all the R-peak detection algorithms, the most well-known one is developed by Pan and Tompkins [5] which is widely analyzed by researchers as a benchmark. It is a good representative of a family of algorithms constructed in two parts: signal preprocessing for reducing noise and enhancing the QRS complex, and a peak searching algorithm by setting proper thresholds. These methods are computationally efficient and can be easily deployed on mobile devices. However, in real-world applications, the ECG signal quality may be influenced by different factors. Data collected by various types of sensors or devices tend to have scaling variations as well as very different noisy property and distribution, caused by electronic or other environmental factors during the data-collection process. In each of those settings, a tuning of hyper-parameters for filters or transformation would have to be performed to retain the accuracy. Moreover, even if the sensory data is being continuously collected, other dynamic factors such as body posture could cause temporal variation. Therefore, an algorithm that is robust in complex conditions is needed in order to deal with sensory data with varying quality.

Deep-learning-based methods for processing the ECG signals are developed in recent years. These methods are trained with large training datasets covering different quality signals and can perform

reasonably well in different conditions. Convolutional Neural Network (CNN) based methods have the ability to draw meaningful features locally from the waveform and are robust to noise. Xiang et al. [11] used a two-layer 1D-CNN network for R-peak detection and reporting an accuracy on MIT-BIH dataset of 99.68%. However, the CNN based method relies on a moving window with human-decided window size, so it cannot predict the location of R-peaks directly. There are other works using Recurrent Neural Network (RNN) based methods for ECG analysis which can better utilize temporal information in the signal. Laitala et al. [12] used a Bi-LSTM model and reported a precision of 99.63% on a subset of the MIT-BIH dataset. However, LSTM model can only take in down-sampled data under efficiency constrains, which suffers from the information loss of the down-sampling process. In this work, we combine CNN and LSTM models, leveraging CNN's capability for spatial feature extractions while using LSTM to capture temporal feature relationship. Zhou et al. [13] used a combination of 6-layer CNN and 1-layer LSTM model to detect R-peak and reported an accuracy of 90.68% in ± 25 ms window. While they used a labeling technique that relies on selected window position on signals, we proposed a Gaussian shape labeling technique that can be applied across the entire sequence.

We utilize a widely used dataset on ECG signals: MIT-BIH arrhythmia classification dataset [14] which contains ECG and location of R-peaks. The dataset is sourced from both normal people and patients with arrhythmia. We also utilize a proprietary Bayland Scientific ECG dataset, whose signals are mainly normal except for a small subset of snoring samples and COVID-19 samples. More details regarding the two datasets will be discussed in Section III.

B. PCG Segmentation

There are already several attempts on PCG segmentation tasks. In earlier works, traditional signal processing techniques were applied for segmentation with two steps: first using envelograms [15] or wavelet transform [16] to extract features related to fundamental heart sounds, then marking the peaks based on features extracted using a threshold, and identifying the boundaries of S1s and S2s. Since these traditional methods rely on threshold-based peak-finding algorithms, they can not be generalized to signals from different sources. And these methods are not robust to the significant noise typically associated with PCG signals will.

Researchers have also developed methods using deep learning framework for PCG segmentation. Significant amount of work focus on temporal modeling for PCG segmentation. For example, Logistic Regression Hidden Semi-Markov Model (LR-HSMM) [17] was used to predict the sequence of fundamental heart sounds; In another work, Recurrent Neural Network (RNN) [18] was shown to perform better than Convolutional Neural Network (CNN) in analyzing the sequential states of PCG signals. Since these temporal methods lack the ability to process raw signals, a feature extraction algorithm will be applied to signals first. Normally, frequency-domain features will be extracted and proven to be effective, for example, wavelet transform was used in [17], and Mel-frequency cepstral coefficients (MFCC) were used in [18], [19].

In our proposed approach we first employ our R-peak detection algorithm to separate data into single heartbeats. Then we apply a U-Net model to arrive at a fine-grain segmentation. In this case, the heartbeat signal feeding the U-Net is not treated as having temporal dependency but rather a static object. Therefore, we can leverage CNN based model's strength in recognizing spatial patterns on the raw signals, which allows us to obtain higher accuracy than other algorithms.

The PhysioNet 2016 heart sound classification dataset [9] is widely used as a benchmark for the heart sound segmentation task since it contains PCG signals with simultaneous labels of states of the heart cycle (S1, systole, S2, diastole). However, only a small set of data includes double-channel ECG and PCG which is different from our Bayland Scientific synchronized dataset. For studying the benefit of multi-modal input for PCG segmentation, we will utilize our private Bayland Scientific dataset for thorough analysis while using the double-channel subset of the PhysioNet dataset for comparison with other state-of-the-art algorithms. More details of the two datasets can be found in Section IV.

C. Downstream task

We focus on two downstream health monitoring problems based on ECG and PCG signals: snoring detection during sleeping, and COVID-19 symptom detection. These two tasks are chosen due to the real medical concern of sleep quality and severity monitoring of COVID-19 infection, as well as the need for 24-hour monitoring with mobile devices. However, there has been limited work done on snoring and COVID-19 detection based on ECG and PCG signal analysis. The existing dataset for snoring detection is the SOMNIA database [20], which uses Polysomnography (PSG) for sleep analysis. However, PSG contains multi-modal physiological signals and requires hospital-level equipment. For the COVID-19 study, the DiCOVA database [21] is widely used whose data contains recordings of users' cough sounds, breath, sustained vowel phonation, and speech audio. Unlike Bayland Scientific's gathering of ECG and PCG signals via small-form-factor personal wearable devices, the collection of the DiCOVA dataset requires high-precision microphones which are not easily obtainable in practice. In our system, the signal analysis and abnormal detection are performed end-to-end on the Bayland Scientific collected dataset with dual-channel signals and labels of normal, snoring, and COVID-19 infection samples.

III. ECG R-peak Detection Task

An ECG signal measures heart electrical activities. The time information and strength of activities are revealed by the location and amplitude of QRS complex which can help in disease diagnosis. The correct identification of R-peaks on ECG signal is the first step which enables subsequent QRS and T wave identification and heart rate analysis. Therefore we first develop an algorithm to correctly identify the timing of a complete heartbeat through localizing the R-peaks on ECG signal.

A. Method

1) Labeling: MIT-BIH arrhythmia database included expert-labeled R-peak labels for every heartbeat. However, the number of points corresponding to the R-peak is relatively small compared to that of other points, such that the classes are imbalanced. For a 360 ms heartbeat, the proportion of R-peak to the heartbeat is $\frac{1}{360}$. If we want to detect the position of R-peaks through a classification task that works through each point in the sequence, a weighted classification loss function will have to be used. However, the weight will be hard to decide since the length of heartbeats varies among individuals. Therefore, we transform the task from classification to regression by converting the label to a Gaussian-shaped target. Let x on the time axis and centered at R-peak position which means $x = 0$ at the R-peak, then the label can be represented in the form:

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

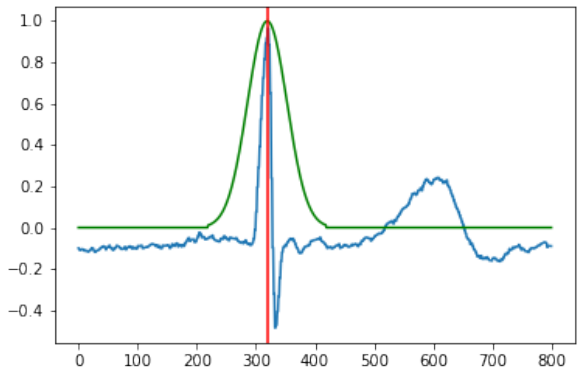

Fig. 3. Illustration of the proposed Gaussian-shaped labeling for R-peak on ECG. The red vertical line shows the true location of R-peaks. The green curve is the Gaussian-shaped label.

TABLE I
CONV-LSTM MODEL STRUCTURE

Layer	Output Dim	Stride	Activation
1D Convolution k=101	8	1	Relu
1D Avg-Pooling	8	2	-
1D Convolution k=101	1	1	Relu
1D Avg-Pooling	8	2	-
LSTM	8	-	Tanh
Fully Connected	1	-	-

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

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

Fig. 3 shows a typical illustration of how Gaussian-shaped label apply on ECG in the range of one heartbeat. In the standard setting, the shape of one Gaussian label covers the R-peak, and at the peak, the Gaussian-shaped label has a value of 1. This setting will help the training process with normalized ECG signal as its input and mean squared loss (MSE) as loss function.

2) CNN-LSTM Model: In this work, we propose a network combining CNN and LSTM for the R-peak detection task, which is illustrated in Table I. The proposed model contains 2 1D-CNN layers each followed by an average pooling layer. These 1D layers will extract high dimensional spatial features from the input ECG signals and perform down-sampling to decrease hidden outputs' dimensions. Following the stacked convolutional layers is the LSTM layer which helps the model to capture temporal information. The output of the LSTM layer is fed into a fully connected layer to make a regression to our Gaussian labels. Relu is used as the activation function of the CNN layers and the fully connected layers. The Tanh activation function is used for the LSTM layer.

3) Loss function: A loss function is used to measure the difference between the model predicted labels and the true label, which is then used for calculating the gradient so to dictate the direction of optimization. In the experiment, we use Mean Squared Error (MSE) Loss function to perform regression task. The optimization object is shown as:

$$argmin_\theta \frac{1}{B}\Sigma_{i=1}^{B}(y_i - f(S_i;\theta))^2, \tag{3}$$

where B denotes the batch size, (S_i, y_i) is a paired sample of input signal sequence and Gaussian-shaped label from the training dataset, and the function $f(\cdot;\theta)$ denotes the operations made by the Conv-LSTM model. Thus our optimization aims to minimize the difference between the model prediction and the smooth approximation of binary R-peak positional labeling.

TABLE II
RESULT OF EVALUATION ON MIT-BIH DATABASE

Method	Precision (%)	Recall (%)	F1 (%)
Pan&Tompkins [5]	99.56	99.76	99.65
CNN [11]	99.91	99.77	99.83
LSTM [12]	99.62	99.53	99.57
Proposed Gaus LSTM	99.70	99.68	99.68

TABLE III
R-PEAK LABELING STRATEGY STUDY ON BAYLAND SCIENTIFIC DATASET

Method	Precision (%)	Recall (%)	F1 (%)
Categorical 50ms	99.96	99.11	99.46
Proposed Gaus 50ms	99.89	99.84	99.84
Categorical 25ms	90.98	90.40	90.64
Proposed Gaus 25ms	96.68	96.63	96.63

B. Experiment

1) Dataset: We use both the MIT-BIH database and our private dataset collected by Bayland Scientific to evaluate our method.

The MIT-BIH database has reportedly been used in many publications. This database includes 48 heartbeat recordings at 360 Hz from 47 different patients. Each recording is 30 minutes in duration and contains two leads: 1) a modified Lead II collected on the chest; 2) another channel from lead V1, V2, V5, or V4. The database has been employed by researchers to test algorithms for QRS detection, arrhythmia detection, and classification. In our experiment, the first channel from Lead II is used as input, and R-peak labeling is used as the target variable.

The second dataset we used is collected using wearable devices by Bayland Scientific company. This private dataset contains 2076 pieces of dual-channel synchronized ECG and PCG signals with an average length of 50 seconds. The sampling rate of recordings is 1000 Hz. We also use this set of multi-modal sequential data for the following PCG segmentation task. However, in the scope of the ECG R-peak identification task, we only use the ECG channel. This ECG signal is a single lead signal of Lead II which is the same one used in the MIT-BIH dataset. All the positions of R-peak are fully labeled by professionals from Bayland Scientific.

2) Evaluation Metrics: The algorithm performance is evaluated with Precision, Recall, and F1. We measure the True Positives, False Positives and False Negatives within a tolerance of 50 ms for the public MIT-BIH dataset. For the private dataset, we examined the performance within a tolerance of 50 ms and 25 m.

3) Implementation: The initial learning rate is set to 0.1, and the SGD optimizer is used to train the proposed model with momentum set to 0.8. If the value of loss on the validation set cannot be reduced within five epochs, the training process will be terminated. And we set the maximum training epochs to 100.

C. Results

1) MIT-BIH Result: Table II summarizes the performance of our algorithm for the MIT-BIH database. Most of the online algorithms achieved high accuracy on this challenge. Our work achieved the same level of accuracy when compared to models that operate on an entire sequential ECG and produce continuous labels.

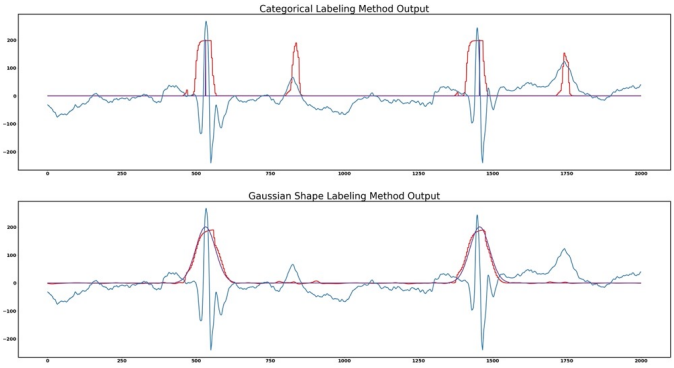

Fig. 4. Illustration of typical Conv-LSTM model prediction for R-peak, the red line is the prediction and the purple line is the true label. The above figure is the output with a categorical labeling strategy and the below one is with Gaussian-shaped labeling.

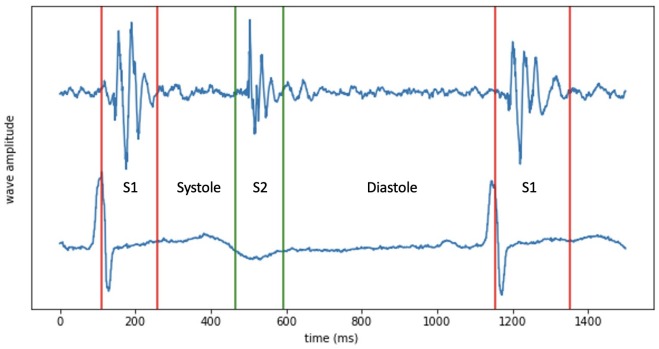

Fig. 5. Illustration of four states of heart cycle with ECG and PCG signal: S1, Systole, S2, and Diastole. R-peaks are labeled as an approximate start of S1. Systole and diastole labeled in between S1 and S2, actually start at the beginning of S1 and S2 respectively.

2) Private Result: On our private dataset, we perform an ablation study on the choice of continuous Gaussian shape labeling and categorical labeling. In this experiment, the categorical label for R-peak is 1 and 0 for the rest in a given sequence. The loss function we used is weighted Cross-Entropy Loss with a weight of $1:550$ for class 0 vs class 1. The predictions made by the model are shown in Fig. 4. Table III shows the results of our experiment. We can observe the lower recall for categorical labeling strategy which indicates the model is likely to incorrectly identify normal points on ECG signals as R-peaks. It is possibly caused by the sub-optimal class weight setting for Cross-Entropy Loss, while the Gaussian-shaped labeling does not have such a problem.

IV. PCG Segmentation Task

Heart sound or its graphical representation phonocardiogram (PCG) is one of the commonly used physiological data for diagnosing cardiac diseases including arrhythmia, heart failure, etc. In computer-aided heart sound analysis, diagnosis can be divided into two parts: signal segmentation and classification. An accurate classification model relies on the precise localization of first/second heart sounds and other states on the heart sound signal. During a cardiac cycle, the heart experiences atrial and ventricular contractions. These vibrations generate sounds, the magnitude of which can be displayed on PCG. Most commonly, we can observe four states in one heartbeat cycle: first (S1) and second (S2) heart sound, systole whose beginning is signaled by the start of S1, and diastole whose beginning is signaled by the start of S2, as shown in Fig. 5. Since we already developed an algorithm that can identify the location of R-peaks on ECG signal, it can also be used to identify the PCG heart sound cycles. Thus, our next task is to segment the individual heart states given a particular heartbeat extracted by our R-peak identification algorithm. PCG segmentation is a more challenging task since the similarity between S1 and S2 peaks, and the occasionally appearing sinus rhythm which is considered as noise compare to our major objects: S1 and S2.

A. Method

1) Data Preprocessing: Let s denote a normalized 2-channel signal from ECG and PCG which contains N heartbeats as segmented by the R-peak identification algorithm. Each heartbeat is defined as starting from 100ms preceding one R-peak and ending at 100ms preceding the next R-peak. In such a setting, the heartbeat period includes all four states of heart cycle,

$$s = [b_1, ..., b_N] \tag{4}$$

Since for CNN processing, the inputs are required to have the same shape, we resize the heartbeat sequence to have the same length. For those heartbeats with a length less than 1536, we pad them to 1536 with zeros after the original sequence; For those longer ones, we only keep the first 1536 samples. After resizing the signals, we obtain a set of equal-length heartbeats from original sequence for our training set: $X = [b'_1, ..., b'_N]$ and $Y = [y_1, ..., y_N]$, where each y_i is an 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 wide usage of 2D U-Net in biomedical image segmentation tasks, we designed a 1D variant of U-Net in our PCG segmentation framework. Similar to the original U-Net, we maintain the Encoder-Decoder structure and the 2 convolutional layers per block architecture with Relu activation function and skip connections. Meanwhile, we change all the convolution layer, max-pooling layer, and up-convolution layer to one-dimensional. Also, we adjust the number of filters in each convolution layer to better extract the spatial features for 1D signals. The kernel size is set to 7 for smooth feature extraction on signals. The convolution is applied with padding for keeping sizes of input and segmented output the same. We use the average pooling layer instead of max-pooling layer. And the step of pooling layer and up-convolution layer is set to 4. The structure of the model is shown in Fig. 6.

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:

$$argmin_\theta \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in (0,1,2)} p_j(y_i) \log f_j(b_i; \theta), \tag{5}$$

where (b_i, y_i) is a pair of point on heartbeat and label at $x = i$; N is the heartbeats 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.

4) Pretraining: Supervised learning models rely on large training samples to boost generalization ability. However, the data labeling process is both expensive and time-consuming. This is particularly a tougher problem in the biomedical field than in the fields of

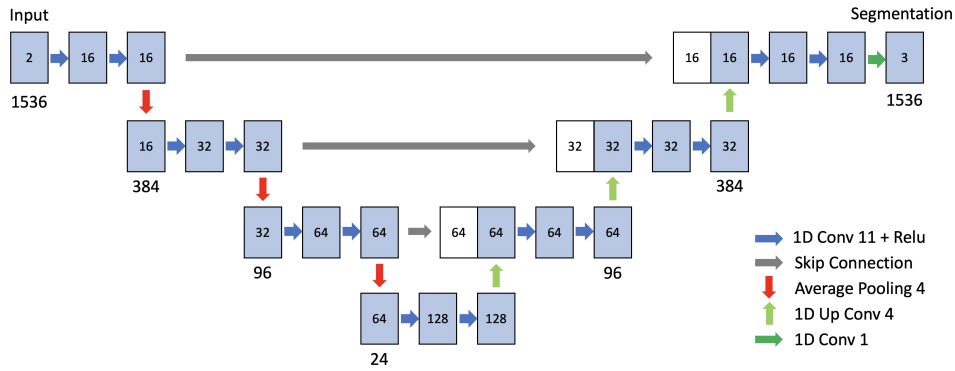

Fig. 6. Illustration of Proposed U-Net Structure. In this example, the input length is set to 1536 with 2 channels of ECG and PCG. Numbers in the box represent the channel of signals, and numbers below the box represent the signal length.

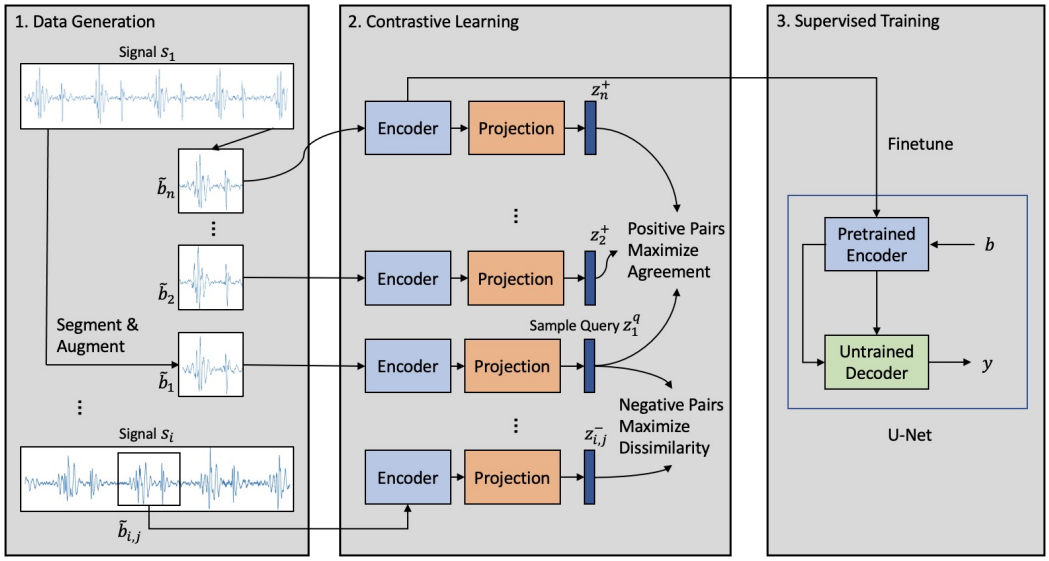

Fig. 7. Illustration of Contrastive Learning Framework

computer vision or natural language, as it requires skilled professionals. Self-supervised pretraining has recently been established as a successful technique in terms of having the capacity to learn better-generalizable representations to improve the accuracy of supervised training through the use of unlabeled data. Among all the pretraining tasks, contrastive learning is one of the advanced techniques in self-supervised pretraining learning and has shown outstanding results on tasks dealing with images. However, for the data associated with a single heartbeat, typical image data augmentation methods like clip or rotation used by SimCLR [22] is not suitable. Here we propose a new contrastive learning method that is effective in training a model with encoder-type structure to obtain general representations for single heartbeat signals.

Our work is inspired by the work proposed by Saeed et al. [23], the general idea is to distinguish the source of segmented signal. The first step is to perform data augmentation on all the heartbeat signals. For each input heartbeat, b, we generate one random augmentation, $b\prime = Aug(b)$. Here we apply two types of augmentation: the frequency masking method from SpecAugment [24]; and masking the ECG channel since we want to guide our encoder to rely more on features from PCG channel. Each of the augmentation does not perturb the location information of the original signals. Then we pass the signals to the encoder of our U-Net, Enc, which maps $b\prime$ to a representation vector, $r = Enc(b\prime)$. Then the representation is passed through a projection network, $Proj$, which maps r to a vector $z = Proj(r)$. We instantiate $Proj$ as a single linear layer perceptron of size 128. We normalize the output of this network to lie on the unit hypersphere, which allows using an inner product to measure distances in the projection space. As in self-supervised contrastive learning [25], we discard $Proj$ at the end of contrastive training and use the output of Enc to perform the downstream supervised training.

Since our contrastive learning strategy involves arbitrary number of positive samples for an anchor one, we refer to the supervised contrastive loss function introduced in [26]. For a set of N heartbeat subset, $\{b_k, y_k\}_{k=1,\ldots,N}$ randomly chosen from the contrastive training dataset, in which the label y_k is the index of source ECG signal of each heartbeat, its corresponding batch consists of $2N$ pairs of training samples, $\{\tilde{b}_l, \tilde{y}_l\}_{l=1,\ldots,2N}$, where $\tilde{b}_{2k}$ and $\tilde{b}_{2k-1}$ are two random data augmentations of b_k and $\tilde{y}_{2k} = \tilde{y}_{2k-1} = y_k$. Then in the supervised contrastive learning setting, within a batch, let $i \in \{1, .., 2N\}$ and $z_l = Proj(Enc(\tilde{b}_l))$, then the loss of a batch can be defined as :

$$L^{CL} = \sum_{i=1}^{2N} L_i^{CL} \tag{6}$$

$$L_i^{CL} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(sim(z_i, z_p)/\tau\right)}{\sum_j^{2N} \mathbf{1}_{j \neq i} \exp\left(sim(z_i, z_j)/\tau\right)}, \tag{7}$$

where the $P(i)$ denotes the set of indices of positives samples in this batch which share the same source signal as anchor sample i, and $cos(z_i, z_l)$ computes the cosine similarity between the latent embeddings z_i and z_j. By optimizing this loss function, the Encoder of the U-Net will pull all the positive samples with their augmentations together and push away the negatives. The framework of contrastive learning is shown in Fig. 7.

B. Experiment

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

The Physionet 2016 database 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 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 [17] and manually decide their correctness.

There are five training sets provided by the challenge while only 'training-a' subsets contain 2 channels signals of ECG and PCG. So we will only use the 'training-a' subset including 409 recordings for training and test purpose.

The second dataset we used is the Bayland Scientific dual-modal dataset introduced earlier. We use the synchronized ECG and PCG signals as input. The positions of S1 start, S1 end, S2 start, and S2 end are identified and labeled by professionals from Bayland Scientific.

2) Evaluation Metrics: We evaluate the performance of our PCG segmentation algorithm using five metrics: positive predictive value (PPV), sensitivity (Se), specificity (Spe), F1 score, and accuracy (Acc). Unlike the ECG task which is evaluated on detecting R-peaks, the performance of PCG segmentation is evaluated for each state of the heart cycle. The evaluation metrics, except the accuracy, for each state in terms of whether our algorithm correctly classifies this target state can be denoted as:

$$PPV = \frac{TP}{TP + FP} \tag{8}$$

$$Se = \frac{TP}{TP + FN} \tag{9}$$

$$Spe = \frac{TN}{TN + FP} \tag{10}$$

$$F1 = 2\frac{PPV \cdot Se}{PPV + Se} \tag{11}$$

Since these metrics are calculated for each of the four states, to evaluate the overall performance of the segmentation algorithm, we arrive at the final metrics by globally averaging among the four classes. The accuracy is calculated globally by the number of the correctly classified states divided by the number of pixels:

$$Acc = \frac{TP}{TP + TN + FP + FN} \tag{12}$$

3) Implementation: The initial learning rate is set to 0.001 for performing Contrastive Learning and training the U-Net segmentation model from scratch, as well as for finetuning the U-Net segmentation model with pretrained encoder to ensure the preservation of the general feature extraction ability of the encoder. The Adam optimizer is used to train the proposed model. If the value of the loss on the validation set cannot be reduced within five epochs, the training process will be terminated. And we set the maximum training epoch to 150 epochs for both encoder training and U-Net training.

TABLE IV
RESULT ON PHYSIONET 2016 DATASET.

Method	PPV	Se	Spe	Acc	F1
U-Net [27]	93.2	92.3	98.2	95.0	92.7
BiLSTM Attention [18]	96.3	97.2	97.5	96.9	96.70
BiLSTM Attention (our imp)	94.2	95.0	95.1	93.5	94.75
GRNN [19]	94.9	95.9	-	-	95.4
Proposed U-Net+Pretrain	96.21	96.04	99.16	97.59	96.12

TABLE V
RESULT ON BAYLAND SCIENTIFIC DATASET.

Input Signal	PPV	Se	Spe	Acc	F1
PCG Only	90.4	91.24	98.48	95.85	90.68
PCG+ECG	90.73	91.76	98.49	95.89	91.11

C. Results

1) PhysioNet Results: Table IV shows the evaluation result of the proposed segmentation algorithm on the PhysioNet dataset compared with other state-of-the-art algorithms. All numbers in the table are in percentage. For our proposed method, contrastive learning and multi-modal inputs are included and applied as aforementioned. The baseline models we compared with include a GRNN model [19], a U-Net model [27] and a Bi-LSTM with attention mechanism [18]. For the GRNN model, we refer to the performances declared by their published results. For the U-Net and LSTM with attention models, due to the unavailability of public implementation and the difference in evaluation metrics, we implement a similar algorithm with the 1D U-Net and LSTM with attention structure, and report the performance with our metrics.

Through the observation, the recurrent neural network achieves overall better results than the regular 1D U-Net. This suggests that the temporal models and their Frequency-domain feature-extraction methods are powerful in processing cardiac sequential data. However, with our proposed pretraining and multi-modal techniques, the U-Net becomes competitive. The reason is that by using an R-peak detection algorithm to select a proper heartbeat-long window size, the CNN-based method is now able to take advantage to perform fine-grain segmentation. As the recurrent network can only make a classification on a selected small window, such approximation is prone to make less accurate predictions at the boundary of fundamental heart sounds.

2) Ablation Study Results:

Ablation to Benefit of Multi-Modal Inputs. Our dataset contains two channels of synchronized ECG and PCG signals, which measure different aspects of cardiac activities. It is natural to assume that analyzing the ECG together with PCG would benefit the segmentation of heart sound. Thus we set out to quantify the benefits of employing

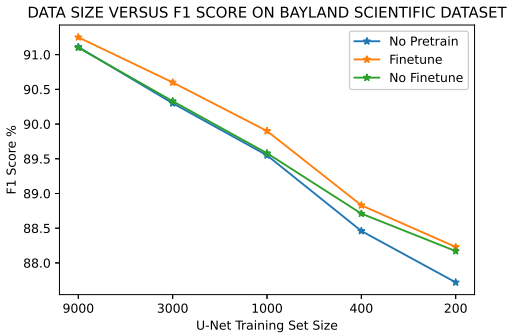

Fig. 8. Illustration of labeled data size influence on the performance of U-Net with or without pretrained Encoder. The Encoder is pretrained with full training data set of 9000 labeled heartbeats.

multi-modality. Table V displays the results of experimenting with single-channel analysis vs dual-channel analysis on the Bayland Scientific dataset. Thus we have demonstrated that by adopting a multi-modal perspective, the model indeed achieves higher accuracy and F1 score as a result of leveraging the joint information content.

Ablation to Pretraining Effect on Reduced Training Data Size Besides improving the performance of the U-Net model, another important reason to utilize contrastive learning pretraining is to reduce the need for labeled training data. Therefore, we design an experiment to examine the efficacy of contrastive learning in terms of data labeling requirements. We train the U-Net from scratch, and with pretrained Encoder respectively with labeled data sizes of 9000, 1000, 400, and 200 heartbeats. In the pretraining setting, the Encoder is trained with 9000 unlabeled data under the contrastive learning framework. When we train the U-Net with pretrained Encoder, we do the experiments with Encoder's gradient frozen and finetuned. We evaluate the result with the F1 score as the metric. Fig. 8 shows the results of this experiment.

As can be observed, utilizing the pretrained Encoder brings about consistent accuracy improvement. The advantage of using pretrained Encoder with finetuning than training the U-Net model from scratch becomes more evident when the training set size is reduced. With the training set size shrinks from 9000 to 3000, 1000, 400 and 200, the differences of F1 score increases from 0.14% to 0.30%, 0.35%, 0.37%, and 0.51%. This enlargement of performance difference demonstrates the Encoder pretrained by the Contrastive Learning method has weights initialized with better generalization ability than random weights initialization. Thus, our proposed pretraining method is able to perform better when the data size is relatively small. Furthermore, by observation, the model with gradient-frozen pretrained Encoder has a similar F1 score as the model trained from scratch when the training data size is large but achieves higher a F1 score when the data size shrinks. This trend indicates our pretrained Encoder is able to produce a better-informed and more-generalized representation of the input.

V. Downstream Tasks

We developed the PCG segmentation algorithm that can be deployed on wearable devices. In order to examine its effectiveness in solving the real-world biomedical problems, we conducted two downstream experiments: COVID-19 sample classification, and snore/non-snore classification from heart sound during sleep. Both tasks try to solve problems based on features extracted from ECG and PCG signals using locations of R-peak, S1 and S2 heart sound localized by R-peak identification and PCG segmentation algorithms.

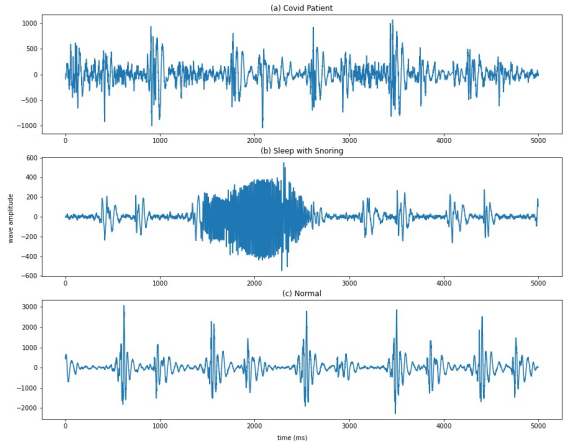

Fig. 9. Example of typical PCG signals. Subfigures (a), (b), and (c) correspond to COVID-19 patient, snoring and normal PCG signals.

A. COVID-19 Patient Classification Description

During the COVID-19 pandemic, physicians employed wearable ECG and PCG sensors to monitor the ECG and heart sound of COVID-19 patients in a way that minimized direct contact. One discovery that was found is that for moderate to severe COVID-19 patients, there were distinctive respiration sounds embedded within the heart sound that was captured by the wearable sensor. This respiration sound is due to lung infections, and many patients exhibiting the symptom were put on ECMO within 48 hours. As for COVID-19 patients self-quarantined at home, the wearable ECG and PCG sensor can be used to remotely monitor the patient. That way, if such respiration sound is detected, the patient will know to seek immediate medical attention. However, due to a large number of COVID-19 patients, it is impossible to have physicians manually process the PCG of every patient and provide an accurate diagnosis in a timely fashion. Using an automated AI-based algorithm to detect the said respiration sound or acoustic signature from the PCG signal expedites the process significantly.

There is already a widely used dataset along with a prediction task proposed online that is intended to detect COVID-19 positive cases based on acoustic signals [21]. Such signals include breathing, cough, voice and etc. The acoustic data needed to be collected intentionally which proves difficult in COVID-19 self-monitoring. Therefore, if the acoustic signature that indicates serious COVID-19 can be detected solely based on the ECG and PCG, it will help people monitor their COVID-19 infection status using personal wearable devices.

In this experiment, we collected a set of synchronized ECG and PCG data from 92 COVID-19 positive patients and 95 negative ones. Recordings were collected by Bayland Scientific wearable devices, and have a standard length of 60 seconds sampled at 1-kHz frequency. Subfigure (a) from Fig. 9 illustrates a typical type of PCG from a COVID-19 positive patient. From observation, a particular characteristic of this type of PCG signal is the overlay of heavy breath noise added on top of a normal waveform, which is caused by the attack on the respiratory system by COVID-19 infection.

B. Snoring Classification Description

Another important application of wearable ECG and PCG sensors is sleep monitoring. PCG recordings contain not only the heart sound but also the snores of the user. Identifying the presence and timing

TABLE VI
FEATURES FOR DOWNSTREAM TASKS

Index	Heartbeat-wise	Extract from Heartbeat
1,2	mean,SD	RR interval
3,4		S1 interval
5,6		S2 interval
7,8		systolic interval
9,10		diastolic interval
11,12		$\frac{\text{systolic interval}}{\text{RR interval}}$
13,14		$\frac{\text{diastolic interval}}{\text{RR interval}}$
15,16		$\frac{\text{systolic interval}}{\text{diastolic interval}}$
17,18		$\frac{\text{amplitude during systole}}{\text{amplitude during S1}}$
19,20		$\frac{\text{amplitude during diastole}}{\text{amplitude during S2}}$

TABLE VII
DOWNSTREAM TASKS

Task	Accuracy	AUROC
COVID Classification	95.06 $\pm$(1.07)	98.44 $\pm$(0.57)
Snore Classification	96.30 $\pm$(1.86)	99.42 $\pm$(0.24)

of the snores can help doctors identify how vital signs (ECG, blood oxygen level, heart sound, and heart function) change before and after the snore. However, because sleep monitoring is a lengthy process of roughly 8-10 hours, manually identifying every snore in the PCG recording is unfeasible and thus an AI-based algorithm is needed to automate the process. While most sleep quality monitoring involves comprehensive polysomnography (PSG) study which captures a multitude of signals including ECG, electromyogram (EMG), eye movements (EOG), electroencephalogram (EEG), etc., snoring detection can be achieved by looking at fewer types of signals. Here we perform a snoring classification task to detect if a snore happens during a 10-seconds sleep period using the dual-channel synchronized ECG and PCG signals. Data we collected includes 320 samples with snoring presence and 414 samples without. Each recording has a length of 10 seconds with a sampling frequency of 1-kHz. Subfigure (b) from Fig. 9 illustrates a typical type of PCG of a snoring sample. From observation, the salient characteristic of this type of PCG signal is the continuous high-frequency noise added on top of the normal waveform, and the noise typically lasts for one to two heartbeat periods.

C. Method

Inspired by the method used in the famous Heart Sound Classification challenge introduced by Liu et al. [9], we extract features based on key points locations on PCG signals. Twenty features that are the same as the one used in PhysioNet 2016 were extracted for each 10-seconds signals as shown in Table VI. For each task, we trained a three-layer Neural Network which contains 68,610 parameters to perform binary classification.

D. Experiment

Considering that the size of the dataset is small and we want to minimize the impact of train-test set partitions, we use K-fold cross-validation where $K = 5$ to test the performance of our model with selected features. Table VII shows the corresponding results from 5-fold cross-validation. On average, our classification algorithm based on features extracted from key points on PCG signals achieved 95.06% accuracy on COVID-19 classification and 96.30% on snore classification. It is worth noticing that the COVID-19 task is relatively harder since the noise signature may be small and subtle to detect in certain cases due to their relatively mild symptoms. The scores of AUC on two tasks are also high since the dataset is relatively balanced between the two classes. Also note that the emphasis here is not to present a new method on these specific downstream problems, but rather to demonstrate that high accuracy can be achieved by simple downstream algorithms leveraging features extracted by our upstream segmentation algorithm.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we introduce a deep-learning-based self-monitoring system. For this system, we developed a deep learning framework aimed at automatic multi-modal physiological signal analysis and abnormal health condition detection. Our proposed method for key points detection on ECG signals with label smoothing achieves state-of-the-art results on the MIT-BIH dataset. We qualitatively analyze the advantages of regression on Gaussian-shaped labels than classification on categorical labels. Based on the R-peak detection method, our proposed beat-wise heart sound segmentation method achieves competitive results compared with other state-of-the-art algorithms on PhysioNet 2016 dataset. Moreover, we examined the benefit of using synchronized multi-modal ECG and PCG signals for segmentation instead of using only a single channel. The evaluation of contrastive learning as an unsupervised pretraining method for the Encoder of U-Net demonstrates both the performance improvement via pretraining on the segmentation tasks and further relieves the reliance on costly labeled data.

Furthermore, our downstream experiment proves the advantage that our system can be easily extended to other types of cardiac diagnosis based on ECG and PCG signals. The pipeline and features can be transferred without modification. This algorithm also possesses the flexibility to support feature modifications and additions, and features crafted for specific downstream tasks can be inserted into the training pipeline appropriately. The only preparation needed is to collect a small dataset for the new tasks. The labeling work is relatively easy for classification tasks like those we did for our snore and COVID-19 detection by assigning the sample signal a normal or abnormal flag. The size of the dataset can also be kept small as demonstrated by the small-size collection of only 320 signals with a length of 10-seconds for our snore detection task.

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 resource available on the target hardware. To enable wider and more efficient deployment of deep-learning-based methods, model compression techniques like pruning [28], [29], quantization [30], [31], and neural architecture search [32], [33] on vital signal processing tasks are worth investigating. Besides efficiency, the deep learning method benefits from the availability of more labeled data, which potentially requires large-scale health data collection, crowdsourcing across different users as well as clinical facilities. In these scenarios, data privacy protection is of paramount importance and typically requires personal health data not leaving the user's own device or associated clinical or research facility. Future research will focus on extending and adapting federated learning methodologies [34] to our multi-modal ECG and PCG processing framework, which will support real-world applications targeting better performance for healthcare applications while preserving data privacy.

REFERENCES

[1] Centers for Disease Control and Prevention, "Heart disease facts," [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm. Accessed on: Apr. 14, 2022.

[2] K. Mc Namara, H. Alzubaidi, and J. K. Jackson, "Cardiovascular disease as a leading cause of death: how are pharmacists getting involved?" *Integrated pharmacy research & practice*, vol. 8, p. 1, 2019.

[3] 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.

[4] H. Ren, H. Jin, C. Chen, H. Ghayvat, and W. Chen, "A novel cardiac auscultation monitoring system based on wireless sensing for healthcare," *IEEE journal of translational engineering in health and medicine*, vol. 6, pp. 1–12, 2018.

[5] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE transactions on biomedical engineering*, no. 3, pp. 230–236, 1985.

[6] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ecg classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2015.

[7] 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.

[8] 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.

[9] 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.

[10] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[11] Y. Xiang, Z. Lin, and J. Meng, "Automatic qrs complex detection using two-level convolutional neural network," *Biomedical engineering online*, vol. 17, no. 1, pp. 1–17, 2018.

[12] J. Laitala, M. Jiang, E. Syrjälä, E. K. Naeini, A. Airola, A. M. Rahmani, N. D. Dutt, and P. Liljeberg, "Robust ecg r-peak detection using lstm," in *Proceedings of the 35th annual ACM symposium on applied computing*, 2020, pp. 1104–1111.

[13] P. Zhou, B. Schwerin, B. Lauder, and S. So, "Deep learning for real-time ecg r-peak prediction," in *2020 14th International Conference on Signal Processing and Communication Systems (ICSPsCS)*. IEEE, 2020, pp. 1–7.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] M. M. van Gilst, J. P. van Dijk, R. Krijn, B. Hoondert, P. Fonseca, R. J. van Sloun, B. Arsenali, N. Vandenbussche, S. Pillen, H. Maass *et al.*, "Protocol of the somnia project: an observational study to create a neurophysiological database for advanced clinical sleep monitoring," *BMJ open*, vol. 9, no. 11, p. e030996, 2019.

[21] A. Muguli, L. Pinto, N. Sharma, P. Krishnan, P. K. Ghosh, R. Kumar, S. Bhat, S. R. Chetupalli, S. Ganapathy, S. Ramoji *et al.*, "Dicova challenge: Dataset, task, and baseline system for covid-19 diagnosis using acoustics," *arXiv preprint arXiv:2103.09148*, 2021.

[22] 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.

[23] 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.

[24] 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.

[25] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European conference on computer vision*. Springer, 2020, pp. 776–794.

[26] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[27] 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.

[28] 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.

[29] H. Yang, W. Wen, and H. Li, "Deephoyer: Learning sparser neural network with differentiable scale-invariant sparsity measures," *arXiv preprint arXiv:1908.09979*, 2019.

[30] Z. Dong, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "Hawq: Hessian aware quantization of neural networks with mixed-precision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 293–302.

[31] 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.

[32] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-all: Train one network and specialize it for efficient deployment," *arXiv preprint arXiv:1908.09791*, 2019.

[33] 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.

[34] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.