Temporal-Framing Adaptive Network for Heart Sound Segmentation Without Prior Knowledge of State Duration

Xingyao Wang, *Student Member, IEEE*, Chengyu Liu, *Senior Member, IEEE*, Yuwen Li, Xianghong Cheng, Jianqing Li, and Gari D. Clifford, *Senior Member, IEEE*

***Abstract*—*Objective:* This paper presents a novel heart sound segmentation algorithm based on Temporal-Framing Adaptive Network (TFAN), including state transition loss and dynamic inference. *Methods*: In contrast to previous state-of-the-art approaches, TFAN does not require any prior knowledge of the state duration of heart sounds and is therefore likely to generalize to non sinus rhythm. TFAN was trained on 50 recordings randomly chosen from Training set A of the 2016 PhysioNet/Computer in Cardiology Challenge and tested on the other 12 independent databases (2,099 recordings and 52,180 beats). And further testing of performance was conducted on databases with three levels of increasing difficulty (LEVEL-I, -II and -III). *Results:* TFAN achieved a superior F_1 score for all 12 databases except for 'Test-B,' with an average of 96.72%, compared to 94.56% for logistic regression hidden semi-Markov model (LR-HSMM) and 94.18% for bidirectional gated recurrent neural network (BiGRNN). Moreover, TFAN achieved an overall F_1 score of 99.21%, 94.17%, 91.31% on LEVEL-I, -II and -III databases respectively, compared to 98.37%, 87.56%, 78.46% for LR-HSMM and 99.01%, 92.63%, 88.45% for BiGRNN. *Conclusion:* TFAN therefore provides a substantial improvement on heart sound segmentation while using less parameters compared to BiGRNN. *Significance:* The proposed method is highly flexible and likely to apply to other non-stationary time series. Further work is required to understand to what extent this approach will provide improved diagnostic performance, although it is logical to assume superior segmentation will lead to improved diagnostics.**

***Index Terms*—Heart sound segmentation, phonocardiogram, deep neural networks, hidden semi-Markov models.**

I. Introduction

CARDIAC auscultation, for identifying heart sounds, is commonly the first step and the most cost-effective measure for screening the various heart dysfunction, even though the final diagnosis is based on the combined analysis from a series of electrophysiologic study or ultrasound recordings. Heart sounds can reflect the hemodynamic processes of the heart and identify some representative symptoms of different diseases, including arrhythmia, valve disease, pulmonary hypertension, heart failure, among other issues [1]. However, only about 20% of medical interns can effectively detect heart conditions using auscultation [2], and extensive training is necessary for human expert evaluation. Automatic and accurate analysis of the recording of heart sounds (the phonocardiogram, or PCG) can be useful for auxiliary diagnosis in clinical applications, and it can potentially assist interns with less developed skills.

The segmentation of heart sounds is a critical step in the automatic analysis of PCG. Accurate localization of fundamental components in PCG is a pre-condition of mining more specific pathological information, including the preliminary diagnosis of specific pathogenic sites and severity levels of these heart diseases [3]. Although unsupervised approaches can facilitate classification or prediction, the lack of interpretability is likely to be a significant barrier to clinical adoption.

Each heart cycle usually consists of a sequence of temporally-constrained states; the first heart sound (S1), the systolic period, the second heart sound (S2) and then the diastolic period (Fig. 1). Segmentation of the PCG into these states facilitates further (pathological) feature extraction within different periods of each heart cycle, e.g., the audible third and fourth heart sound (S3 and S4), murmurs, ejection clicks, pericardial "knock," etc. In addition, segmentation into these states allows for the detection of abnormalities in the timing of different sounds. For example, a mid or late systolic click is most likely a diagnostic indicator

Manuscript received December 31, 2019; revised April 24, 2020, June 30, 2020, and July 11, 2020; accepted July 13, 2020. Date of publication July 20, 2020; date of current version January 20, 2021. This work was supported in part by the Distinguished Young Scholars of Jiangsu Province under Grant BK20190014, in part by the National Natural Science Foundation of China under Grant 81871444 and in part by the Primary Research & Development Plan of Jiangsu Province under Grant BE2017735, and in part by the National Institutes of Health-sponsored Research Resource for Complex Physiologic Signals under Grant R01GM104987. *(Corresponding authors: Chengyu Liu; Xianghong Cheng.)*

Xingyao Wang is with the School of Instrument Science and Engineering, Southeast University.

Chengyu Liu is with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China, and also with the State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China (e-mail: chengyu@seu.edu.cn).

Yuwen Li and Jianqing Li are with the School of Instrument Science and Engineering, Southeast University.

Xianghong Cheng is with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: xhcheng@seu.edu.cn).

Gari D. Clifford is with the Department of Biomedical Informatics, Emory University School of Medicine, and also with the Department of Biomedical Engineering, Georgia Institute of Technology.

Digital Object Identifier 10.1109/TBME.2020.3010241

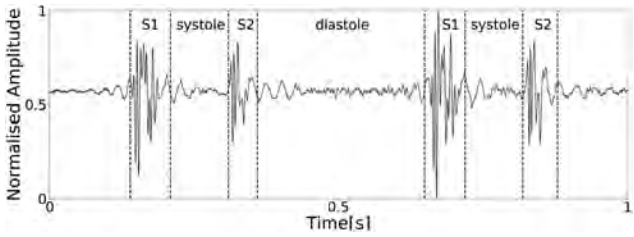

Fig. 1. Example of a recorded PCG signal and four states of the heart cycle (S1, systole, S2, diastole).

of mitral (or tricuspid) valve prolapse, even though echocardiograms may fail to confirm this finding [1].

In earlier works, segmentation of heart sounds have leveraged features from both the time and frequency domain [4]–[8], including: Shannon energy [9], wavelet envelope [10], Hilbert transform [11], time-frequency transform [12]–[14], and cepstral coefficients [15]–[17], etc. These features have been used both directly (on a sliding window) or to generate observable sequences from heart sounds for probabilistic sequence models, like hidden Markov models (HMMs) and their variations [18]–[20]. Although HMM-based methods have the advantage of modeling the sequential and periodic nature of heart sounds, this can result in false positives when noise or artifacts occur with similar features to the heart sounds. To mitigate this problem, Gill *et al.* [21] proposed the incorporation of timing durations within HMM for heart sound segmentation. Schmidt *et al.* [22] were the first to explicitly model the expected duration of heart sounds within the HMM framework using a hidden semi-Markov model (HSMM). Springer *et al.* [18] extended this work by modifying the Viterbi algorithm to include the duration densities and adding a logistic regression emission layer. This logistic regression-HSMM-based (LR-HSMM) method was evaluated on 10,172 seconds of heart sounds collected from 112 (healthy and pathological) subjects (with simultaneous electrocardiogram (ECG) as a gold standard) and demonstrated an average F_1 score of 98.5% for segmenting S1 and 97.2% for segmenting S2 [23]. This method was adopted as the reference segmentation method in the 2016 PhysioNet/Computet in Cardiology (CinC) Challenge for the classification of normal and abnormal heart sounds [24].

Nevertheless, the dependence of prior knowledge of state duration makes the method prone to false negatives during arrhythmia, particularly for tachycardia and bradycardia. As reported by the authors in the assessment of the 2016 PhysioNet/CinC Challenge [24], Not all researchers adopted Springer's segmentation method as the first step in the required classification task [25]–[28]. Whereas, the accuracy of anomaly recognition was not distinguished by this. It is indicated that segmentation does not necessarily result in an obvious improvement in classification. And further development of the segmentation algorithm may result in superior performance.

Recently, various approaches for reducing the explicit restrictions on state duration in heart sound segmentation have been proposed. Messner *et al.* [29] suggested an event detection approach using bidirectional gated recurrent neural network (BiGRNN) and achieved a similar performance compared to Springer's method. Renna *et al.* [30] utilized 1D U-Net [31] as transformation for HMMs and HSMMs. Meanwhile, with the progress in convolution neural network (CNN) for temporal data [32], [33], it seems obvious to apply these techniques in this context. However, such deep-learning-based (DL-based) approaches are known to overfit on the differences in noise levels between databases, due to recording conditions and device heterogeneity.

In this work, we propose an algorithm that combines both automated feature learning and sequential modeling. In order to eliminate the instability of the segmentation on pathological and noisy PCG signals, the proposed method disuses prior knowledge of heart sound state duration. The main contributions of this paper are:

1) Designing an adaptive Wiener filter to reduce the variabilities of the characteristics from different stethoscopes on heart sounds.
2) Developing an adaptive learning method to detect the four states of heart sounds, including building a temporal-framing adaptive network (TFAN) for the frame-level recognition, and designing state transition loss and dynamic inference.
3) Testing and comparing the proposed method with two state-of-the-art methods, the LR-HSMM method [18] and the BiGRNN-based method [29], over the whole database from PhysioNet/CinC Challenge 2016 and data sets with different segmentation difficulties.

TABLE I
STATISTICS OF RE-ANNOTATED DATABASES AFTER EXCLUDING UNSURE RECORDINGS

Database	Total Recordings	Total Beats	Recordings Used	Beats Used
Training-A	392	14,560	392	14,560
Training-B	368	3,353	368	3,353
Training-C	27	1,801	27	1,801
Training-D	52	853	52	853
Training-E*	1,927	59,637	500	15,256
Training-F	108	4,452	108	4,452
Test-B	206	1,247	206	1,247
Test-C	15	1,007	15	1,007
Test-D	24	268	24	268
Test-E*	882	25,261	200	6,060
Test-G	174	2,116	174	2,116
Test-I	33	1,207	33	1,207
Total	4,208	115,762	2,099	52,180

in Training-E and Test-E* indicates that part of original Training-E and Test-E were utilized in this work.

II. DATABASE

A. General Introduction

The 2016 PhysioNet/CinC Challenge [34] contains 12 independent data sets collected by different research teams. These were used to develop and test the proposed method (see Table I). Among them, Training-A is the only database which contains simultaneously recorded PCGs and ECGs. The other 11 sub

data sets only contains PCGs, namely Training-B˜E* and F, and Test-B˜E*, G and I.

The state labels of data sets were assigned as onsets of S1, systole, S2 and diastole. S1 occurs immediately after R-peaks (ventricular depolarization) of the ECG, while S2 occurs at approximately at the end-T-waves of the ECG (the end of ventricular depolarization) [18], [35]. Therefore, for PCGs with synchronous ECGs, the automated-detected R peaks and end-T-waves were the basis of annotations of S1 and S2 onsets. The segmented results solved by LR-HSMM were utilized as automatic marks for annotation reference in recordings containing only PCGs. The incorrect annotations happen in Training-A when R peaks and end-T-wave positions of an abnormal ECG period were misdetected. For data sets besides Training-A, although the annotations provided for the challenge were manually corrected by the organizers, some of them were still questionable and required re-annotations. These annotations were hand-corrected by visual and audible inspection of PCG waveforms. The re-annotation instances in Training-A are illustrated with reference ECG in Fig. 2.

Since the majority of heart sounds in Training-E (N = 4,074) and Test-E (N = 901) are normal, a total of 500 and 200 recordings were randomly extracted from Training-E and Test-E respectively to alleviate the re-annotation work while ensuring the accuracy of the evaluation.

B. Derivative Date Set

In order to further excavate characteristics of each data set, several indicators were designed as follows:

$$W = \frac{1}{N}\sum_{n=1}^{N} y^2[n], \tag{1}$$

$$F_{\text{S2}} = \frac{W_{S2}}{W_{diastole}}, \tag{2}$$

$$D_{\text{noise\&murmur}} = \frac{W_{\text{systole}} + W_{\text{diastole}}}{W_{\text{S1}} + W_{\text{S2}}}, \tag{3}$$

$$D_{\text{rhythm}} = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M-1}(ss[i] - E\{ss\})^2}, \tag{4}$$

$$D_{\text{rate}} = \left\| \max(1.2, E\{ss\}) + \min(0.6, E\{ss\}) - 1.8 \right\|, \tag{5}$$

where W is the average power of the heart sound $y[n]$ with N points and M heart beats. F_{S2} defines S2 sound articulation as the ratio of S2 sound's energy to diastolic energy. $D_{\text{noise\&murmur}}$ is the ratio of power between systole/diastole and S1/S2, which represents the index of murmur severity and signal quality. D_{rhythm} is estimated by the summation of the standard deviation of S1 onset intervals (ss) to measure the severity of arrhythmia. D_{rate} is relative difference of $E\{ss\}$ from normal range, which reflects the degree of abnormality in heart rate (bradycardia and tachycardia). In Equation (5), 1.2 s and 0.6 s are average S1 onset intervals of 50 and 100 heart beats per minute, respectively.

According to the statistical results of the indicators, for normal heart sounds, $D_{\text{noise\&murmur}}$ is below 0.3 and F_{S2} is larger than 2.0. And for the cases contaminated by sever murmurs or noise, $D_{\text{noise\&murmur}}$ is always over 0.8. Therefore, the noise&murmur level can be divided into low ($D_{\text{noise\&murmur}} \leq 0.8$) and high ($D_{\text{noise\&murmur}} > 0.8$). During arrhythmia, D_{rhythm} is over 0.12, which is also the indicated value for PP interval deviation of arrhythmic ECG. D_{rate} would be greater than 0 when the heart rate is abnormal.

Three derivative data sets were constructed according to the designed indicators. They were named LEVEL-I, LEVEL-II and LEVEL-III, corresponding to easy, medium and difficult in terms of both automatic and manual heart sound segmentation. The threshold of $D_{\text{noise\&murmur}}$ is set to 0.8 to distinguish heart sounds with complicated severe noise and murmur. $D_{\text{rhythm}} + D_{\text{rate}}$ is an indicator of abnormal heart rhythm and heart rate. Thus, the threshold of $D_{\text{rhythm}} + D_{\text{rate}}$ is assigned a value of 0.2 s. Fig. 3 provides a graphical illustration of the above partition rules.

All of the heart sound recordings chosen for different levels were segmented into multiple 10 second files. The resultant numbers of recordings and beats are summarized in Table II. The specific instances in the three difficulty levels are displayed in Fig. 4.

We counted the proportion of various anomalies in each data set based on the indicators, including heart sounds with high-level noise&murmur, arrhythmic heart sounds, heart sounds with abnormal heart rate and heart sounds with vague S2. It was found that Training-B and Test-B have the lowest signal quality among the data sets and all of the data sets contain a certain amount of heart sounds with arrhythmia and abnormal heart rate expect for Training-A. The specific data is shown in Table III.

III. Methods

The proposed method involves two main parts: a signal pre-processing routine, and the TFAN segmentation. The signal pre-processing employed three filters: an adaptive Wiener filter, a bandpass filter and a wavelet filter. The TFAN is an original network designed for heart sound segmentation with an encoder-decoder architecture. In order to learn the state transition information in PCG, the loss function of the TFAN was carefully designed.

A. Signal Pre-Processing

The segmentation algorithm used a combination of three filtered PCG signals as inputs. The three filters included an adaptive Wiener filter, a bandpass filter and a wavelet filter. As shown by the instances reported in Fig. 5, the adaptive Wiener filter was designed to suppress the in-band noise, especially reduce the impact of tail sounds in systole and diastole. This approach increased the amplitude resolution of alternate segments between heart sound states. The bandpass filter and the wavelet filter were applied to enhance S1 and S2 sounds and provided complementary information of their waveform features.

1) Adaptive Wiener Filter: Consider a zero-mean clean heart sound signal $x(n)$ contaminated by noise $v(n)$ (uncorrelated with $x(n)$), so that the noisy heart sound $y(n)$ at the

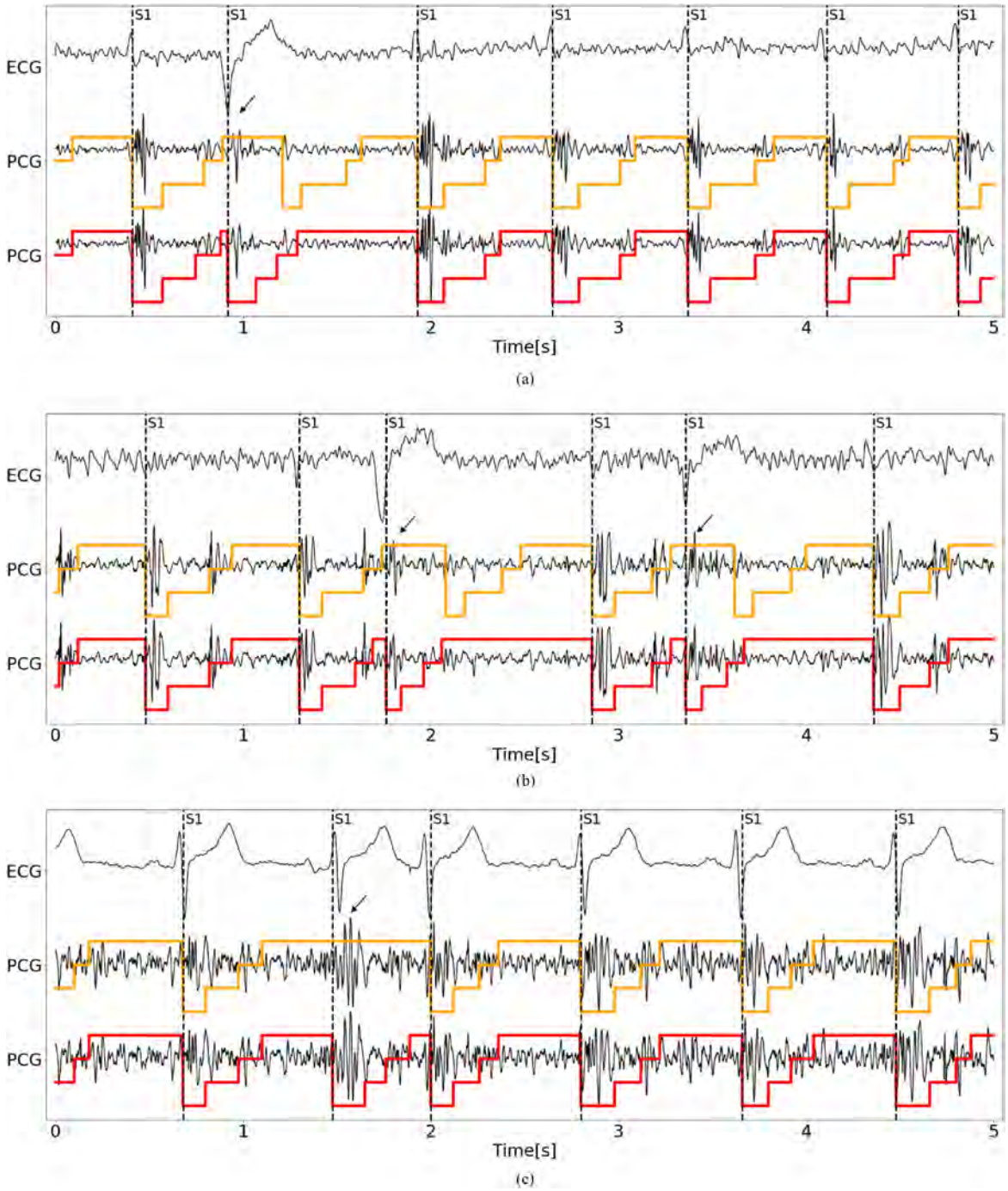

Fig. 2. Illustrations of the ECG and simultaneous PCG with automatically derived states in Training-A. Sub-figures (a) and (b) are from patients with premature ventricular contractions (PVCs), and sub-figure (c) is from a patient with premature atrial contractions (PACs). Arrhythmia-like PACs or PVCs always induce the false annotations in PCG signals (middle waveform in yellow) and the arrows point out the mistakes in the original annotation. The corrected annotations are shown in each sub-figure as a repeated waveform and a red staircase plot overlaid. Each level in the staircase plot represents S1, systole, S2 and diastole (in ascending value).

discrete time n is

$$y(n) = x(n) + v(n), n = 0, \ldots, N-1, \tag{6}$$

The estimation of the error signal $e_x(n)$ between the clean heart sound at the discrete time n is given by

$$e_x(n) = x(n) - \hat{x}(n) = x(n) - \boldsymbol{h^T y}(n), \tag{7}$$

where superscript ^{T} denotes transpose of a vector or a matrix,

$$\boldsymbol{h} = [h_0, h_1, \ldots, h_(L-1)]^T$$

is an finite impulse response (FIR) filter of length L, and

$$\boldsymbol{y(n)} = [y(L-1), y(L-2), \ldots, y(0)]^T$$

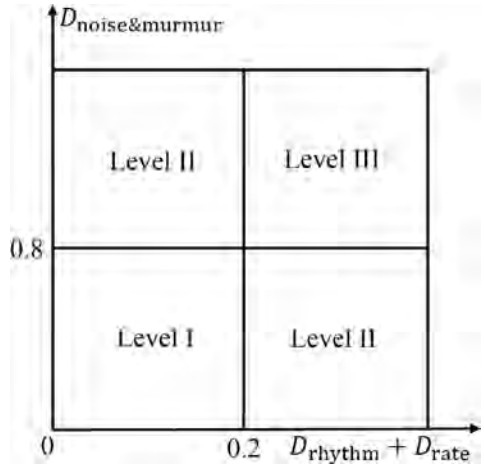

Fig. 3. The partition rules of the three difficulty levels (LEVEL-I, LEVEL-II and LEVEL-III) based on the extent of noise/murmur and evaluation of heart rhythm/rate in heart sounds.

TABLE II
SUMMARY OF DATABASES CORRESPONDING TO THREE DIFFICULTY LEVELS

Database	Recordings	Beats
LEVEL-I	200	2,296
LEVEL-II	200	2,459
LEVEL-III	150	1,906
Total	550	6,661

is a vector of window from observation signal y(n) containing L samples.

Assuming the optimal estimate of the clean heart sound $x(n)$ is $\hat{x}_o(n)$, the optimal filter $\boldsymbol{h_o}$ for $\hat{x}_o(n)$ is the Wiener filter which is obtained by

$$\boldsymbol{h_o} = \arg\min_{\boldsymbol{h}} E\{e_x^2(n)\}. \tag{8}$$

According to Wiener-Hopf equation, we have

$$\boldsymbol{R_y h_o} = E\{y(n)x(n)\} = \boldsymbol{r_y} - \boldsymbol{r_v}, \tag{9}$$

where $\boldsymbol{R_y}$ is the correlation matrix of the observed signal $y(n)$. $\boldsymbol{r_y}$ and $\boldsymbol{r_v}$ are the correlation vectors, which are also the first columns of $\boldsymbol{R_y}$ and the correlation matrix of the noise $\boldsymbol{R_v}$ respectively.

Now $\boldsymbol{h_o}$ can be inferred as

$$\boldsymbol{h_o} = \boldsymbol{u_1} - \boldsymbol{R_y^{-1} r_v}, \tag{10}$$

where $\boldsymbol{u_1} = [1, 0, \ldots, 0]^T$.

Assuming that the additive noise is white over a very short time duration in comparison to the heart sounds, we have

$$\boldsymbol{r_v} = \sigma_v^2 \boldsymbol{u_1}, \tag{11}$$

and

$$\begin{aligned} \boldsymbol{h_o} &= \boldsymbol{u_1} - \sigma_v^2 \boldsymbol{R_y^{-1} u_1} \\ &= \left[1 - \frac{\sigma_v^2}{\boldsymbol{R_y}[0]}, 1 - \frac{\sigma_v^2}{\boldsymbol{R_y}[1]}, \ldots, 1 - \frac{\sigma_v^2}{\boldsymbol{R_y}[L-1]}\right], \end{aligned} \tag{12}$$

where $\sigma_v^2 = E\{v^2(n)\}$.

Because the noise $v(n)$ is not directly observable, σ_v^2 is ideally calculated while there is no heart sound signal. In order to avoid being disturbed by sudden changes in the recording, the local window is segmented in fixed length and σ_v^2 is estimated as the lower quartile of the local variances $Q1(\boldsymbol{lvar})$ for all segments. Finally the estimated heart sound of the local window $\hat{x}(n)$ is given by

$$\begin{aligned} \hat{x}(n) &= \boldsymbol{h_o}(n)^T \boldsymbol{y}(n) \\ &= \left(1 - \frac{Q1(\boldsymbol{lvar})}{\boldsymbol{R_y}[n]}\right) \boldsymbol{y}(n), n = 0, 1, \ldots, L. \end{aligned} \tag{13}$$

2) Bandpass Filter: The majority of the frequency content in S1 and S2 sounds is below 150 Hz, usually with a peak around 50 Hz [36]. Thus, a Bandpass filter was applied to create a signal with 30–60 Hz pass-band, to be used as one input channel to provide the potential optimal positions of S1 and S2.

3) Wavelet Filter: The first step in the wavelet filter for heart sounds is a discrete time wavelet transform (DWT). Following Springer *et al.* [18], the reverse biorthogonal wavelet with three vanishing moments for the decomposition (analysis) wavelet and nine vanishing moments for the reconstruction (synthesis) wavelet ($'rbio3.9'$) was used. In order to remove the insignificant noise, the detail coefficients below an adaptive threshold at some scales were set to zero. The threshold was set to be the median energy, which was estimated by averaging the absolute coefficients at different scales. The final filtered heart sounds were reconstructed by the inverse DWT.

B. Temporal-Framing Adaptive Network

1) Model Architecture: The TFAN was designed with an encoder-decoder architecture. The encoder (Fig. 6) is a transformer module for the purpose of mapping the original signals into a feature space. The decoder (Fig. 6) is designed to segment the output feature mapped by the encoder into four states (S1, systole, S2, diastole). Within the network, a framing module is deployed between the encoder and the decoder.

Before loaded to the TFAN model, each processed heart sound recording was sliced into segments of two seconds and resampled to 1,000 Hz. The input data is denoted as $x(n)$ for $n = 0, \ldots, 1, 999$ and $x(n) \in \mathbb{R}^3$.

2) Encoder: A residual convolution block is used as a basic unit for feature mapping (Fig. 6). The residual block [37] contains a branch leading out to a series of transformations F, whose outputs are added to the input x of the block, so the original mapping is recast into $x + F(x)$. This effectively allows layers to learn modifications to the identity mapping, rather than the entire transformation, which is more advantageous for identifying similar states in the heart sound (e.g., S1 and S2).

In the TFAN, instance normalization (IN) and dilated convolution were utilized in each residual block. The reasons for using IN are: 1) The segmentation model is trained with limited batch size and IN normalizes across each training sample instead of the mini-batch, therefore biased estimations of mean and variance of mapped features are avoided; 2) IN normalizes across each channel, so the independence of each channel is maintained. For the input data $x \in R^{N \times T \times C}$, IN calculates the mean and

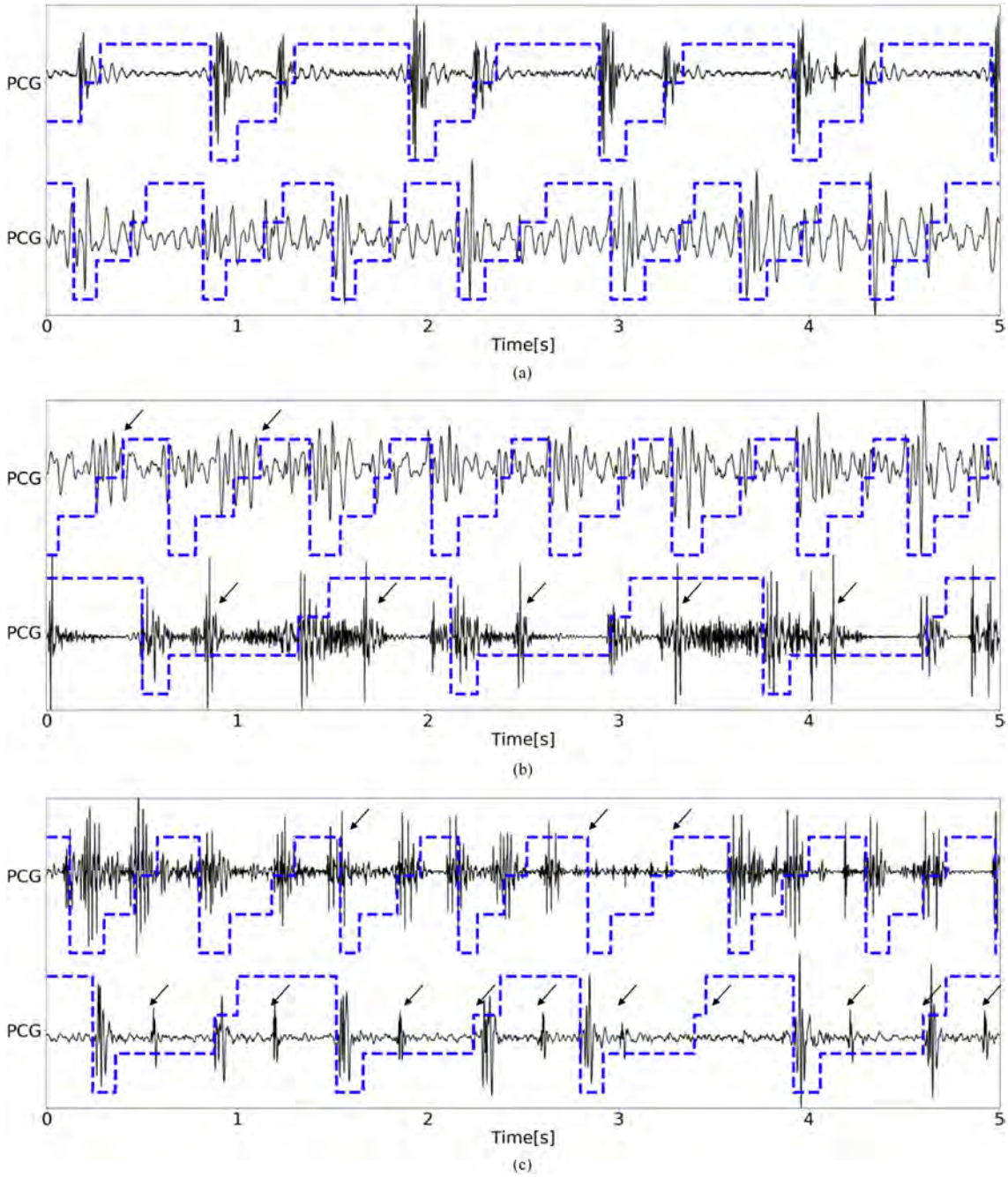

Fig. 4. Sub-figures (a), (b) and (c) correspond to three instances from LEVEL-I, LEVEL-II and LEVEL-III respectively. The blue dashed line indicates the states assigned by the LR-HSMM method and the arrows identify the unsuccessful segmentation. Note that the increased difficulty significantly impacts the performance of the LR-HSMM method.

variance across the time dimension of each sample and retains the dimensions of the batch N and channel C as

$$\mu_{nc}(x) = \frac{1}{T}\sum_{t=1}^{T} x_n ct, \tag{14}$$

$$\sigma_{nc}(s) = \sqrt{\frac{1}{T}\sum\nolimits_{t=1}^{T}(x_{nct} - \mu_{nc}(x))^2 + \varepsilon}, \tag{15}$$

where ε is the biased value to avoid division by 0 when normalizing the weights.

Dilated convolution can enlarge the receptive field of convolution layers and preserve the size of feature maps without loss of resolution. This is critical for the subsequent framing module and decoder. Meanwhile, bidirectional padding is chosen as the padding strategy in dilated convolution. The padding length is decided by the convolutional kernel size and dilation rate. Fig. 7 illustrates the padding method for different dilation rates (d) in the case of a convolution kernel size (ks) of 3.

3) Decoder: Before passed through the Decoder, the feature map of the heart sound produced by Encoder is framed by a fixed

TABLE III
SUMMARY OF THE CHARACTERISTICS IN EACH DATA SET. THE NUMBER OF RECORDINGS AND PROPORTION IS REPORTED. THE CHARACTERISTICS INCLUDE NOISE&MURMUR, ARRHYTHMIA, ABNORMAL HEART RATE(ABN. HR) AND VAGUE S2

Database	Recordings	High-level Noise&Murmur		Arrhythmia		Abn. HR		Vague S2	
		Count	Prop.(%)	Count	Prop.(%)	Count	Prop.(%)	Count	Prop.(%)
Training-A	392	3	0.77	25	6.38	21	5.36	32	8.16
Training-B	368	184	50.0	4	1.09	30	8.15	274	74.47
Training-C	27	2	7.41	4	14.85	2	7.41	6	22.22
Training-D	52	1	1.92	8	15.38	11	21.15	5	9.62
Training-E*	500	24	4.80	13	2.60	95	19.00	51	10.20
Training-F	108	1	0.93	22	20.37	11	10.19	0	0.00
Test-B	206	72	34.95	6	2.91	22	10.68	118	57.28
Test-C	15	2	13.33	2	13.33	1	6.67	2	13.33
Test-D	24	0	0.00	3	12.50	3	12.50	1	4.17
Test-E*	200	1	6.67	5	2.50	39	19.50	25	12.50
Test-G	174	3	1.72	11	6.32	11	6.32	23	13.22
Test-I	33	6	18.18	0	0.00	1	3.03	14	42.42
Level-I	200	0	0.00	0	0.00	13	6.50	23	11.50
Level-II	200	33	16.50	27	13.50	14	7.00	49	24.50
Level-III	150	105	70.00	45	30.00	14	9.33	19	12.67

in Training-E and Test-E* indicates that part of original Training-E and Test-E were utilized in this work.

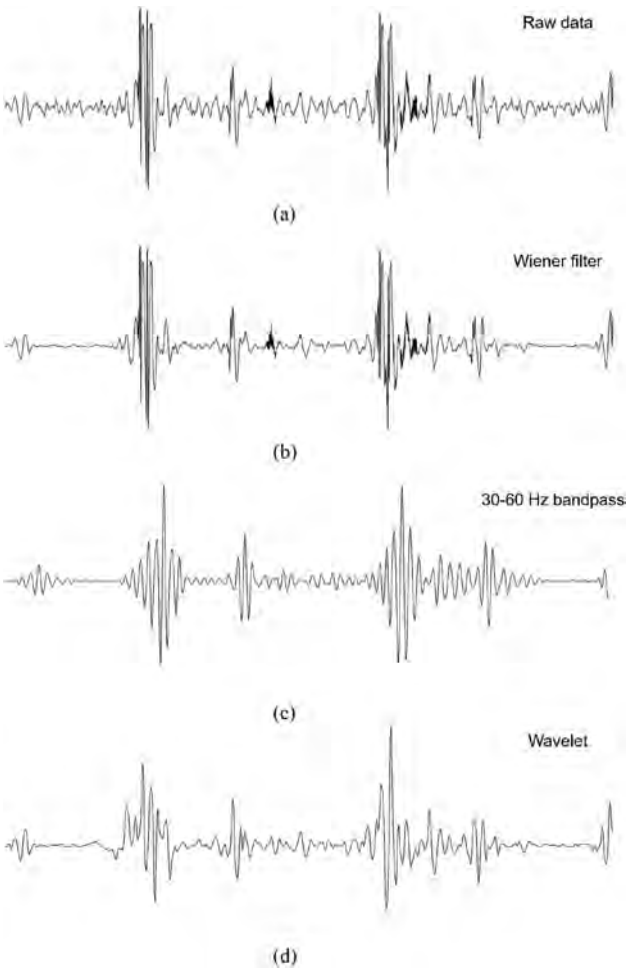

Fig. 5. Sub-figures (a), (b), (c) and (d) demonstrate the examples of raw PCG signal, the signal processed by an adaptive Wiener filter, the signal processed by a 30–60 Hz bandpass filter and the signal processed by a reverse biorthogonal wavelet filter. The output of each filter constitute the three channels of input data for temporal framing network.

length τ. Then the frame-level features can be further mapped by 2D convolution blocks (Fig. 6) of decoder. The output is then fed to a bidirectional long short-term memory (Bi-LSTM) layer to learn sequential characteristics of the frame-level features.

Assuming the mappings of the encoder and decoder are denoted as $f = \mathbb{F}(x(n))$ and $g = \mathbb{G}(f)$, the feature map is transformed as below after the frame-level decoder:

$$f \to [f_0, f_1, \ldots, f_{\frac{N}{\tau}}], \tag{16}$$

$$\hat{s}(m) = [g(f_0), g(f_1), \ldots, g(f_{\frac{N}{\tau}})], m = 0, \ldots, \frac{N}{\tau}. \tag{17}$$

$\hat{s}(m)$ is defined as the sequence of the logits output from the model, where $s(m)$ is the ground truth of the heart sound states.

4) Loss Function: According to the periodic nature of heart sounds, the identification of state for each frame is determined not only based on the features but the state transition information between the current and preceding frames. Therefore, the loss in the TFAN is the combination of the classification loss and the state transition loss between the current frame and the previous frame:

$$L(y, \hat{y}) = -\frac{1}{T}\frac{1}{N}\sum_{\tau=1}^{T}\sum_{i=1}^{N}\Big\{C_1 \times y_{i\tau}\log\hat{y}_{i\tau} + C_2 \times \frac{y_{i\tau} + y_{i(\tau-1)}}{2}\log\frac{\hat{y}_{i\tau} + \hat{y}_{i(\tau-1)}}{2}\Big\}, \tag{18}$$

where y and $\hat{y}$ represent the annotated state and the predicted mask of the frame, respectively. T and N are the number of frames and the number of heart sound states, separately, which $T = 100$ and $N = 4$ in the TFAN-based method. y_{i0} and $\hat{y}_{i0}$ are padded as the ground truth and the predicted logit of the first frame. The weighting parameters C_1 and C_2 could help adjust the constraint degree of the state transition information and the features of each frame in state prediction. In the TFAN-based method, $C_1 = 1$ and $C_2 = 2$.

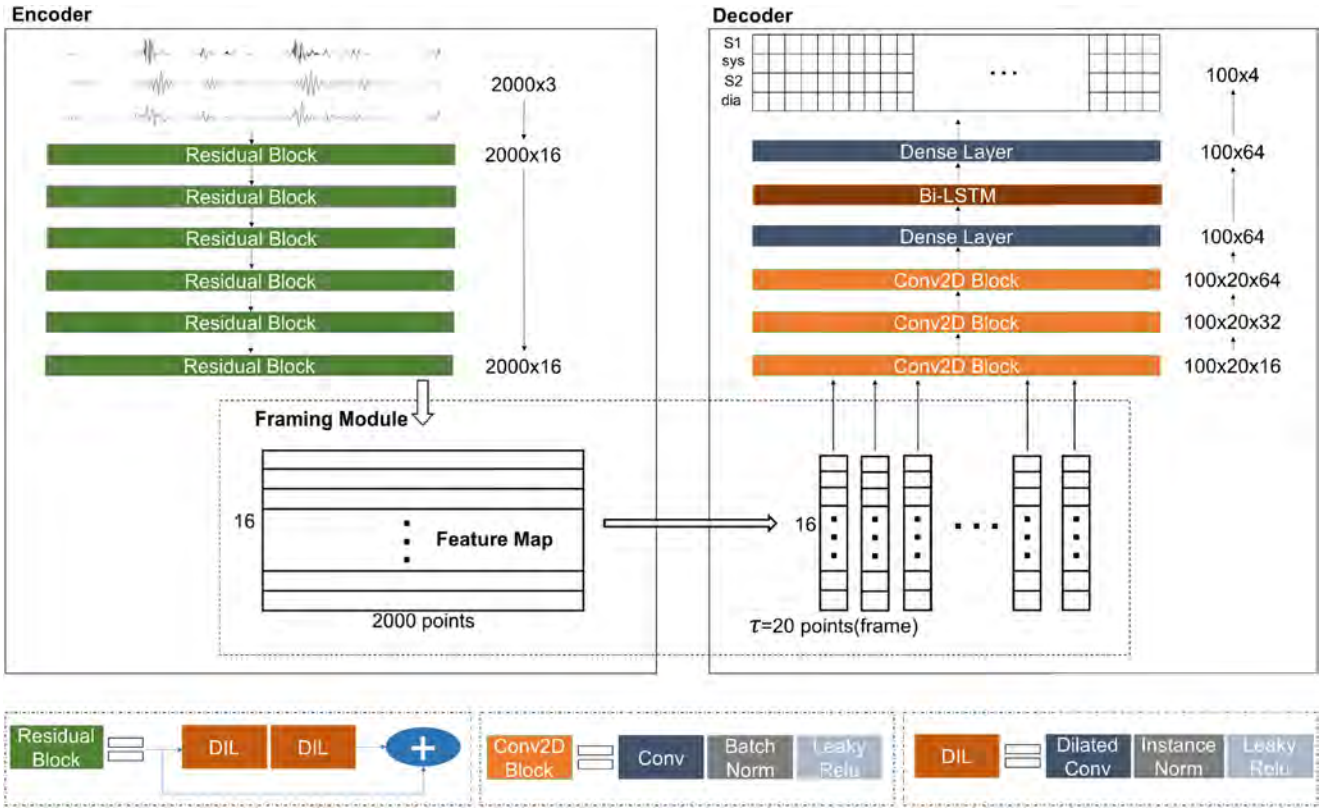

Fig. 6. The architecture of the proposed TFAN. The PCG signal is framed into multi-frames after feature mapping in the Encoder. Then the whole frames in one batch generate a new batch of features to be input into the decoder. The final outputs are the predicted logits of four states in each frame.

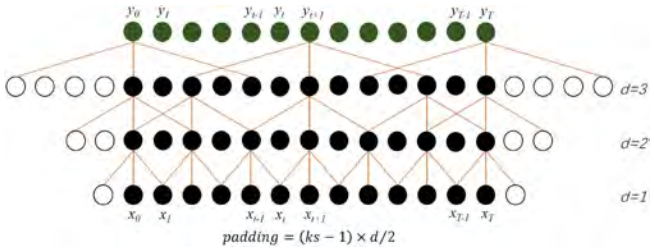

Fig. 7. The padding strategy of the convolution operation in temporal residual block. The purpose using the above padding pattern is to maintain the dimension while feature mapping.

5) Dynamic Inference: Since the length of the input data of our model is fixed, the heart sound recording needs to be divided into segments for dynamic inference. In order to minimize the impact on segmentation of data around slicing boundary, 50% overlapping windows are adopted. For the overlapping windows, the logits of different states are simply averaged. If the length of the remaining recorded data is less than the 50% overlapping duration, the input segment of fixed length is taken before the last point. Meanwhile, the logits of the remaining data are retained and concatenated with the previous results so that all points of the recording can be detected.

Knowning $s(t) \in \{0, 1, 2, 3\}$, each element in $s(t)$ corresponds to S1, systole, S2, diastole respectively. The labels are then one-hot encoded. The outputs from TFAN are the logits of four states for each frame. Assuming the input data is F and the length of state sequence after framing is M, the inference step needs to find out $s(m)$ by $\arg\max_s P(s_1, s_2, \ldots, s_M|F)$. Since the total search of $s_{1\sim M}$ for the best state sequence required 4^M times, the search time complexity would be high when M is large. The Viterbi algorithm is therefore adopted to shorten the solving time. Based on the Viterbi algorithm, the inference method can be transformed to

$$\max P(s_1, \ldots, s_M|F) = \max\{q(v, M)|v\}, \tag{19}$$

where

$$q(v = j, m) = \max\{q(v = i, m-1) \times a(i, j, m)|i\}, \tag{20}$$

for $v = 0, 1, 2, 3$ and $i = 0, 1, 2, 3$. Note that $q(v, M)$ represents the maximum probability for the state sequence ending with v, and $a(i, j, m)$ defines the transforming probability from state i at step $m-1$ to state j at step m. For heart sound states, the state transition probability matrix is given by

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{14} \\ a_{21} & a_{22} & \cdots & a_{24} \\ \vdots & \vdots & \vdots & \vdots \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0 & 0.5 \end{bmatrix}. \tag{21}$$

The predicted state sequence s is inferred by the following function

$$s_m = \arg\max_v q(v, m). \tag{22}$$

IV. Experiments

The proposed segmentation methods were compared with two methods appeared in the literature. Namely, the BiGRNN-based method using spectrogram and envelop features described in [29] and the LR-HSMM method, which is currently considered as the state-of-art PCG segmentation method. For fairly comparing the performance of TFAN and BiGRNN, the proposed dynamic inference approach was conducted in both methods.

Besides, as a generative model, LR-HSMM is essentially trained to explicitly model the probability distribution of each heart sound state given four extracted features on each time step. Therefore, the LR-HSMM method is not sensitive to the size of training set. In the previous study, LR-HSMM is trained by recordings from only 60 patients [18]. Instead, for DL-based methods, all-round data sets are generally required for precisely learning data distribution. Such as [29], the training set for BiGRNN consisted of recordings randomly selected from Training-B˜F of the challenge. In order to improve the extensiveness of DL-based methods, we introduced the framing module in the proposed TFAN, converting estimation of state sequence to estimation of state on each time step. For finding out the influence of various feature learning approaches on segmentation and the capacity of the proposed method in few-shot learning, we limited the number of training recordings.

The experiments comparing the performances of the three methods were conducted in two scenarios. The first scenario was to test the methods on Training-A* and other independent sub data sets from 2016 PhysioNet/CinC Challenge. The second scenario was to test on the data sets of three difficulty levels (LEVEL-I, LEVEL-II and LEVEL-III).

A. Training Setup

Since the gold-standard reference positions of onsets of S1 and S2 sounds were derived from the synchronous ECGs [18], to ensure the preciseness of the experiment, the training set was consisted by heart sound recordings with synchronous ECGs splitted from Training-A. The ultimate size of training set was restricted to 50 recordings for all the methods, and the remaining recordings were utilized as Training-A* for testing.

Five-fold cross-validation was adopted as the training approach. For ensuring the recording used for validation is not used to train, 50 heart sound recordings were split into 40 recordings for training and 10 recordings for validation in each fold at first. Then the heart sounds in training and validation set were pre-processed and sliced into 2 s segments for neural network training. After five-fold cross-validation training, the model with best performance on the validation fold was chosen.

The loss functions of BiGRNN and TFAN were unified into the proposed one. Weights of both models were updated by the Nesterov Momentum optimizer with factor of 0.9 and learning rate of 0.001. In order to avoid overfitting, the following early stop strategy was adopted. When the model failed to achieve the best validation accuracy in 20 consecutive epochs, the training is terminated.

B. Evaluation Metrics

To evaluate the performance of the TFAN-based method against the LR-HSMM method, three measurements are considered, which are defined as:

$$SE = \frac{TP}{TP + FN} \times 100\%, \tag{23}$$

$$P_+ = \frac{TP}{TP + FP} \times 100\%, \tag{24}$$

$$F_1 = \frac{2 \times SE \times P_+}{SE + P_+} \times 100\%, \tag{25}$$

where TP (true positive), FP (false positive) and FN (false negative) are determined by the following rules [23]:

Let $y = y_0, y_1, \ldots, y_i, \ldots y_N$ denotes the manually annotated onset positions for one of the four heart sound states while $\hat{y}$ represents the state onsets based on the predicted states $\hat{s}$. Assuming the tolerance parameter is σ, the predicted segmented onset is expected to appear in the time region $y_i - \sigma \leq \hat{y_i} < y_i + \sigma$ and should not in the time interval $y_i + \sigma \leq \hat{y_i} < y_{i+1} - \sigma$. N_1 and N_2 would denote the counted numbers of the predicted start positions within the two time intervals. Therefore, a successful prediction happens when $N_1 = 1$ and $N_2 = 0$. The TP, FP and FN are then counted as:

$$TP = TP + 1, \qquad \text{if } N_1 > 0, \tag{26}$$

$$FP = \begin{cases} FP + N_1 - 1, & \text{if } N_1 > 1, \\ FP + N_2, & \text{if } N_2 > 0, \end{cases}$$

$$FN = FN + 1, \qquad \text{if } N_1 = 0. \tag{27}$$

The tolerance parameter σ was set to 100 (ms) to evaluate different heart sound segmentation methods [18]. The tolerance is based on the ECG R-peak detection tolerance of 150 (ms) [38], which, as is approximately the length of the fundamental heart sounds, is shortened to 100 (ms).

Significance testing was performed using a two-sided paired t-test on the F_1 scores from LEVEL-I, LEVEL-II and LEVEL-III.

V. Results

The gross performance results of the LR-HSMM method, the BiGRNN-based method and the TFAN-based method on all the test sets were presented in Table IV. The TFAN-based method was tested with and without the adaptive Wiener filter. Table IV illustrates the performance for the combined four states (S1, systole, S2 and diastole), as well as the F_1 scores for each state separately to give an indication of performances on different states.

The average performance of the TFAN-based method, the BiGRNN-based method and the LR-HSMM method on LEVEL-I, LEVEL-II and LEVEL-III were reported in Table V. These gross scores were calculated on a per recording basis,

TABLE IV

TOTAL EXPERIMENTAL RESULTS (%) OF THE LR-HSMM METHOD, THE BIGRNN-BASED METHOD, THE TFAN-BASED METHOD WITHOUT THE ADAPTIVE WIENER FILTER AND THE TFAN-BASED METHOD WITH THE ADAPTIVE WIENER FILTER (PROPOSED) ON ALL OF THE DATA SETS FROM THE 2016 PHYSIONET/CINC CHALLENGE. THE METRICS WERE CALCULATED FROM THE TOTAL NUMBER OF TP, FP AND FN IN EACH DATABASE

Database	Method	F_1 measurement for each state				Overall evaluation metrics		
		F_1^{S1}	F_1^{sys}	F_1^{S2}	F_1^{dia}	Se	P_+	F_1
Training-A*	LR-HSMM	**97.56**	97.42	96.45	95.59	97.37	97.10	96.75
	BiGRNN	97.04	97.40	97.21	96.14	97.13	96.50	97.06
	TFAN	97.36	97.58	97.05	96.40	97.42	96.78	97.10
	TFAN+Adaptive Wiener Filter (proposed)	97.35	**97.69**	**97.26**	**96.54**	**97.49**	**96.94**	**97.21**
Training-B	LR-HSMM	99.43	99.47	98.64	98.55	99.37	98.68	99.02
	BiGRNN	93.62	93.90	91.45	91.69	93.25	92.07	92.66
	TFAN	98.81	99.16	98.55	98.56	99.16	98.38	98.77
	TFAN+Adaptive Wiener Filter (proposed)	**99.65**	**99.61**	**99.29**	**99.09**	**99.41**	**99.41**	**99.41**
Training-C	LR-HSMM	93.84	91.84	87.14	85.57	88.25	90.97	89.59
	BiGRNN	95.47	94.44	88.33	88.01	90.91	92.20	91.55
	TFAN	98.19	96.43	**92.38**	**91.84**	**94.09**	**95.33**	**94.71**
	TFAN+Adaptive Wiener Filter (proposed)	**98.27**	**96.48**	91.21	90.71	93.32	95.03	94.16
Training-D	LR-HSMM	96.04	96.04	93.94	91.69	93.21	95.67	94.43
	BiGRNN	96.88	96.67	96.97	95.02	96.19	96.06	96.35
	TFAN	96.37	96.66	96.80	95.11	96.11	96.36	96.23
	TFAN+Adaptive Wiener Filter (proposed)	**96.97**	**97.66**	**97.04**	**95.53**	**96.90**	**96.70**	**96.80**
Training-E*	LR-HSMM	98.32	98.19	96.69	96.04	96.23	98.41	97.31
	BiGRNN	96.15	96.24	92.56	92.53	93.26	95.50	94.37
	TFAN	97.50	97.51	95.87	94.75	95.48	97.35	96.41
	TFAN+Adaptive Wiener Filter (proposed)	**98.37**	**98.60**	**97.49**	**96.59**	**97.10**	**98.43**	**97.76**
Training-F	LR-HSMM	90.19	90.51	87.89	87.45	87.61	90.45	89.01
	BiGRNN	86.49	86.80	86.31	86.37	87.64	85.38	86.50
	TFAN	90.69	90.63	89.64	88.56	90.61	89.16	89.88
	TFAN+Adaptive Wiener Filter (proposed)	**92.91**	**93.29**	**92.70**	**91.30**	**92.63**	**92.47**	**92.55**
Test-B	LR-HSMM	**97.31**	**97.59**	**95.44**	**94.70**	**96.60**	**95.90**	**96.25**
	BiGRNN	94.60	94.09	90.41	90.56	93.10	91.72	92.40
	TFAN	96.57	95.70	93.99	93.67	95.80	94.16	94.97
	TFAN+Adaptive Wiener Filter (proposed)	96.00	95.88	92.02	92.47	94.51	93.66	94.09
Test-C	LR-HSMM	95.60	95.60	85.06	83.69	88.24	91.76	89.97
	BiGRNN	97.92	97.68	95.18	94.00	95.46	96.92	96.19
	TFAN	98.15	97.56	95.18	93.76	95.43	96.89	96.16
	TFAN+Adaptive Wiener Filter (proposed)	**98.45**	**98.22**	**94.89**	**94.42**	**95.93**	**97.05**	**96.49**
Test-D	LR-HSMM	96.36	95.45	92.24	92.04	93.11	94.90	94.00
	BiGRNN	98.43	98.43	98.88	98.10	99.11	97.81	98.45
	TFAN	98.43	98.43	98.21	97.89	99.00	97.48	98.24
	TFAN+Adaptive Wiener Filter (proposed)	**98.65**	**99.10**	**98.88**	**98.94**	**99.44**	**98.35**	**98.90**
Test-E*	LR-HSMM	98.02	98.01	96.59	95.93	96.95	97.33	97.14
	BiGRNN	98.61	98.59	96.63	96.79	97.28	98.03	97.66
	TFAN	99.15	98.98	98.15	97.87	98.63	98.45	98.54
	TFAN+Adaptive Wiener Filter (proposed)	**99.22**	**99.24**	**98.11**	**98.03**	**98.63**	**98.67**	**98.65**
Test-G	LR-HSMM	**96.59**	**96.36**	93.90	93.64	94.15	**96.10**	95.12
	BiGRNN	91.93	92.83	92.56	92.57	92.80	92.15	92.47
	TFAN	94.29	94.22	93.77	93.94	94.71	93.41	94.06
	TFAN+Adaptive Wiener Filter (proposed)	96.22	96.13	**95.90**	**95.52**	**96.05**	95.84	**95.94**
Test-I	LR-HSMM	**99.61**	99.08	92.44	93.60	96.10	96.24	96.18
	BiGRNN	95.63	96.12	93.92	94.15	94.95	95.42	94.49
	TFAN	99.52	**99.61**	**98.08**	97.37	98.67	**98.61**	98.64
	TFAN+Adaptive Wiener Filter (proposed)	99.22	99.39	97.99	**98.19**	**98.85**	98.55	**98.70**
Global Average	LR-HSMM	96.57	96.30	93.04	92.37	93.93	95.29	94.56
	BiGRNN	95.24	95.27	93.37	93.00	94.26	94.15	94.18
	TFAN	97.09	96.87	95.64	94.98	96.26	96.03	96.14
	TFAN+Adaptive Wiener Filter (proposed)	**97.61**	**97.61**	**96.07**	**95.94**	**96.69**	**96.76**	**96.72**

$*$ in Training-A* indicates that the 50 recordings in training set were excluded from Training-A for testing.

$*$ in Training-E* and Test-E* indicates that part of original Training-E and Test-E were utilized for testing.

TABLE V
STATISTICAL RESULTS (%) OF THE LR-HSMM METHOD, THE BIGRNN-BASED METHOD AND THE TFAN-BASED METHOD AMONG ALL THE RECORDINGS IN LEVEL-I, LEVEL-II AND LEVEL-III. THE PERFORMANCE METRIC MEANS AND STANDARD ERRORS ARE COMPUTED OVER EACH RECORDING OF THE DATABASE RESPECTIVELY

Database	Method	F_1 measurement for each state				Overall evaluation metrics		
		F_1^{S1}	F_1^{sys}	F_1^{S2}	F_1^{dia}	Se	P_+	F_1
LEVEL-I	LR-HSMM	99.15±0.27	98.71±0.31	98.38±0.68	97.23±0.70	97.55±0.48	99.26±0.36	98.37±0.44
	BiGRNN	99.27±0.15	99.62±0.12	99.35±0.17	97.78±0.25	98.53±0.15	99.51±0.12	99.01±0.13
	TFAN+Adaptive Wiener Filter (proposed)	**99.43±0.12**	**99.78±0.09**	**99.60±0.11**	**98.04±0.21**	**98.73±0.13**	**99.71±0.09**	**99.21±0.10**
LEVEL-II	LR-HSMM	89.35±1.54	89.13±1.51	86.88±1.74	84.81±1.77	86.37±1.61	89.49±1.42	87.56±1.54
	BiGRNN	93.41±0.82	93.50±0.79	92.30±0.76	91.28±0.78	93.17±0.68	92.27±0.81	92.63±0.73
	TFAN+Adaptive Wiener Filter (proposed)	**94.88±0.74**	**95.39±0.70**	**94.03±0.76**	**92.40±0.80**	**94.39±0.66**	**94.06±0.77**	**94.17±0.71**
LEVEL-III	LR-HSMM	85.22±1.81	82.97±1.99	73.59±2.71	71.68±2.70	76.30±2.19	82.13±1.82	78.46±2.05
	BiGRNN	91.13±1.09	90.44±1.25	86.40±1.38	85.76±1.49	88.62±1.22	88.52±1.23	88.45±1.20
	TFAN+Adaptive Wiener Filter (proposed)	**94.73±0.77**	**93.27±1.05**	**89.32±1.29**	**87.83±1.43**	**90.64±1.03**	**92.12±0.99**	**91.31±1.00**

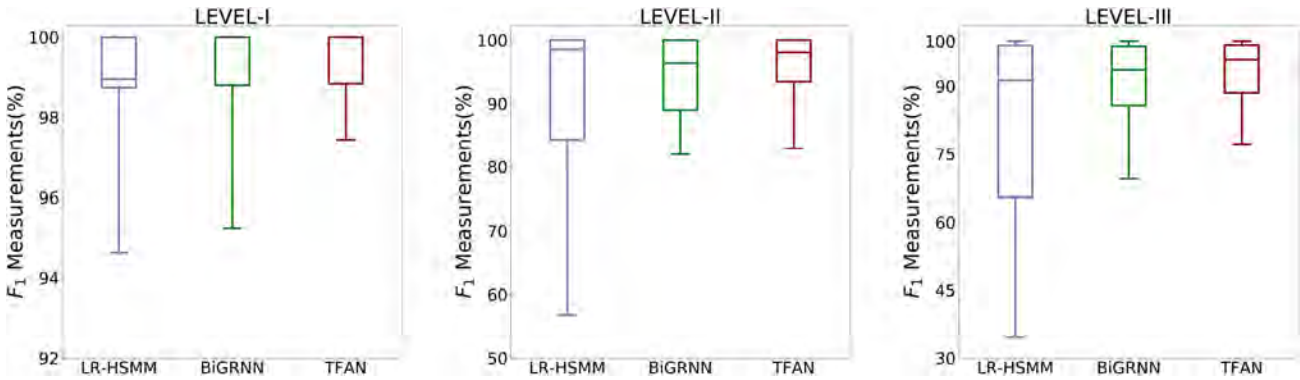

Fig. 8. F_1 measurements of the LR-HSMM method, the BiGRNN-based method and the TFAN-based method on databases with three levels of difficulty.

calculating the different metrics for each recording, then average over recordings in each of the data sets. The standard error of the averages results was also shown.

Fig. 8 illustrated the discrepancy of the performance stability over each heart sound recording across the TFAN-based method, the BiGRNN-based method and the LR-HSMM method on LEVEL-I, LEVEL-II and LEVEL-III.

1) Comparison With the LR-HSMM Method: According to Table IV, the TFAN-based method outperformed the LR-HSMM method on most of the test sets, especially on Training-C, Training-D, Training-F, Test-C, Test-D and Test-I. The LR-HSMM method achieved the total F_1 score of 89.59% on Training-C, 94.43% on Training-D, 89.01% on Training-F, 89.97% on Test-C, 94.00% on Test-D, 96.18% on Test-I, while an enormously improvement can be seen for the TFAN-based method with the F_1 score of 94.71%, 96.80%, 92.55%, 96.49%, 98.90% and 98.70% respectively. However, the total F_1 score of the TFAN-based method on Test-B is slightly lower than LR-HSMM (96.25%), which was 94.97% without the adaptive Wiener filter and 94.09% with the adaptive Wiener filter.

In Table V, a significant improvement of performance on the TFAN-based method compared to LR-HSMM could be observed On Level-II and Level-III (94.17% to 87.56%, $p < 0.0001$ and 91.31% to 78.46%, $p < 0.0001$). In comparison of standard errors, the TFAN-based method reduced the standard error by at least a factor of two comparing to the LR-HSMM method.

Fig. 9 showed two examples of automatically segmented heart sound recordings by the TFAN-based method and the LR-HSMM method. Repeated mistakes happened in Fig. 9(a) and (b) for the LR-HSMM method when segmenting PCG signals of arrhythmia and tachycardia.

2) Comparison With the BiRNN-Based Method: According to Table IV, the TFAN-based method outperformed the BiGRNN-based method on the whole data sets. Their overall F_1 scores approximated on Training-A*, Training-D, Test-C and Test-D. Meanwhile evident improvement in performance could be observed on Training-B (99.4% to 92.66%), Training-E* (97.76% to 94.37%), Training-F (92.55% to 86.50%), Test-G (95.94% to 92.47%) and Test-I (98.70% to 94.49%).

Table V showed that the TFAN-based method outperformed the BiGRNN-based method on LEVEL-I, LEVEL-II and LEVEL-III. As difficulty of segmentation escalated, the average F_1 scores of the TFAN-based method increased by around 2% compared to the BiGRNN-based method (94.17% to 92.63% on LEVEL-II and 91.31% to 88.45% on LEVEL-III). Note that the both methods showed comparable stability on each data set based on the standard errors reported in Table V.

3) Comparison of DL-Based Methods and the LR-HSMM Method: The BiGRNN-based method and the proposed TFAN-based method were both DL-based methods, sharing the loss function and inference function in our experiments. In Table IV, the DL-based methods performed better on Training-C and Test-C compared to the LR-HSMM method. And the both

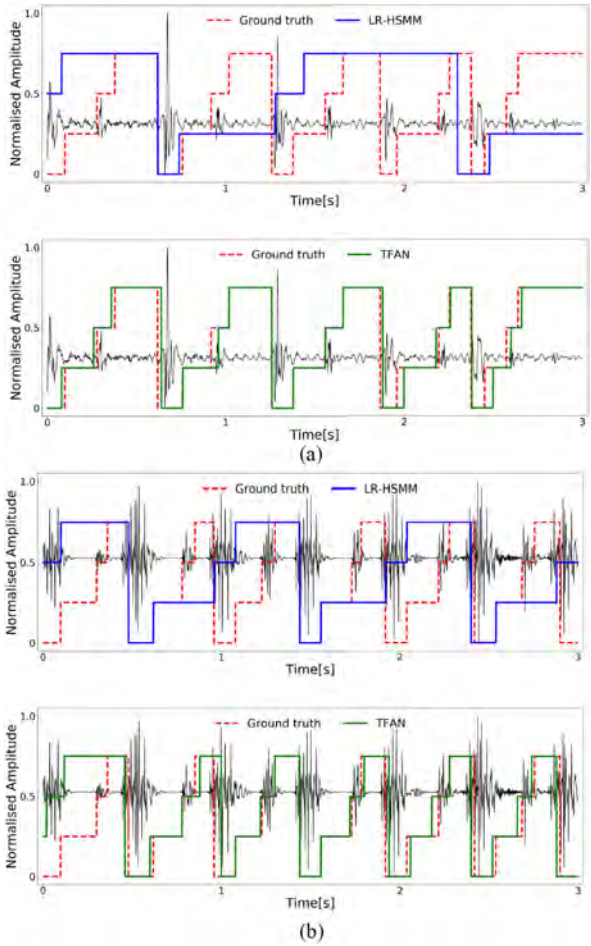

Fig. 9. Example of segmented pathological PCG signal with arrhythmia (a) and tachycardia (b) using the LR-HSMM method and the TFAN-based method. The four states of the heart cycle, S1, systole, S2, diastole, are represented by the staircase plot with the value of 0, 0.25, 0.5, 0.75 respectively.

methods failed to exceed the segmentation performance of the LR-HSMM method on Test-B. Moreover, according to Table V and Fig. 8, the DL-based methods provided noticeable improvement to the segmentation performance on LEVEL-II and LEVEL-III.

4) Evaluation of Adaptive Wiener Filter: The introduction of the adaptive Wiener filter in the TFAN-based method resulted in better performance on most of the test sets (except Training-C and Test-B), especially with a nearly 3% increase on Training-F. For Training-C and Test-B, the adaptive Wiener filter caused slightly drop of around 1% on performance when segmenting S2 and diastole.

VI. Discussion

The results reported in Section V provide a comprehensive comparison between the TFAN-based method, the LR-HSMM method and the recent BiGRNN-based method. From Table IV we can see that the TFAN-based method matched or outperformed the LR-HSMM method except for Test-B. The same situation happened to another DL-based method, the BiGRNN-based method. We hypothesize this is because Test-B is significantly different from the rest of the data, and in some way contains unusual noise or timing in the S2 and diastole periods (where the performance was most affected).

Notably, the characteristics statistics of each data set reveal that the majority heart sound recordings in Training-B and Test-B have obscure S2 sounds (Table II). It may be caused by contamination of noise and murmur or the stethoscopes with poor sensing performance. Since the TFAN-based method do not use a probability distribution to constrain the duration of the states, as well the BiGRNN-based method, the methods may have failed to locate S2 unlike the LR-HSMM which can infer the most probable location even in high noise (the sensitivity-specificity trade-off).

The performances of the DL-based methods and the LR-HSMM method were both outstanding when dealing with heart sounds from patients with normal sinus rhythm. Therefore, the global average overall F-scores of the three methods are both in the mid 90's, with the TFAN-based method ($F_1 = 96.72\%$) outperforming the BiGRNN-based method ($F_1 = 94.18\%$) and the LR-HSMM method ($F_1 = 94.56\%$). For data sets containing a certain amount of heart sound recordings with abnormal heart rhythms, such as Training-C, Training-D, Training-F and Test-C, we observe that the DL-based methods are always superior to the LR-HSMM method. The examples in Fig. 9 further highlight the distinctions.

From Fig. 9(b), we observe that although the TFAN-based method miss detects the S1 sound at the beginning, the following segmentation will be corrected in time. Unlike the TFAN-based method, the LR-HSMM method always fails to detect events when the intervals were irregular or incompatible with the prior probabilistic distribution of the state duration. This is a significant result, since we are looking to diagnose abnormality and the DL-based methods (TFAN and BiGRNN) are more applicable in real-world clinical environment.

Utilization of adaptive Wiener filter in the TFAN-based method was effective in most of the situations except for Training-C and Test-B, which led to a slight drop in performance for segmenting S2 and diastole. The most likely reason is that the adaptive Wiener filter may attenuate the weak murmurs and disappearing S2 sounds. This can be improved by redesigning the Wiener filter to include these specific features in the pass band.

The improvement of performance was reinforced when the three methods were tested on the data sets with different difficulty levels (LEVEL-I, LEVEL-II and LEVEL-III). We refer to the results reported for LEVEL-III, with the TFAN-based method ($F_1 = 91.31\%$) outperforming the LR-HSMM method ($F_1 = 78.46\%$) and the BiGRNN-based method ($F_1 = 88.45\%$). Moreover, according to Table V and Fig. 8, we observe that the stability of the TFAN-based method and the BiGRNN-based method is superior to the LR-HSMM method in segmenting complicated heart sounds.

The BiGRNN-based method matches the TFAN-based method on a certain number of data sets with relatively higher signal quality. But obviously, the generalization of the

BiGRNN-based method is inferior to the TFAN-based method. In comparison of model size, the weight file of TFAN is 2.3 Mb with 290,453 parameters, superior to BiGRNN model, which is 8.1 Mb with 1,016,805 parameters. Considering the both methods shared the proposed loss function and the designed dynamic inference approach, the advantage of TFAN over BiGRNN is in model structure.

The proposed framing module in TFAN slices the feature map into frames after the encoder (see Section III-B). Then the decoder implicitly learns the conditional probability distribution of state given encoded feature matrix for each time step. Comparing to BiGRNN, this structure enables TFAN to be more flexible in learning the decision boundaries between distributions of different heart sound states and brings the advantage of dealing with out-of-distribution data. Therefore, although the training set was limited into 50 recordings, the TFAN-based method still achieved the best performance.

The inner framing operation is also equivalent to incorporating the feature transformation of the signal during the model learning process. On the premise of ensuring the time resolution as much as possible, the feature expression dimension in each frame is improved. Unlike non-adaptive static feature extraction methods, such as envelop filters and spectrogram transform, this structure makes the model capable of capturing the features of the inter-state variability and the state transition information dynamically. This results in a high sensitivity for detecting the onsets and offsets of S1 and S2 precisely and reduces errors introduced by other heart sounds (e.g. S3) and noise. Moreover, the proposed TFAN-based method does not introduce the current error information into subsequent calculations for identifying the S1 and S2 in the next cycle.

However, there are two key limitations to the TFAN-based method. Firstly, the TFAN-based method was prone to missing weak or disappearing S2 sounds and identified the subsequent S1 sound as S2 in such cases. Secondly, the TFAN-based method tended to falsely identify some brief or transient noise as S1 or S2 sounds if the noises were similar to S1 or S2. This is basically an inherent problem of any classification technique, although with enough data we expect to be able to remove such events that appear at implausible times in the sequence of states, considering all pathological states.

VII. Conclusion

This paper proposed a novel method for heart sound segmentation of S1, systole, S2 and diastole. The method built up a frame-level feature classifier for the four components by an original temporal framing network. The study was focused on how to incorporate the state transition information into algorithm without using HMMs. The introduction of state transition loss and dynamic inference effectively addressed the problem within one model. Moreover, the TFAN-based method did not require explicit modeling of timing and was therefore able to generalize to arrhythmia and other high variability recordings more effectively than the current state of the art. Even though the training set was restricted to a small database with 50 single-source recordings randomly selected from Training-A, it was noted that the TFAN-based method provided a substantial improvement, particularly for more difficult cases, and on data sets not represented in the public training data. Future work will examine how increasing the number of training patterns and modeling the distribution of latent space to improve the performance. However, we note that the more data we use, the more we must use lower quality data, or make enormous effort to improve the labels.

Further work is also required to understand how this approach will provide improved diagnostic performance, although it is logical to assume better segmentation will lead to improved diagnostics.

Acknowledgment

The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] M. E. Tavel, "Cardiac auscultation: A glorious pastbut does it have a future?" *Circulation*, vol. 93, no. 6, pp. 1250–1253, 1996.
[2] S. Mangione *et al.*, "The teaching and practice of cardiac auscultation during internal medicine and cardiology training: A nationwide survey," *Ann. Internal Med.*, vol. 119, no. 1, pp. 47–54, 1993.
[3] C. Y. Liu *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiol. Meas.*, vol. 37, no. 12, pp. 2181–2213, 2016.
[4] P. Carvalho *et al.*, "Low complexity algorithm for heart sound segmentation using the variance fractal dimension," in *Proc. IEEE Int. Workshop Intell. Signal Process.*, 2005, pp. 194–199.
[5] J. Vepa, P. Tolay, and A. Jain, "Segmentation of heart sounds using simplicity features and timing information," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 469–472.
[6] J. Pedrosa, A. Castro, and T. T. Vinhoza, "Automatic heart sound segmentation and murmur detection in pediatric phonocardiograms," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2014, pp. 2294–2297.
[7] S. Ari, P. Kumar, and G. Saha, "A robust heart sound segmentation algorithm for commonly occurring heart valve diseases," *J. Med. Eng. Technol.*, vol. 32, no. 6, pp. 456–465, 2008.
[8] Z. Yan *et al.*, "The moment segmentation analysis of heart sound pattern," *Comput. Methods Programs Biomed.*, vol. 98, no. 2, pp. 140–150, 2010.
[9] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in *Proc. Comput. Cardiol. Conf.*, 1997, pp. 105–108.
[10] L. Huiying, L. Sakari, and H. Iiro, "A heart sound segmentation algorithm using wavelet decomposition and reconstruction," in *Proc. 19th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 4, 1997, pp. 1630–1633.
[11] S. Sun *et al.*."Automatic moment segmentation and peak detection analysis of heart sound pattern via short-time modified Hilbert transform," *Comput. Methods Programs Biomed.*, vol. 114, no. 3, pp. 219–230, 2014.
[12] A. Moukadem *et al.*."A robust heart sounds segmentation module based on s-transform," *Biomed. Signal Process. Control*, vol. 8, no. 3, pp. 273–281, 2013.
[13] A. Castro *et al.*."Heart sound segmentation of pediatric auscultations using wavelet analysis," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2013, pp. 3909–3912.
[14] H. Naseri and M. Homaeinezhad, "Detection and boundary identification of phonocardiogram sounds using an expert frequency-energy based metric," *Ann. Biomed. Eng.*, vol. 41, no. 2, pp. 279–292, 2013.
[15] J. Vepa, "Classification of heart murmurs using cepstral features and support vector machines," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2009, pp. 2539–2542.
[16] D. Kumar *et al.*., "A new algorithm for detection of S1 and S2 heart sounds," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, 2006, pp. 1180–1183.
[17] T. E. Chen *et al.*."S1 and S2 heart sound recognition using deep neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 372–380, Feb. 2017.

[18] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-HSMM-based heart sound segmentation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 822–832, Apr. 2016.
[19] J. Oliveira *et al.*."Adaptive sojourn time HSMM for heart sound segmentation," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 642–649, Mar. 2019.
[20] F. M. Noman *et al.*."A Markov-switching model approach to heart sound segmentation and classification," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 3, pp. 705–716, Mar. 2020.
[21] D. Gill, N. Gavrieli, and N. Intrator, "Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model," in *Proc. Comput. Cardiol. Conf.*, 2005, pp. 957–960.
[22] S. E. Schmidt *et al.*."Segmentation of heart sound recordings by a duration-dependent hidden Markov model," *Physiol. Meas.*, vol. 31, no. 4, pp. 513–529, 2010.
[23] C. Y. Liu, D. Springer, and G. D. Clifford, "Performance of an open-source heart sound segmentation algorithm on eight independent databases," *Physiol. Meas.*, vol. 38, no. 8, pp. 1730–1745, 2017.
[24] G. D. Clifford *et al.*."Recent advances in heart sound analysis," *Physiol. Meas.*, vol. 38, no. 8, pp. E10–E25, 2017.
[25] C. Potes *et al.*."Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds," in *Proc. Comput. Cardiol. Conf.*, 2016, pp. 621–624.
[26] E. Kay and A. Agarwal, "Dropconnected neural networks trained on time-frequency and inter-beat features for classifying heart sounds," *Physiol. Meas.*, vol. 38, no. 8, pp. 1645–1657, 2017.
[27] M. Zabihi *et al.*."Heart sound anomaly and quality detection using ensemble of neural networks without segmentation," in *Proc. Comput. Cardiol. Conf.*, 2016, pp. 613–616.
[28] T. C. I. Yang and H. Hsieh, "Classification of acoustic physiological signals based on deep learning neural networks with augmented features," in *Proc. Comput. Cardiol. Conf.*, 2016, pp. 569–572.
[29] E. Messner, M. Zöhrer, and F. Pernkopf, "Heart sound segmentation-an event detection approach using deep recurrent neural networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1964–1974, Sep. 2018.
[30] F. Renna, J. H. Oliveira, and M. T. Coimbra, "Deep convolutional neural networks for heart sound segmentation," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2435–2445, Nov. 2019.
[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
[32] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5069–5073.
[33] V. Peddinti *et al.*."Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 373–377, Mar. 2018.
[34] G. D. Clifford *et al.*."Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016," in *Proc. Comput. Cardiol. Conf.*, 2016, pp. 609–612.
[35] S. Y. Lee *et al.*."Electrocardiogram and phonocardiogram monitoring system for cardiac auscultation," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1471–1482, Dec. 2019.
[36] P. Arnott, G. Pfeiffer, and M. Tavel, "Spectral analysis of heart sounds: Relationships between some physical characteristics and frequency spectra of first and second heart sounds in normals and hypertensives," *J. Biomed. Eng.*, vol. 6, no. 2, pp. 121–128, 1984.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[38] American National StandardsInstitute, "Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms," *ANSI/AAMI Standard EC57*, 2012.