A Fully Convolutional Hybrid Fusion Network for Identification of S1 and S2 from Phonocardiogram

Juyeong Jung

Department of Medical Science

The Graduate School, Yonsei University

A Fully Convolutional Hybrid Fusion Network for Identification of S1 and S2 from Phonocardiogram

Directed by Professor Hyuk-Jae Chang

The Master's Thesis
submitted to the Department of Medical Science,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Master of Medical Science

Juyeong Jung

Jun 2024

This certifies that the Master's Thesis

of Juyeong Jung is approved.

Thesis Supervisor : Hyuk-Jae Chang

Thesis Committee Member#1 : Inhyun Jung

Thesis Committee Member#2 : Hackjoon Shim

The Graduate School

Yonsei University

December 2022

ACKNOWLEDGEMENTS

2 년이 넘는 시간 동안 값진 경험을 하며, 배울 점이 많은 교수님들 및 연구원분들과 함께 연구할 수 있어 매우 뜻깊은 시간이었습니다. 먼저 연세대학교 CONNECT-AI 연구실에서 연구할 수 있게 해 주신 장혁재 지도교수님께 진심으로 감사드리며, 항상 연구원들 지도편달에 힘써 주시는 심학준 교수님과 바쁘신 와중에도 제 석사 학위 심사에 시간을 할애해 주신 정인현 교수님께도 감사의 말씀을 올립니다.

저에게 지금의 연구실에서 인턴 생활을 할 수 있는 기회를 주신 홍영택 박사님께 감사의 말씀을 드리며, 지식이 얕은 저에게 많은 가르침을 주신 장영걸, 김세근, 홍승균, 하성민, 정성희, 전병환 박사님께도 감사드립니다. 힘들 때 의지하며 같이 연구한 한경훈, 안경진, 이지나, 정현석, 김가은, 정다운, 김지연, 최자영, 정시현, 구희준, 전재익 연구원에게도 고맙다는 말을 전합니다. 같이 일하며 고생하시는 행정팀, 통계팀 및 온택트 직원분들께도 감사의 인사를 드립니다.

집에서 저를 항상 걱정하고 응원해주시는 어머니, 아버지와 대학원생인 저를 자랑스럽게 여겨 주시는 할머니, 할아버지께도 항상 감사드리며, 표현을 자주 못해 죄송합니다. 고민이 많을 시기를 겪고 있는 동생에게도 격려의 말을 전합니다.

제 곁에서 도움을 주신 모든 분들께 다시 한번 감사드리며, 늘 건강하시고 행복한 일이 가득하시길 바라겠습니다.

정주영 올림.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

A Fully Convolutional Hybrid Fusion Network for Identification of S1 and S2 from Phonocardiogram

Juyeong Jung

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Hyuk-Jae Chang)

Cardiac auscultation is simple, inexpensive, and helps in the early diagnosis of heart diseases. However, because it requires extensive training, only a few specialists can detect abnormal heart sounds via auscultation. A phonocardiogram (PCG) is a recording of heart sounds, and a computerized algorithm for PCG analysis can support the clinical use of cardiac auscultation. It is important to detect the fundamental components—i.e., the first heart sound (S1) and second heart sound (S2)—in PCG analysis.

In this study, we propose a fully convolutional deep fusion network that comprehensively analyzes heterogeneous envelopes and scalogram features. We evaluated three variants of the proposed method—early, late, and hybrid fusion—and found that multimodal factorized bilinear pooling-based hybrid fusion produced the best results. Specifically, it exhibited state-of-the-art segmentation performance, with an accuracy of 0.9455, positive predictive value of 0.9688, and

sensitivity of 0.9832. To the best of our knowledge, this is the first study to completely interpret the heterogeneous features in PCG segmentation.

Key words : heart sound, phonocardiogram(pcg), deep learning, convolutional neural network(cnn), hybrid fusion, envelope, scalogram

A Fully Convolutional Hybrid Fusion Network for Identification of S1 and S2 from Phonocardiogram

Juyeong Jung

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Hyuk-Jae Chang)

I. INTRODUCTION

1. The importance of various feature-based heart sounds analysis

Heart auscultation is a simple and inexpensive first-line diagnostic test for early screening of heart abnormalities. The two fundamental components of heart sound, the first heart sound (S1) and second heart sound (S2), are produced by the mechanical activities of the heart valves. S1 denotes the start of the systolic phase, which is caused by the closure of the mitral and tricuspid valves. S2 occurs early in the diastolic phase and is caused by the closure of the aortic and pulmonary valves. **Figure 1** shows an example of heart sound with S1 and S2 labels. Abnormal states S3 and S4 are observed as noise patterns between S1 and S2 in the case of heart abnormalities (mainly valve abnormalities) and appear as physiological heart sounds such as murmurs, clicks, and splitting. Based on these characteristics, heart abnormalities can be detected by heart auscultation in a non-

invasive manner. However, despite the accessibility and simplicity of heart auscultation, its role is gradually diminishing because of the time and effort required to master it, diagnostic variances, and the advent of more accurate downstream tests.

A phonocardiogram (PCG) is a digital recording of analog heart sound by an electronic stethoscope, allowing repetitive listening and signal analysis. The computerized algorithm for PCG analysis supports the detection of subtle abnormal signal patterns, helping with precise diagnosis and contributing to the clinical use of heart auscultation. The first step in PCG analysis is identifying the location of heart sounds, S1 and S2. It is crucial to accurately localize S1 and S2 in order to identify abnormal signals and provide explanations [1], [2]. In practice, PCG signals are often obtained under conditions that make it difficult to identify S1 and S2, such as weak amplitude signals or motion-induced fricatives (breathing, dialog, etc.). Consequently, accurately identifying S1 and S2 from PCG signals is challenging.

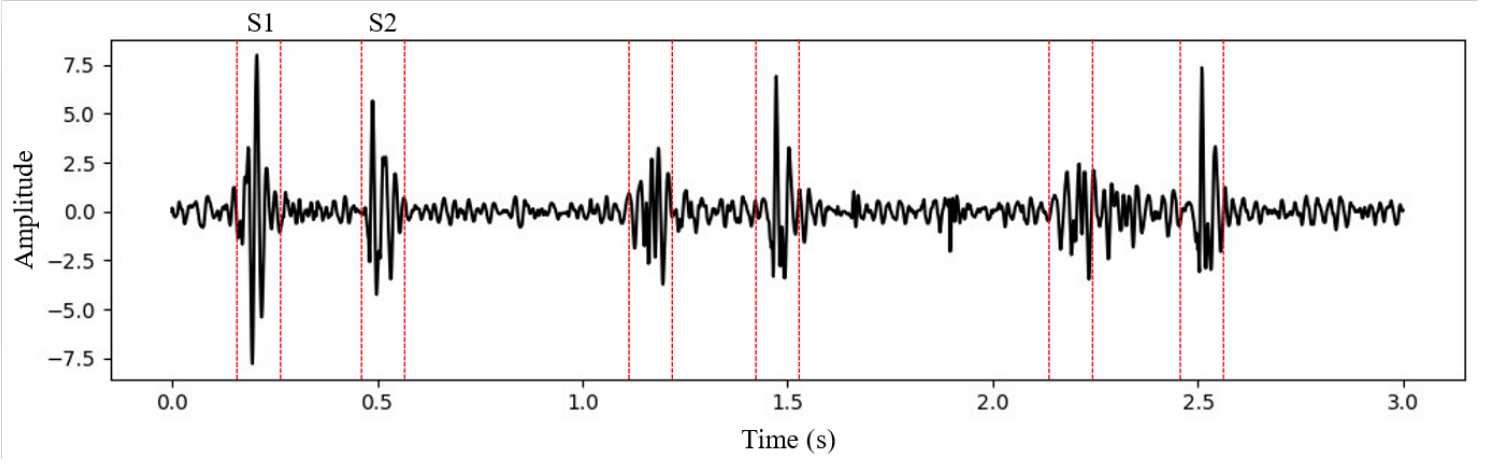

Figure 1. Normal heart sound pattern with states S1 and S2. In the normal heart sounds, there is little noise except for S1 and S2 sounds.

2. Related works

Many methods have been proposed to identify S1 and S2 from PCG signals [3]. According to their methodological characteristics, these methods can be divided into three categories: 1) envelope feature-based method, 2) spectral feature-based method, 3) deep neural network-based method.

The envelope is one of the most widely used features for signal processing. The envelope represents a smooth curve outlining its amplitude extremes from an oscillating input signal. The envelope features are extracted from the results applied with signal processing techniques such as Hilbert transformation and Shannon energy calculation. Rezek et al. [4] and Gupta et al. [5] used homomorphic envelopes to extract the primary amplitude of signals, and Liang et al. [6] used a Shannon energy-based envelope to segment PCG signals. However, specific rules and assumptions were required to identify S1 and S2 from the envelope features unless the probabilistic modeling is used.

The Hidden Markov Model (HMM) series has been used as a representative probabilistic framework for PCG segmentation. Gill et al. [7] proposed an HMM-based PCG segmentation method that uses a homomorphic envelope as an observed input sequence. Ricke et al. [8] proposed an HMM-based method that uses multiple features: the average Shannon energy, delta Shannon energy, and delta-delta Shannon energy of the heart sound signal. Springer et al. [9] proposed a Hidden Semi-Markov Model (HSMM)-based method that uses multiple envelope features: homomorphic, Hilbert, wavelet, and power spectral density (PSD) envelopes. Because the probabilistic modeling does not rely on experimental assumptions, they perform better than without probabilistic modeling. However, HMM series has limitations in the long-range dependencies within the sequences, and high-level feature fusion is not supported even though multiple envelope features are used simultaneously.

Spectral feature, another frequently used feature for PCG analysis, can be extracted by Fourier transform and wavelet transform (WT). Short-time Fourier transform (STFT) effectively erases unnecessary frequency components by converting signals from the time domain to the time-frequency domain. However, it's temporal resolution is limited due to fixed-size windows. In contrast, WT represents the signal's frequency components over time by controlling the translation and scale parameters of the wavelets and has an optimal balance between time and frequency resolution.

Patidar et al. [10] used CWT-based features and a least-squares support vector machine to identify heart-valve disorders. Nivitha Varghees et al. [11] used WT to segment heart sounds and classify abnormal heart sound. Vikhe et al. [12] used STFT and continuous wavelet transform (CWT) to detect heart sound abnormalities. Ghosh et al. [13] compared the various types of spectral features, including short-time Fourier transform (STFT) and wavelet transform (WT).

Recently, deep neural networks—including convolutional neural networks (CNNs) and recurrent neural networks (RNNs)—have shown promising performance in the field of not only computer vision but also signal processing. Renna et al. [14] proposed a 1D convolutional U-Net based PCG segmentation method with envelope features [9]. In [9], the first convolutional layer integrates multiple envelope features on the channel axis while simultaneously performing temporal modeling on the time axis. RNNs, representative frameworks for sequential data, have been used for various signal-processing tasks, such as speech recognition and enhancement [15]. RNNs have been successfully applied to long-range and multi-level sequential modeling, which are the limitations of HMMs. Fernando et al. [16] proposed bi-directional long short-term memory (BiLSTM) with an attention mechanism for PCG segmentation. In [16], the proposed method takes the envelope and time-frequency features such as

homomorphic, Hilbert, wavelet, PSD, and Mel frequency cepstral coefficients (MFCCs) as input and estimates the location of states S1 and S2 from them. However, some input data, including the spectral features used in the studies above, are linearly interpolated to increase temporal resolution. That may lead to detailed information loss and degrade the performance of the segmentation model.

Kay and Agarwal [17] proposed an abnormal heart sound detection method that interprets four different features from PCG signals: 1) CWT, 2) MFCCs, 3) inter-beat statistics (the mean and standard deviation of the length), and 4) complexity (the spectral entropy, skewness, and kurtosis of the signal). They used principal component analysis to reduce the dimension size of all features and combine them as an input to a fully connected neural network. The interpretation of the integrated input features improved the detection performance, but the performance gains were not significant owing to the shallow level of integration.

There have been several fusion techniques proposed in recent years. Fukui et al. [18] proposed multimodal compact bilinear pooling (MCB) for visual question answering. MCB projects the image and text features in high dimensions. Then the projected features are fused via the elementwise product. However, MCB requires a high computational cost to project features into high-dimensional space. Zhou et al. [19] proposed a multimodal factorized bilinear (MFB) pooling layer to fuse image and text feature. The MFB consisted of a feature expand stage and a feature squeeze stage. In the squeeze stage, sum pooling was used to approximate the representations of the input data.

Peng et al. [20] compared early, late, slow fusion to find proper feature fusion. In early fusion, input data is integrated into one input, and in late fusion, the results are integrated in the final layer after each network has learned the input data. They proposed slow fusion to integrate two spatial-temporal information. In

[21], they improves natural language processing performance by incorporating CNN and RNN submodules as combined network and it named hybrid fusion model. The necessity for feature fusion has been recognized by many researchers, but deep-level of feature fusion for PCG signal segmentation has yet to be studied. We believed it was essential to study PCG signal segmentation using deep learning-based high-level feature fusion.

3. Contribution of this paper

In this study, we propose a novel, fully convolutional deep fusion network for identifying S1 and S2 from a PCG. The proposed network takes envelope and scalogram features, which are complementary and dimensionally heterogeneous, as input. As in [9], [14], we used four envelopes: Hilbert, homomorphic, wavelet, and PSD. These multiple envelopes help identify the temporal characteristics of the signal. As a scalogram feature, the CWT feature of the 2D representation is used, which has an optimal balance between time and frequency localization in the time-frequency domain.

We explicitly employ two sub-modules for intermediate representations and a fusion module that integrates them to achieve effective deep-level fusion from two heterogeneous but complementary features. The two sub-modules are composed of multi-layered convolutional layers, allowing parameter- and computational-efficient temporary modeling. The fusion module uses a convolutional multimodal factorized bilinear (MFB) pooling, a modified version of [19], to effectively consider all interactions between the intermediate representations from two sub-modules. Finally, S1 and S2 are identified through the last 1 x 1 convolution after time-wise information fusion.

To the best of our knowledge, this is the first study to focus on the integrated interpretation of the heterogeneous features at a deep level to analyze PCG signals.

The proposed network effectively integrates these heterogeneous features at a deep level, thereby improving the identification performance for S1 and S2.

II. PROPOSED METHOD

In this section, we describe the proposed novel deep-fusion method for identifying S1 and S2 from heterogeneous features for PCG signals. It consists of three main parts: 1) preprocessing and input feature extraction, 2) learning the IRs, and 3) fusion and inference. **Figure 2** shows the workflow of the proposed method. Preprocessing is a fundamental step in the elimination of these noise components. Because PCG signals often contain undesired noise components such as fricative sounds and environmental noise. The proposed fusion method takes four envelopes in the 1D time domain and a scalogram in the 2D time-frequency domain as input features. These features are extracted from preprocessed PCG signals. Preprocessing and input feature extraction are described in Section II-1.

The proposed method produces IRs from two input features to integrate complementary information. Since they contain information that have different dimensions and characteristics of PCG, it is more effective to integrate them after making them into IRs rather than direct fusion and then integrate them. The module structure for learning IR is described in Section III-2. After fusing them, the proposed method outputs the final inference from the integrated features. The details of the fusion and inference module are described in Section III-3.

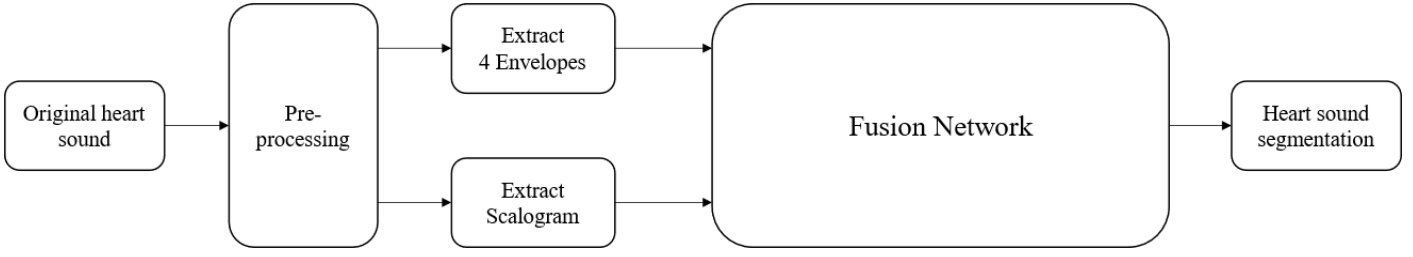

Figure 2. Workflow of the proposed phonocardiogram (PCG) segmentation method.

1. Preprocessing and feature extraction

Heart sounds are usually contaminated by various sources and noise levels, such as the voices of patients and staff, friction between stethoscopes, and clothing. These noises make it difficult to accurately detect principal heart sounds components. Therefore, signal preprocessing is an essential part of heart-sound segmentation. The fundamental heart-sound components S1 and S2 have dominant low-frequency characteristics (20--150 Hz). As the first preprocessing step, Butterworth filters [22] are used to remove undesired noise components in the signal. The cut-off frequencies for the Butterworth filters are 25 Hz and 400 Hz for the high-pass and low-pass bands, respectively. Spike removal is then applied to the signal to remove the abnormal spike amplitudes.

As in [9], [14], we used four different envelopes: Hilbert, homomorphic, wavelet, and power spectral density (PSD). These multiple envelopes help identify the temporal characteristics of the signal. The Hilbert envelope is computed as the absolute value of the Hilbert transform, which can transform a real-valued signal into a complex one [23]. The homomorphic envelope is computed as homomorphic filtering, which can remove certain noise [24]. The wavelet envelope is computed using a wavelet transform, which can overcome the disadvantages of the Fourier transform [25]. The PSD envelope is computed by multiplying each frequency bin in a fast Fourier transform by its complex conjugate [26]. **Figure 3** shows an example of the four different envelopes extracted from a preprocessed PCG signal.

The continuous wavelet transform (CWT) provides a 2D representation in the time--frequency domain, called a scalogram. The scalogram represents the frequency components of the signal over time by controlling the translation and scale parameters of the wavelets and has an optimal trade-off between time and frequency localization. Among them, CWT have continuous values for the scale

and translation parameters, which are usually used in the scalogram analysis of signals. The CWT can be defined as

$$\Psi_x(s,\tau) = \frac{1}{\sqrt{s}} \int x(t)\ \psi^*\left((t-\tau)/s\right)dt \qquad (1)$$

where $\psi(t)$ is the mother wavelet function, $*$ is the complex conjugate, and s and τ are the scale and translation parameters, respectively. In this study, we employed a Morlet wavelet as the mother wavelet [12]:

$$\psi(t) = e^{2j\pi ft}\, e^{-t^2/2\sigma} \qquad (2)$$

where t denotes the time parameter, f denotes the frequency of the wavelet, and σ denotes the Gaussian width [12]. Envelope and scalogram features are down-sampled at 50 Hz to synchronize with the electrocardiogram (ECG) as a ground truth signal for the S1 and S2 labels and normalized using the mean and standard deviation [9], [14]. **Figure 4** shows an example of a scalogram extracted from a preprocessed PCG with the labels S1 and S2.

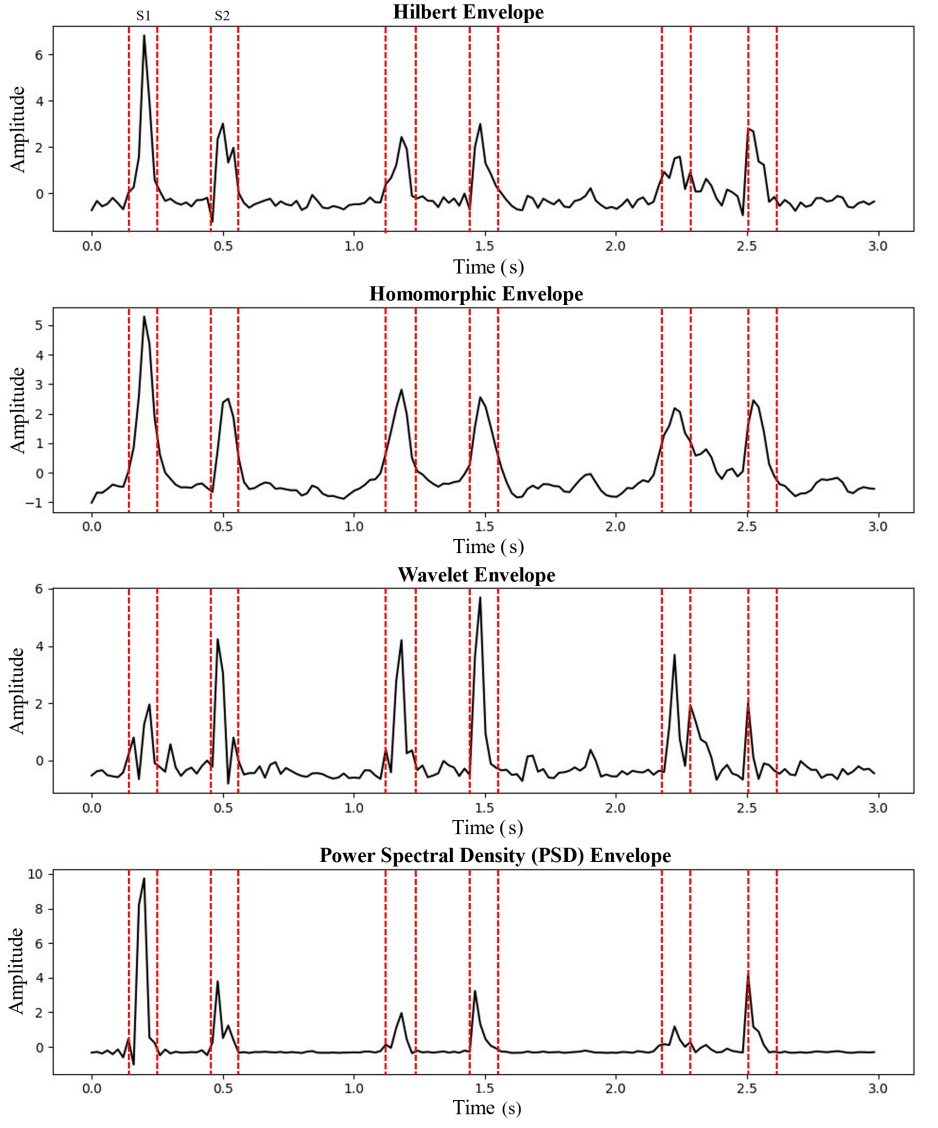

Figure 3. Four envelope features extracted from a PCG signal with labels S1 and S2. Certain amplitudes are strong in the states S1 and S2.

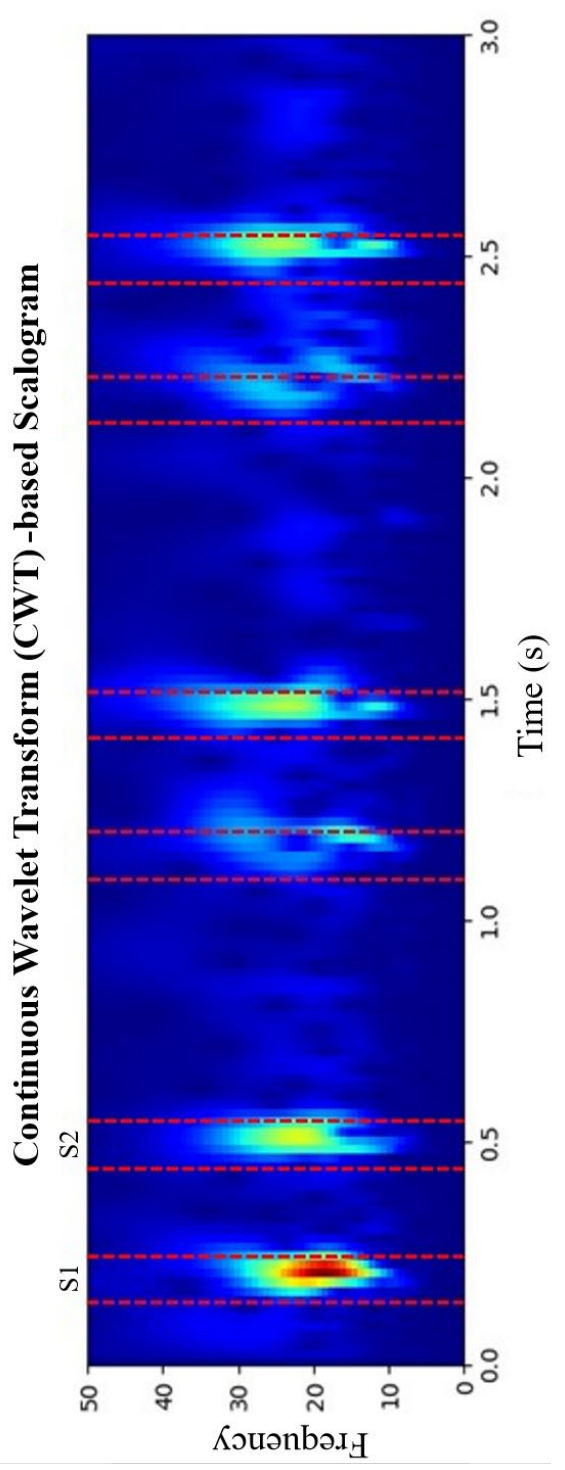

Figure 4. Example of scalogram with heart sound states S1 and S2. Blue and red indicate the lowest and highest energies, respectively. Certain frequency energy appears strongly in the states S1 and S2.

2. Structure of learning IR module

The proposed network has two submodules that generate IRs from heterogeneous inputs and another fusion module that integrates the IRs. The architecture of the proposed network is shown in **Figure 5**. Before presenting the details of the proposed network, we define the symbols for the two heterogeneous features. The envelope feature $\mathrm{E} \in \mathbb{R}^{4\times \mathrm{T}}$ consists of the following four envelopes: Hilbert $E_{hb} \in \mathbb{R}^{1\times T}$, homomorphic $E_{hm} \in \mathbb{R}^{1\times T}$, wavelet $E_{wv} \in \mathbb{R}^{1\times T}$, and PSD $E_{psd} \in \mathbb{R}^{1\times T}$, where T denotes the time of the sample. These four envelopes are concatenated along the channel axis. The scalogram feature $\mathrm{S} \in \mathbb{R}^{\mathrm{F}\times \mathrm{T}}$ is transformed using a continuous wavelet transform. F is the number of frequency bins (scale bins) and T is the time length of the sample.

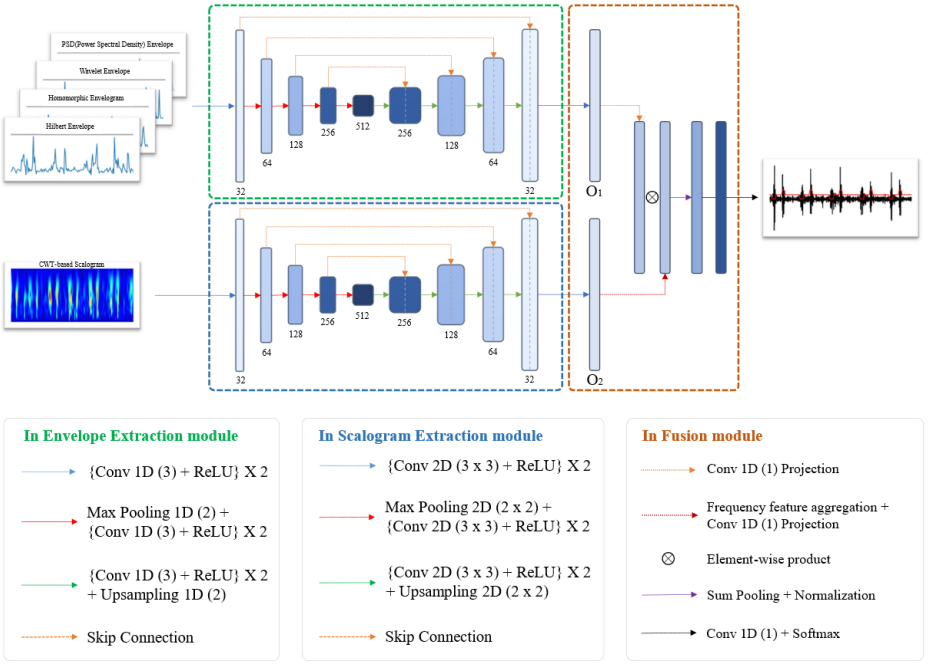

Figure 5. The architecture of the proposed fully convolutional hybrid fusion network.

A. Intermediate representation for envelope

Inspired by [14], the envelope representation module $\mathrm{M_e}$ is a U-Net architecture. The convolution block consists of two 1D convolution layers with a kernel of size of 3 and is activated with a rectified linear unit (ReLU) [27]. Zero padding is applied before the convolution operation to prevent dimensional reduction of the feature map. We employ a dropout [28] between the convolution layers to prevent overfitting. In the encoding path, 1D max-pooling is applied, and the number of feature maps is doubled. A skip connection between the encoder and decoder blocks is used to improve the gradient flow. The decoding path uses 1D upsampling to recover the original time resolution.

$$\mathbf{E} = \left[E_{\{\mathrm{hb}\}}, E_{\{hm\}}, E_{\{\mathrm{wv}\}}, E_{\{psd\}}\right],$$
$$\mathrm{R_E} = \mathrm{M_e}(\mathbf{E}; \theta_\mathrm{e}) \quad (3)$$

where [] denotes concatenation, θ_e are the parameters of the envelope representation module, and $\mathrm{R_E} \in \mathrm{R}^{\mathrm{C_E} \times \mathrm{T}}$ is the IR for the envelope.

B. Intermediate representation for scalogram

The scalogram representation module $\mathrm{M_s}$ has the same architecture as $\mathrm{M_e}$; however, 2D convolution is applied instead of 1D convolution. The convolution block consists of two 2D convolution layers with a 3×3 kernel and is activated with ReLU. Batch normalization [29] is used for training stability. An adaptive average pooling (AAP) [30] layer is attached to the end of the $\mathrm{M_s}$ module. The AAP aggregates the deep scalogram representation effectively along the frequency axis, enabling the integration of two IRs from the envelope module $\mathrm{M_e}$ and the scalogram module $\mathrm{M_s}$ for all times.

$$\mathrm{R}_S = \mathrm{AAP}\left(\mathrm{M_s}(\mathrm{S}; \theta_s)\right) \quad (4)$$

where θ_s is a parameter of the scalogram representation module.

3. Structure of fusion and inference module

A. Naïve fusion

Before the fusion with IRs from envelope and scalogram, we tested naïve feature fusion to test whether the combination has a complementary effect. This naïve fusion has two approaches early fusion and late fusion. 1) Early fusion: This combines the heterogeneous input features, and then the combined input feature is interpreted by the submodule M_e. 2) Late fusion: This combines the output probability from each submodule M_e and M_s. This means that each module is trained independently, and then the output probabilities from M_e and M_s are averaged, as in the model ensemble methods.

In this study, we tested three naïve feature fusions: 1) Early fusion between 1D envelope and 1D scalogram (early fusion), 2) Late fusion between 1D envelope and 1D scalogram (late fusion 1), and 3) Late fusion between 1D envelope and 2D scalogram (late fusion 2).

For the experiments, we used 1D scalogram by transforming 2D scalogram. In 1D scalogram, the frequency axis of 2D scalogram is concatenated along the channel axis, which induces various frequency features at that time point. The 1D envelope and 1D scalogram are concatenated along the channel axis as inputs for early fusion. Therefore, the envelope representation module M_e, which consists of a 1D convolution layer, has interpreted envelope and frequency features over the same time point. For late fusion, we tested two late fusions. The difference between 1D scalogram and 2D scalogram is clear from the results of this experiment. The 1D convolutional kernel of M_e strides along the time axis and analyzes all frequency features. By contrast, the 2D convolutional kernel of M_s analyzes adjacent time and frequency features.

B. Hybrid fusion

There are many ways to combine the features of two different domains, such as element-wise summation and outer product. We integrate two IRs from the envelope (R_E) and a scalogram (R_S) instead of naïve fusion. This fusion approach, called hybrid fusion [20], [21], is effective for fusing heterogeneous features from different perspectives on different dimensions.

In Hybrid fusion, The IRs R_E and R_S from each submodule M_e and M_s are interpreted by fusion module M_f. In this study, we tested four IR fusion methods. 1) element-wise concatenation, 2) element-wise summation, 3) element-wise product, and 4) multi-modal factorized bilinear pooling (MFB) [19]. To implement the MFB-based fusion, we refer to the following formula from [19] :

$$z_i = x^T W_i y \qquad (5)$$

where $x \in \mathbb{R}^m$ is the IR from envelope feature, $y \in \mathbb{R}^n$ is the IR from scalogram feature, $W_i \in \mathbb{R}^{m \times n}$ is feature projection matrix, $z_i \in \mathbb{R}$ is the output of MFB. The aim of MFB is to obtain o-dimensional output by learning $W = [w_i, \dots, W_o] \in \mathbb{R}^{m \times n \times o}$. After several modifications in [19], Equation 5 can be rewritten as follow:

$$Z = SumPooling(\tilde{U}^T x \circ \tilde{V}^T y, k) \quad (6)$$

where $\tilde{U} \in \mathbb{R}^{m \times ko}$ and $\tilde{V} \in \mathbb{R}^{n \times ko}$, $\circ$ is element-wise product and k is the latent dimension of factorized matrices. $SumPooling$ is the method of using a 1D non-overlapping window with size k to perform sum pooling on x. After $SumPooling$, Z was normalized with power normalization and L2 normalization.

At the end of the fusion module M_f, there is a 1D convolution layers with a kernel of size of 1, which takes the fused IRs and outputs the class probabilities $\hat{p} \in \mathbb{R}^{C \times T}$ for all times. C is the number of classes (S1, S2, systolic, and

diastolic) and T is the time length of the sample.

$$\hat{p} = M_f(Z; \theta_f) \quad (7)$$

where θ_f is a parameter of the inference module.

III. EXPERIMENTAL EVALUATION

1. Dataset

In this study, we used the public dataset from PhysioNet/CinC Challenge 2016 [31]. This dataset consists of 792 heart sound recordings from 135 patients. Among them, 406 heart sounds were collected from patients with heart disease and the remaining 386 were collected from healthy individuals. Heart sounds were recorded at different points in the chest, with recording times varying from 1 to 35s, and all signals were sampled at 1 kHz. Based on a synchronized electrocardiogram, heart sounds are annotated for S1, S2, systolic, and diastolic. The synchronization was performed with the agreement between the five R-peak and final T-wave detectors.

To evaluate general performance, external validation was performed with an open dataset [32], which is not used for training. The dataset consists of 69 paired PCGs and ECGs, obtained from a total of 24 subjects. Eight of these were recorded in a comfortable (stressless) environment for 30 seconds, and the remaining 61 were recorded in a walking, running and cycling environment for 30 minutes.

2. Implementation details

We performed 10-fold cross-validation to evaluate segmentation performance. For the 10-fold cross-validation, the training dataset was divided into ten subsets. Then, nine subsets were used to train the network, and the remaining subset was used for the validation. The result of each validation fold was evaluated using the best-performance model, which was determined by the lowest validation loss. To prevent performance overestimation, it is important to ensure that each patient's data are not duplicated in the training and validation datasets. The process was repeated ten times for cross-validation.

We sampled a local patch from the PCG signal for generalized segmentation performance. Local patches were sampled from the preprocessed PCG signal using the defined sampling rules. The sampling rules were as follows: length of sample patch = 64, 128, 256, and 512, and stride = 1/8 of the patch length. For example, if the sample length is 64, approximately 0.64 s of heart sound is sampled. The adaptive moment estimation (Adam) optimizer [33] with a weight decay was used. The learning rate was 1e-4, and the weight decay rate was 1e-2. The batch size was 64, number of training epochs was 150, and dropout rate was 0.3. The weights of each layer were determined using the categorical cross-entropy loss function. We set MFB factor number k to 2, and output dimension o to 32. The proposed network was implemented using PyTorch 1.5.0 [34]. The model training took approximately one hour on our workstation, with 32 GB of RAM, an Intel Xeon(R) E-2174G CPU, and an Nvidia RTX-5000 GPU.

3. Evaluation Metrics

We employed three evaluation metrics to compare our proposed method with conventional methods [9], [14], [16]. The three metrics were accuracy (ACC), positive predictive value (PPV), and sensitivity (SEN). ACC is calculated by comparing the predicted sequence $\hat{s}(t)$ labels with the ground truth sequence $s(t)$ labels, which means that the predicted labels are correctly positioned compared to the ground truth labels. PPV and SEN are focused on the centers of S1 and S2. A true positive (TP) is counted when the center of S1 and S2 of $\hat{s}(t)$ is closer than 60 ms from the center of the corresponding S1 and S2 of $s(t)$. Any positive except for true-positive was considered a false-positive. PPV was calculated using the following equation:

$$\mathrm{PPV} = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FP}} \quad (6)$$

where # is the number of values. SEN is calculated using the following equation:

$$\mathrm{SEN} = \frac{\#\mathrm{TP}}{(\#\mathrm{S1} + \#\mathrm{S2})\text{in G.T}} \qquad (7)$$

The evaluation was performed for each patient. The inference was performed with sampled patches and then merged to obtain the original length.

4. Results

A. Performance comparison by feature

We compared the segmentation performance of the 1D envelope feature and the 2D scalogram feature. Segmentation was performed in four classes: S1, systole, S2, and diastole. The post-processing method MAX [14] was applied after the segmentation. The MAX operation replaces the out-of-order predicted values with appropriate values to fit the context.

For 1D envelope-based segmentation, the best performance was ACC = 0.9377, PPV = 0.9625, and SEN = 0.9673 when the patch length N was 512 and the stride was 64. The segmentation performance improved when the length of the sample patch increased. For 2D scalogram based segmentation, the best accuracy was superior to 1D envelope-based segmentation, with 1D envelope vs. 2D scalogram values as follows: ACC: 0.9377 vs. 0.9385, PPV: 0.9625 vs. 0.9604, SEN: 0.9673 vs. 0.9614 with a patch length N of 512 and a stride of 64. The performance is summarized in **Table 1**.

Table 1. Experimental results of single methods

	Envelope based			Scalogram based		
	ACC	PPV	SEN	ACC	PPV	SEN
N=64, S=8	0.9204	0.9429	0.9436	0.9211	0.9259	0.9261
N=128, S=16	0.9334	0.9539	0.9552	0.9359	0.9528	0.9534
N=256, S=32	0.9368	0.9579	0.9594	0.9359	0.9553	0.9567
N=512, S=64	0.9377	**0.9625**	**0.9673**	**0.9385**	0.9604	0.9614

N is the length of each patch. S is the stride between patches. The performance evaluation metrics are Sample Accuracy (ACC), Positive Predictive Value (PPV), and Sensitivity (SEN). Boldface indicates the best value.

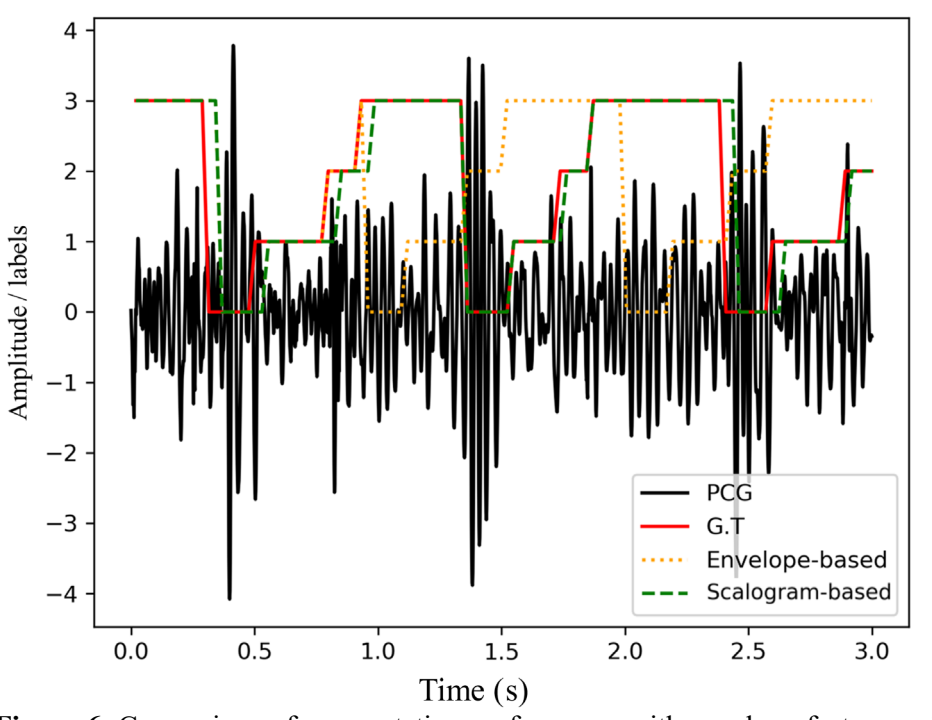

Figure 6. Comparison of segmentation performance with envelope feature and scalogram feature. Black line indicates PCG signal, red line indicates ground truth labels, yellow dotted line indicates 1D envelope-based method, and green dashed line indicates 2D scalogram-based method.

B. Performance comparison of the fusion methods

The best performance for early fusion was ACC = 0.9321, PPV = 0.9608, and SEN = 0.9788, with a patch length N of 512 and a stride of 64. When early fusion was compared with scalogram-based segmentation, PPV and SEN were higher in early fusion, but the ACC was higher in 2D scalogram (early fusion vs. 2D scalogram, ACC: 0.9321 vs. 0.9385, PPV: 0.9608 vs. 0.9604, SEN: 0.9788 vs. 0.9614). In early fusion, the 1D convolution kernel strides along the time axis and analyzes all concatenated envelope and frequency features simultaneously.

Late fusion 1 exhibited higher performance than early fusion in all metrics (early fusion vs. late fusion 1, ACC: 0.9321 vs. 0.9359, PPV: 0.9608 vs. 0.9717, SEN: 0.9788 vs. 0.9792), with a patch length N of 512 and stride of 64. Late fusion 2 exhibited higher performance than late fusion 1 on ACC and PPV (late fusion 1 vs. late fusion 2, ACC: 0.9359 vs. 0.9456, PPV: 0.9717 vs. 0.9767, SEN: 0.9792 vs. 0.9774), with a patch length N of 512 and a stride of 64. From this **Table 2**, analyzing a scalogram in a 2D domain is more helpful in improving performance than analyzing it in a 1D domain. These results indicate that a processing the scalogram in 1D is not suitable for inducing IR that has sufficiently learned the characteristics of the scalogram. Conversely, the 2D scalogram is appropriate for inducing proper IR and hybrid fusion 2 showed the best segmentation accuracy.

Table 2. Experimental results of Naïve fusion methods

	Naïve fusion								
	Early Fusion (1D Envelope + 1D Scalogram)			Late Fusion 1 (1D Envelope + 1D Scalogram)			Late Fusion 2 (1D Envelope + 2D Scalogram)		
	ACC	PPV	SEN	ACC	PPV	SEN	ACC	PPV	SEN
N=64, S=8	0.9207	0.9426	0.9624	0.9219	0.9062	0.9635	0.9270	0.9653	0.9370
N=128, S=16	0.9344	0.9489	0.9735	0.9292	0.9352	0.9679	0.9425	0.9626	0.9665
N=256, S=32	0.9327	0.9572	0.9728	0.9342	0.9645	**0.9809**	0.9429	0.9714	0.9732
N=512, S=64	0.9321	0.9608	0.9788	0.9359	0.9717	0.9792	**0.9456**	**0.9767**	0.9774

N is the length of each patch. S is the stride between patches. The performance evaluation metrics are Accuracy (ACC), Positive Predictive Value (PPV), and Sensitivity (SEN). Boldface indicates the best value.

For hybrid fusion, we tested four hybrid fusion types: 1) elementwise concatenation, 2) elementwise summation, 3) elementwise product, and 4) MFB pooling. To find the optimal hyperparameters of MFB, we experimented with various size of k and o in the **Table 3**. If o is greater than 32, one more convolution layer was added in the final convolution block.

Table 3. Experimental results of MFB hyperparameters

MFB hyperparameters			
k / o	ACC	PPV	SEN
1 / 32	0.9434	0.9656	0.9824
2 / 8	0.9412	0.9631	0.9823
2 / 16	0.9439	0.9628	0.9822
2 / 32	**0.9455**	**0.9688**	**0.9832**
2 / 64*	0.9430	0.9662	0.9823
3 / 32	0.9437	0.9639	0.9826
3 / 64*	0.9431	0.9637	0.9831
4 / 32	0.9388	0.9632	0.9820
5 / 32	0.9429	0.9650	0.9830

All experiments in this table used patch size 512, and stride 64. The performance evaluation metrics are Sample Accuracy (ACC), Positive Predictive Value (PPV), and Sensitivity (SEN). Boldface indicates the best value. * means that two convolution layer were used in the last convolution block.

After finding an optimal hyperparameters of MFB, we compared the performance by fusion types in **Table 4**. As a result, MFB outperforms other compared fusion types: ACC: 0.9455, PPV: 0.9688, and SEN: 0.9832 with a patch length N of 512 and stride of 64. This MFB fusion also superior to late fusion 2 in terms of the SEN, which was the best fusion methods among the tested naïve feature fusion methods (MFB fusion vs. late fusion 2, ACC: 0. 9455 vs. 0.9456, PPV: 0. 9688 vs. 0.9767, SEN: 0. 9832 vs. 0.9774)

Table 4. Experimental results of Hybrid fusion methods

	Hybrid fusion											
	Concatenate			Sum			Product			MFB		
	ACC	PPV	SEN	ACC	PPV	SEN	ACC	PPV	SEN	ACC	PPV	SEN
N=64, S=8	0.9201	0.9479	0.9622	0.9182	0.9420	0.9648	0.9155	0.9400	0.9638	0.9158	0.9417	0.9676
N=128, S=16	0.9354	0.9549	0.9736	0.9314	0.9539	0.9764	0.9286	0.9451	0.9775	0.9331	0.9515	0.9760
N=256, S=32	0.9382	0.9612	0.9776	0.9375	0.9614	0.9790	0.9341	0.9530	0.9822	0.9354	0.9566	0.9792
N=512, S=64	0.9409	0.9636	0.9806	0.9381	0.9652	0.9652	0.9401	0.9664	0.9664	**0.9455**	**0.9688**	**0.9832**

N is the length of each patch. S is the stride between patches. The performance evaluation metrics are Sample Accuracy (ACC), Positive Predictive Value (PPV), and Sensitivity (SEN). Boldface indicates the best value.

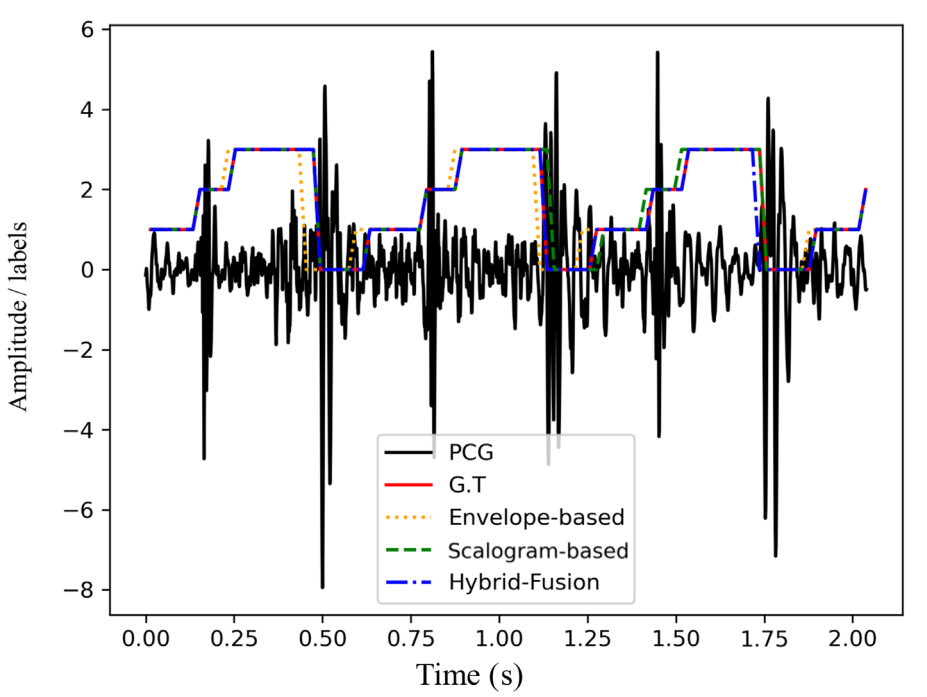

Figure 7. Comparison between the hybrid fusion method and the single feature-based methods. Black line indicates PCG signal, red line indicates ground truth labels, yellow dotted line indicates envelope-based method, green dashed line indicates scalogram-based method, and blue dashed-dotted line indicates MFB-based hybrid fusion method.

The result validates that the proper IRs from different features have complementary impacts on segmentation performance, and we experimentally demonstrate that the efficient and effective pooling approach can maximize the feature analysis performance. The experimental results are presented in **Table 4**.

C. External validation

To find points S1 and S2, the code provided by the challenge [32] was used. In this code, an R-peak is calculated from the synchronized ECG, and a systolic period is obtained by calculating an interval from the R-peak to the next R-peak. The points S1 and S2 of the PCG are obtained through the calculated systolic period and R-peak. However, several S1 and S2 points were located incorrectly due to the subject's voice and various noises. In this case, it is explained that the expert modified it manually. The evaluation was conducted with the best performance model and the result of the external validation are as follows: ACC: 0.7721, PPV: 0.7854, and SEN: 0.9743. It was validated with the most optimal patch size 512 and stride 64 proved in the above experiments.

IV. DISCUSSION

In this paper, we proposed a novel deep fusion network for PCG segmentation that jointly analyzes envelope and scalogram features. Furthermore, we experimentally demonstrated that the heterogeneous features from the PCG signal have complementary impacts, and the proposed fusion method shows state-of-the-art PCG segmentation performance.

In the independent use of features, 2D scalogram-based segmentation outperformed 1D envelope-based segmentation (**Table 1**). The signal envelope describes the change in the peak amplitude over time, whereas the scalogram describes the change in frequency over time. Therefore, the scalogram feature can consider a frequency feature that are not in the envelope feature.

The results indicate that the 2D scalogram provides better time--frequency resolution features for PCG segmentation than the combination of multiple 1D envelope features. This means that it is more efficient to analyze frequency components than amplitude features of PCG when considering a single feature.

Figure 6 shows the segmentation results of the 1D envelope-based method and 2D scalogram-based method on abnormal heart sound in the test set. The scalogram-based segmentation showed robust performance against noise because it focused on the frequency components of the signal. On the other hand, the envelope-based method is vulnerable to noise because it depends mainly on the amplitude of the signal. However, since the envelope feature can consider the patterns of PCG over time, it is important to consider the overall characteristics of the heart sounds.

In the next experiment, we tested naïve fusion and hybrid fusion with IRs. In naïve fusion, early fusion was used to validate whether combinations of different features can have complementary effects. And late fusion was used to validate whether it is better to consider both features at the same time or to combine the results after learning separately.

In early fusion, we transformed the 2D scalogram into a 1D scalogram. Early fusion was more efficient in terms of computational cost because it used only the 1D convolutional layers of M_e. By fusing envelope features and frequency features on the same network, we demonstrated that using heterogeneous features has complementary effects and provides better performance than using a single feature.

In late fusion, fusion occurred after the learning and evaluation of each feature-based model were completed. It shows better performance than early fusion, indicating that it is more effective to fuse after sufficiently learning the characteristics of the input data. However, late fusion is not intended to have proper IRs owing to independent optimization and the combination of outputs such as an ensemble. In addition, since two networks need to be trained, it takes more time and cost of training than training a single network.

The aim of hybrid fusion is to find a cost-effective, high-performance fusion method by fusing IRs that have sufficiently learned the features of two heterogeneous input datasets within one network. We used the best performance fusion method by experimentally comparing recently proposed methods with classical feature fusion methods. The experimental results demonstrated the importance of IRs in combining heterogeneous features. We found that combining the IRs after proper conversion is more effective than simply calculating and combining the IRs. The proposed deep fusion network exhibited

state-of-the-art PCG segmentation performance as a single network, with ACC = 0.9455, PPV = 0.9688, and SEN = 0.9832

Figure 7 compares hybrid fusion, envelope-based segmentation, and scalogram-based segmentation. Envelope and scalogram-based methods in heart sound with noise do not accurately detect S1 and S2 regions. On the other hand, the hybrid fusion-based method proposed in this paper shows stronger S1 and S2 segmentation performance by comprehensively considering the continuous temporal features and the frequency features of the heart sounds.

We also compared the segmentation performance of the proposed method with the methods proposed by Fernando et al. [16] and Renna et al. [14]. The method by Fernando et al. [16], based on a BiLSTM for sequential modeling and attention techniques, achieved ACC = 0.969, PPV = 0.963, and SEN = 0.972 on the same PhysioNet/CinC Challenge 2016 dataset. By contrast, the proposed deep fusion network (MFB pooling based hybrid fusion) exhibited better performance in terms of PPV and SEN (hybrid fusion vs. [16], ACC: 0.946 vs. 0.969, PPV: 0.969 vs. 0.963, SEN: 0.983 vs. 0.972). The proposed method achieved a more robust segmentation performance without a sequential modeling method.

The method proposed by Renna et al. [14] achieved a segmentation performance of ACC = 0.937, PPV = 0.958, SEN = 0.958. Our fusion method outperformed their method on all metrics. Their method is based on 1D CNN with 1D envelope features, and the HMM and HSMM were additionally utilized for sequential modeling. The sequential modeling methods utilized contribute to the performance improvement but are not effective on all metrics.

V. CONCLUSION

In this paper, we proposed a fully convolutional deep fusion network that comprehensively analyzes heterogeneous envelopes and scalogram features. We demonstrated the benefit of a comprehensive analysis of heterogeneous features and state-of-the-art PCG segmentation accuracy on a single network. Since only a few specialists can detect abnormal heart sounds using auscultation, various diagnostic modalities such as electrocardiography, ultrasound imaging, and Doppler techniques have contributed to reducing the clinical use of auscultation.

To use it in an actual medical environment, data and verification from the actual medical field are required; this will be the subject of future research. Furthermore, since the segmentation of S1 and S2 in heart sounds is finally intended to help detect effective cardiac murmur, so we plan to develop artificial intelligence technology for detecting cardiac murmur based on the method proposed in this paper.

REFERENCES

[1] G.D. Clifford, C. Liu, B. Moody, J. Millet, S. Schmidt, Q. Li, I. Silva, R.G. Mark, Recent advances in heart sound analysis, Physiol. Meas. 38 (2017) E10–E25.

[2] Y. Duck Shin, K. Hoon Yim, S. Hi Park, Y. Wook Jeon, J. Ho Bae, T. Soo Lee et al., The correlation between the first heart sound and cardiac output as measured by using digital esophageal stethoscope under anaesthesia, Pak J Med Sci. 30 (2014) 276–281.

[3] A.K. Dwivedi, S.A. Imtiaz, E. Rodriguez-Villegas, Algorithms for automatic analysis and classification of heart sounds–a systematic review, IEEE Access. 7 (2018) 8316–8345.

[4] I. Rezek, S.J. Roberts, Envelope extraction via complex homomorphic filtering, Technical Report TR-98-9 Technical Report. (1998).

[5] Neural network classification of homomorphic segmented heart sounds - ScienceDirect, (n.d.). https://www.sciencedirect.com/science/article/abs/pii/S1568494605000694 (accessed March 17, 2022).

[6] H. Liang, S. Lukkarinen, I. Hartimo, Heart sound segmentation algorithm based on heart sound envelogram, in: Computers in Cardiology 1997, 1997: pp. 105–108.

[7] D. Gill, N. Gavrieli, N. Intrator, Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model, Computers in Cardiology, 2005. (2005) 957–960.

[8] A.D. Ricke, R.J. Povinelli, M.T. Johnson, Automatic segmentation of heart sound signals using hidden markov models, in: Computers in Cardiology, 2005, 2005: pp. 953–956.

[9] D.B. Springer, L. Tarassenko, G.D. Clifford, Logistic Regression-HSMM-Based Heart Sound Segmentation, IEEE Transactions on Biomedical Engineering. 63 (2016) 822–832.

[10] S. Patidar, R.B. Pachori, A Continuous Wavelet Transform Based Method for Detecting Heart Valve Disorders Using Phonocardiograph Signals, in: G. Lee, D. Howard, D. Ślęzak, Y.S. Hong (Eds.), Convergence and Hybrid Information Technology, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012: pp. 513–520.

[11] V. Nivitha Varghees, K.I. Ramachandran, Effective Heart Sound Segmentation and Murmur Classification Using Empirical Wavelet Transform and Instantaneous Phase for Electronic Stethoscope, IEEE Sensors J. 17 (2017) 3861–3872.

[12] P.S. Vikhe, N.S. Nehe, V.R. Thool, Heart Sound Abnormality Detection Using Short Time Fourier Transform and Continuous Wavelet Transform, in: 2009 Second International Conference on Emerging Trends in Engineering & Technology, IEEE, Nagpur, India, 2009: pp. 50–54.

[13] S.K. Ghosh, R.K. Tripathy, R. Ponnalagu, A study on time-frequency analysis of phonocardiogram signals, Microelectronics and Signal Processing. (2021) 189–202.

[14] F. Renna, J. Oliveira, M.T. Coimbra, Deep Convolutional Neural Networks for Heart Sound Segmentation, IEEE J. Biomed. Health Inform. 23 (2019) 2435–2445.

[15] A. Graves, A. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013: pp. 6645–6649.

[16] T. Fernando, H. Ghaemmaghami, S. Denman, S. Sridharan, N. Hussain, C. Fookes, Heart Sound Segmentation Using Bidirectional LSTMs With Attention, IEEE J. Biomed. Health Inform. 24 (2020) 1601–1609.

[17] E. Kay, A. Agarwal, DropConnected neural networks trained on time-frequency and inter-beat features for classifying heart sounds, Physiol. Meas. 38 (2017) 1645–1657.

[18] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach,

Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding, (2016). http://arxiv.org/abs/1606.01847 (accessed October 6, 2022).

[19] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2017: pp. 1821–1830.

[20] M. Peng, C. Wang, T. Chen, G. Liu, X. Fu, Dual temporal scale convolutional neural network for micro-expression recognition, Frontiers in Psychology. 8 (2017) 1745.

[21] M.U. Salur, I. Aydin, A novel hybrid deep learning model for sentiment classification, IEEE Access. 8 (2020) 58080–58093.

[22] P.J. Arnott, SPECTRAL ANALYSIS OF HEART SOUNDS: RELATIONSHIPS BEWEEN SOME PHYSICAL CHARACTERISTICS AND FREQUENCY SPECTRA OF FIRST AND SECOND HEART SOUNDS IN NORMALS AND HYPERTENSIVES, (n.d.) 8.

[23] S.R. Messer, J. Agzarian, D. Abbott, Optimal wavelet denoising for phonocardiograms, Microelectronics Journal. 32 (2001) 931–941.

[24] I.A. Rezek, S.J. Roberts, Envelope Extraction via Complex Homomorphic Filtering, (n.d.) 9.

[25] S. Choi, Detection of valvular heart disorders using wavelet packet decomposition and support vector machine, Expert Systems with Applications. 35 (2008) 1679–1687.

[26] R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics, IEEE Transactions on Speech and Audio Processing. 9 (2001) 504–512.

[27] A.F. Agarap, Deep Learning using Rectified Linear Units (ReLU), ArXiv:1803.08375 [Cs, Stat]. (2019). http://arxiv.org/abs/1803.08375 (accessed March 24, 2022).

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov,

Dropout: A Simple Way to Prevent Neural Networks from Overfitting, (n.d.) 30.

[29] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in: Proceedings of the 32nd International Conference on Machine Learning, PMLR, 2015: pp. 448–456. https://proceedings.mlr.press/v37/ioffe15.html (accessed March 21, 2022).

[30] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: pp. 8759–8768.

[31] C. Liu, D. Springer, Q. Li, B. Moody, R.A. Juan, F.J. Chorro et al., An open access database for the evaluation of heart sound algorithms, Physiol. Meas. 37 (2016) 2181–2213.

[32] Kazemnejad, Arsalan, Gordany, Peiman, Sameni, Reza, EPHNOGRAM: A Simultaneous Electrocardiogram and Phonocardiogram Database, (n.d.).

[33] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, ArXiv:1412.6980 [Cs]. (2017). http://arxiv.org/abs/1412.6980 (accessed March 24, 2022).

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan et al., Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019: pp. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

ABSTRACT (IN KOREAN)

심장음의 S1과 S2 식별을 위한 완전 컨볼루션 기반 하이브리드 융합 네트워크

<지도교수 장혁재>

연세대학교 대학원 의과학과

정 주 영

심장 청진법은 간단하고 저렴하며 심장 질환의 조기 진단에 도움을 준다. 그러나 심장 청진음을 듣고 심장 질환을 구별하는 것은 상당한 노력과 훈련이 필요하기 때문에, 심장 청진음을 통해 비정상적인 심장 소리를 감지하는 것은 소수의 전문가만이 가능했다. PCG (Phonocardiogram)는 심장 청진음을 녹음한 것으로, 심장 청진의 임상적 사용을 지원하기 위해 디지털화된 PCG 분석 알고리즘은은 꾸준히 연구되어 왔다. PCG 분석에서 가장 중요한 것 중에 하나는, 심장음의 주성분인 첫 번째 심장음(S1)과 두 번째 심장음(S2)을 구별하는 것이다.

본 연구에서는 심장음의 서로 다른 특성을 갖는 포락선 특징과 스칼로그램 특징을 종합적으로 분석하는 완전 컨볼루션(합성곱) 연산

기반의 하이브리드 융합 네트워크를 제안한다. 본 논문에서 제안한 융합 방법의 강건함을 증명하기 위해 세 가지 변형인 초기, 후기 및 하이브리드 융합 방법 또한 평가했다. 특징 융합 방법들 중, 서로 다른 모달리티를 입력으로 받은 후 이중 선형 풀링 기반(MFB)으로 특징을 융합하는 하이브리드 융합이 최상의 결과를 보였다. 구체적으로 0.9455의 정확도, 0.9688의 양의 예측값, 0.9832의 민감도로 현존하는 가장 높은 성능을 보였다.

핵심되는 말 : 심장음, 심장 청진음, 인공지능, 합성곱 신경망, 하이브리드 융합, 포락선, 스칼로그램