International Journal of Basic and Applied Sciences, 14 (7) (2025) 1-10

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/k0nqfx70
Research paper

MW-Net: A Deep Framework for Segmentation and Detection of S_1 and S_2 Heart Sounds

Madhwendra Nath *, Subodh Srivastava

Department of Electronics and Communication Engineering, National Institute of Technology Patna, Patna-800005, India
**Corresponding author E-mail: madhwendran.phd19.ec@nitp.ac.in*

Received: October 7, 2025, Accepted: October 12, 2025, Published: November 2, 2025

Abstract

This paper proposes a Mish W-Net deep framework for precise segmentation of the primary heart sounds, S_1 and S_2, from PCG signals. The proposed deep framework incorporates a synchro-squeezed wavelet transform for signal-to-image conversion. Additionally, a random search-optimized complex diffusion unsharp masking is proposed to mitigate the issues of noise and image quality after the signal-to-noise conversion. Besides this, to achieve self-regularizing feedback learning, the Mish activation function is employed at the architectural level of the proposed Mish W-Net. The effectiveness of the proposed deep framework is evaluated on the publicly available PhysioNet/Computing in Cardiology PCG dataset. Comparative qualitative and quantitative assessment outperforms the existing methodologies. The proposed method achieves accuracy, precision, recall, F1-score, and intersection over union of 98.28%, 98.77%, 99.50%, 99.13%, and 98.28%. It shows an effective framework that offers precise segmentation and detection of heart sounds.

Keywords: *Complex Diffusion, Mish, Synchro-Squeezed Wavelet Transforms, Random Search, Unsharp Masking, W-net.*

1. Introduction

Clinical assessment of heart sounds (HS) primarily relies on cardiac auscultation, which offers the early diagnosis of cardiovascular diseases (CVDs) [1]. The major CVDs detectable via auscultation include hypertension, arrhythmias, valvular stenosis, and regurgitation. According to the Global Burden of Disease (GBD) study, CVD was responsible for 438 million disabilities and 19 million mortalities worldwide in 2021 [2]. Moreover, CVDs are responsible for 60% of adult mortality in India. [3]. Literature shows the primary analysis of HS relies on the clinical expertise of healthcare professionals using acoustic stethoscopes.

The phonocardiogram (PCG) [4] Has emerged as a complementary diagnostic tool that provides a digital and objective representation of the first and second heart sounds: S_1 and S_2 As illustrated in Fig.1, these HS are essential auditory indicators of the cardiac cycle. These are linked to particular valve-closing occurrences. [5]. These HS generally exist within the 20–200 Hz frequency range. The S_1 It is produced by the closing of the mitral and tricuspid valves. In addition, S_2 It is generated by the closing of the aortic and pulmonic valves. However, both conventional auscultation and PCG analyses require significant human intervention, which led to automated HS analysis.

To develop an accurate automated HS model based on PCG signals, segmentation plays a crucial role. The studies reveal that a PCG recording typically captures multiple cardiac cycles embedded with respiratory noise, ambient interference, and pathological murmurs. [6]. These challenges make accurate interpretation more difficult. The segmentation approaches are categorized as the energy envelope, frequency/time–frequency analysis, probabilistic, machine learning (ML), and deep learning (DL) methods. [7]. The energy envelope-based approaches include the Hilbert transform and Shannon energy. However, noise significantly affects these methods, and their effectiveness hinges on the choice of thresholding. The Short-Time Fourier Transform (STFT) and Wavelet Transform (WT) are the primary methods that belong to frequency/time–frequency analysis methods. [8]. Similarly, temporal and frequency resolution influence these methods' performance. Furthermore, these methods frequently require manual adjustment of their threshold values. The probabilistic models include Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). However, they rely heavily on accurate initialization and prior assumptions, which may not generalize across varying recording conditions. The performance of the ML models is reliant on handcrafted features, whereas the DL model suffers from overfitting. In addition to these issues, both approaches exhibit limited performance because of signal variability and noise. To overcome these challenges, a Mish W-Net deep framework has been proposed for precise segmentation of HS. S_1 and S_2. The major contributions are as follows:

- A Mish W-Net automated deep framework has been proposed for the segmentation and detection of S_1 and S_2 Heart sounds.
- A synchro-squeezed wavelet transform has been introduced for 1D PCG signal conversion to a 2D image.
- To improve the image quality and lessen the impact of noise, a random search-optimized complex diffusion unsharp masking has been proposed.
- To achieve self-regularizing feedback learning, the Mish activation function is integrated at the architectural level.

- A comprehensive qualitative and quantitative assessment has been conducted for the publicly available PCG dataset.

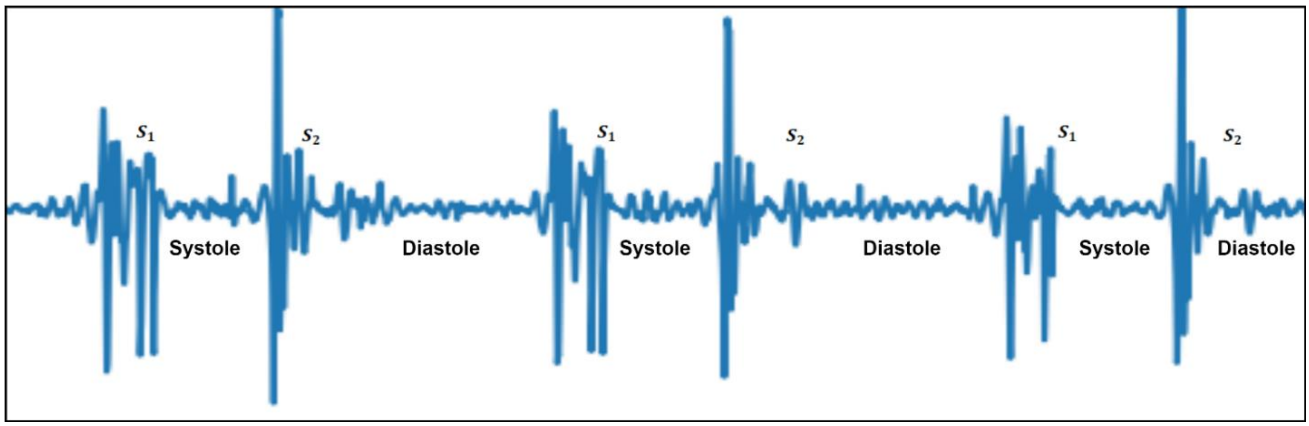

Fig 1: Illustration of the digital representation of heart sounds corresponds to the first heart sound S_1 and the second heart sound S_2

The remainder of this paper is structured as follows: Section 2 reviews existing literature on PCG segmentation and detection. Section 3 details the methods and models, including the dataset and the proposed methodology. Section 4 presents experimental design, evaluation metrics, and comparative performance assessment. Section 5 concludes the paper and outlines future research.

2. Related Work

In recent years, a wide range of techniques have been explored for the segmentation of heart sounds (HS) from phonocardiogram (PCG) signals. Chen et al. [9] employed a convolutional long short-term memory (Conv-LSTM) network for HS segmentation using the Massachusetts Institute of Technology heart sound database. Although the model captured sequential dependencies effectively, the baseline configuration was adopted without modification, limiting its ability to learn fine-grained spectral-temporal structures, especially under varying noise conditions. Park et al. [10] integrated the convolutional Fourier transform (CFT) within a U-Net architecture to segment heart sounds from PCG signals in both encoder and decoder paths. While the method demonstrated improved frequency localization, it was developed exclusively for one-dimensional PCG sequences, constraining its spatial-spectral representation capability. Liu et al. [11] proposed a time–frequency domain framework that employed envelope extraction after denoising to achieve segmentation. However, the method required manual threshold tuning, making it dataset-specific and sensitive to operator selection. Similarly, Xiao et al. [12] utilized the wavelet transform (WT) for denoising and the Hilbert transform for envelope extraction, enhanced by first-order Shannon energy. Although effective, its dependence on threshold values limited adaptability across varying recording environments. Renna et al. [13] customized a convolutional neural network (CNN) integrated with hidden Markov (HMM) and hidden semi-Markov (HSMM) models for HS segmentation. Despite improving temporal labeling, the approach relied on one-dimensional PCG inputs, restricting feature diversity. Messner et al. [14] compared several deep recurrent neural networks (DRNNs) such as RNN, LSTM, and gated RNNs, combined with virtual-adversarial training for robust segmentation; yet, these models inherited the baseline limitations of fixed sequential dependencies and lacked spatial feature extraction.

Attention-based architectures have recently gained traction for biomedical signal modeling. Fernando et al. [15] integrated bidirectional LSTM with an attention mechanism to emphasize salient cardiac cycles and improve segmentation precision. While effective, the approach remained dependent on hand-crafted envelopes and Mel-frequency cepstral coefficient (MFCC) features, limiting full automation. Mukherjee et al. [16] introduced a U-Net-based network for PCG denoising using short-time Fourier transform (STFT) representations of 1-D signals. Similarly, Venegas et al. [17] employed a U-Net combined with a Butterworth bandpass filter and Shannon energy-based segmentation, but the model architecture largely followed the baseline 1-D formulation without feature-level refinements.

Collectively, the literature indicates that Energy-envelope, probabilistic, time–frequency, machine learning (ML), and deep learning (DL) approaches dominate heart sound segmentation research. Envelope-based techniques such as Shannon energy and Hilbert transforms are highly susceptible to background noise. Probabilistic frameworks (e.g., HMM, HSMM) rely on strict prior assumptions regarding state transitions and emission distributions, often failing in non-stationary or noisy conditions. Time–frequency methods (e.g., STFT, WT) exhibit threshold-dependent performance, while ML models depend on hand-crafted feature design and domain expertise. Many DL-based methods adopt standard 1-D network baselines (e.g., U-Net, BiLSTM) without architectural adaptation, thereby limiting spatial–temporal generalization.

Recently, emerging transformer-based and hybrid attention models (e.g., Fernando et al., 2020; Park et al., 2025) have demonstrated improved context modeling in Biomedical signal segmentation. However, such frameworks are computationally intensive and require large datasets to converge, which is impractical for PCG recordings that are often short and noisy.

To address these limitations, this work proposes the Mish W-Net (MW-Net), a novel deep framework that combines synchro-squeezed wavelet transform (SSWT) based scalogram representations with random search-optimized complex diffusion unsharp masking (RSCDUM) and the Mish activation function. MW-Net is designed to enhance spectral-temporal discrimination while maintaining computational efficiency, enabling accurate segmentation and detection of S_1 and S_2 heart sounds from PCG signals.

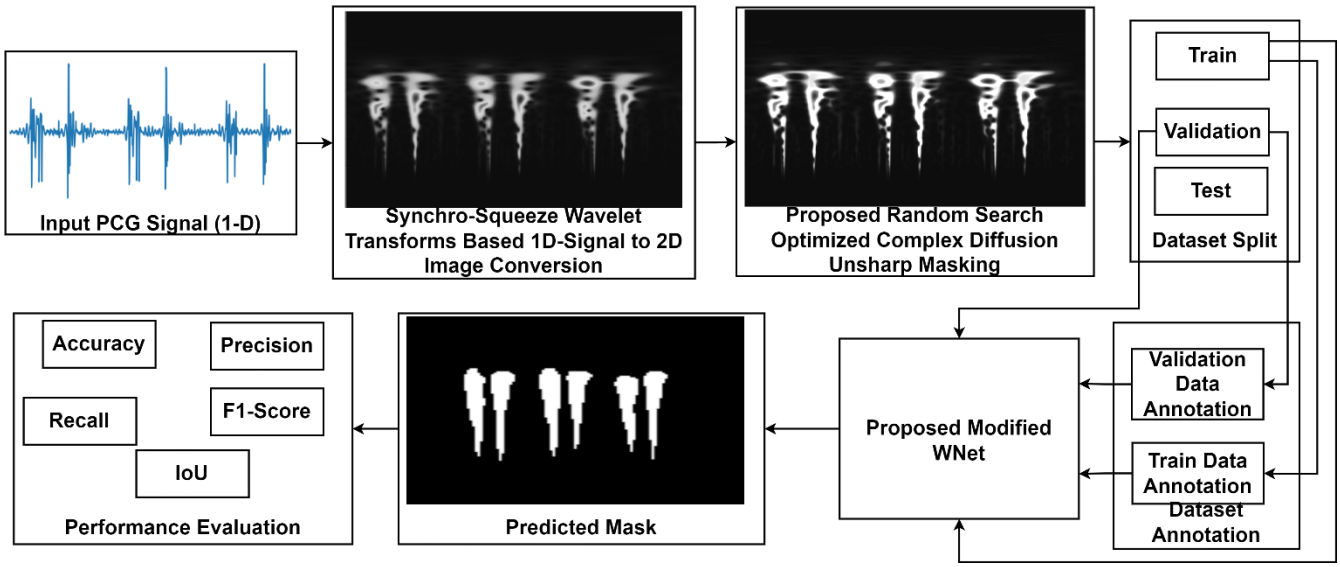

Fig. 2: Schematic outline of the proposed deep framework for the segmentation and detection of the heart sounds

3. Methods and Materials

The development of the proposed Mish W-Net (MW-Net) begins with PCG collection and preprocessing, followed by signal-to-image conversion, image enhancement using random search-optimized complex diffusion unsharp masking (RSCDUM), dataset splitting and annotation, model training, and quantitative/qualitative evaluation. The data is divided as train (70%), validation (15%), and test (15%). The training and validation of the proposed MW-Net has been done in the next step. Furthermore, the performance assessment has been presented. A schematic overview of the proposed deep framework for the segmentation and detection of the heart sound is shown in Fig. 1.

3.1 Dataset

The publicly available PhysioNet/Computing in Cardiology Challenge 2016 [18] The dataset is utilized for the development of the proposed MW-Net. It contains 3,153 heart sound recordings from 764 subjects. The subjects range from newborns to elderly adults, capturing a wide variety of normal and pathological cardiac conditions. The signals are recorded at 2,000 Hz and resampled to 1,000 Hz for consistency. Table 1 summarizes dataset attributes.

Table 1: Description of PCIN Challenge 2016 dataset

Attributes	Description
Number of subjects	764
Number of recordings	3153
Signal format	.wav
Recording locations	Aortic, pulmonic, tricuspid, and mitral auscultation sites

3.2 Proposed methodology for heart sounds Segmentation

Signal-to-image conversion is the first step toward implementing the proposed method after acquisition of the raw PCG signal. To accomplish this objective, Synchro-Squeezed Wavelet Transform (SSWT) [19] Is employed. It offers 2-D scalograms for the 1-D raw PCG signal. However, the literature demonstrates that raw PCG signals commonly suffer from variable cardiac-cycle lengths and background noise; these factors can blur spectral features in scalograms. To reduce such variability, a structured preprocessing pipeline is applied before SSWT. It includes the cropping of cardiac cycle length, resampling of the signal, and denoising of the signal through band-pass Butterworth filtering. [20], and the normalization of the signal. All these steps improve the effectiveness of the SSWT conversion of 1-D PCG signals into 2-D scalograms.

Let $x(t)$ represents the 1-D raw time domain PCG signal with t_c Cardiac cycle. The variable length of the cardiac cycle may cause the signal to lose its periodicity. Moreover, it results in blurred features in the generated scalograms. Therefore, cropping of the raw signal is performed to achieve a uniform scalogram. It is mathematically expressed as:

$$x_c(t) = x(t),\ t_c \in [t_s, t_e] \tag{1}$$

where, $x_c(t)$ is a cropped PCG signal, and $[t_s, t_e]$ Interval of the cropped signal. The raw signals are cropped to a fixed number of cardiac cycles (3–6 cycles) to improve temporal consistency and reduce scalogram blurring. Afterward, the cropped signal is resampled. The resampling standardizes the temporal resolution and reduces computational variability. The overall expression is governed as:

$$x_r[n] = x_c\left(\frac{n}{f_s}\right), n = 0, \dots, N-1 \tag{2}$$

where, $x_r[n]$ is a discrete resampled signal, N is the total sample and f_s Is the sampling frequency, and it is set to $1000\ Hz$. Afterward, to suppress irrelevant frequency components, a band-pass Butterworth filter is applied. It isolates the dominant spectral bands corresponding to S_1 and S_2 While attenuating noise. It is mathematically represented in Eq. (3).

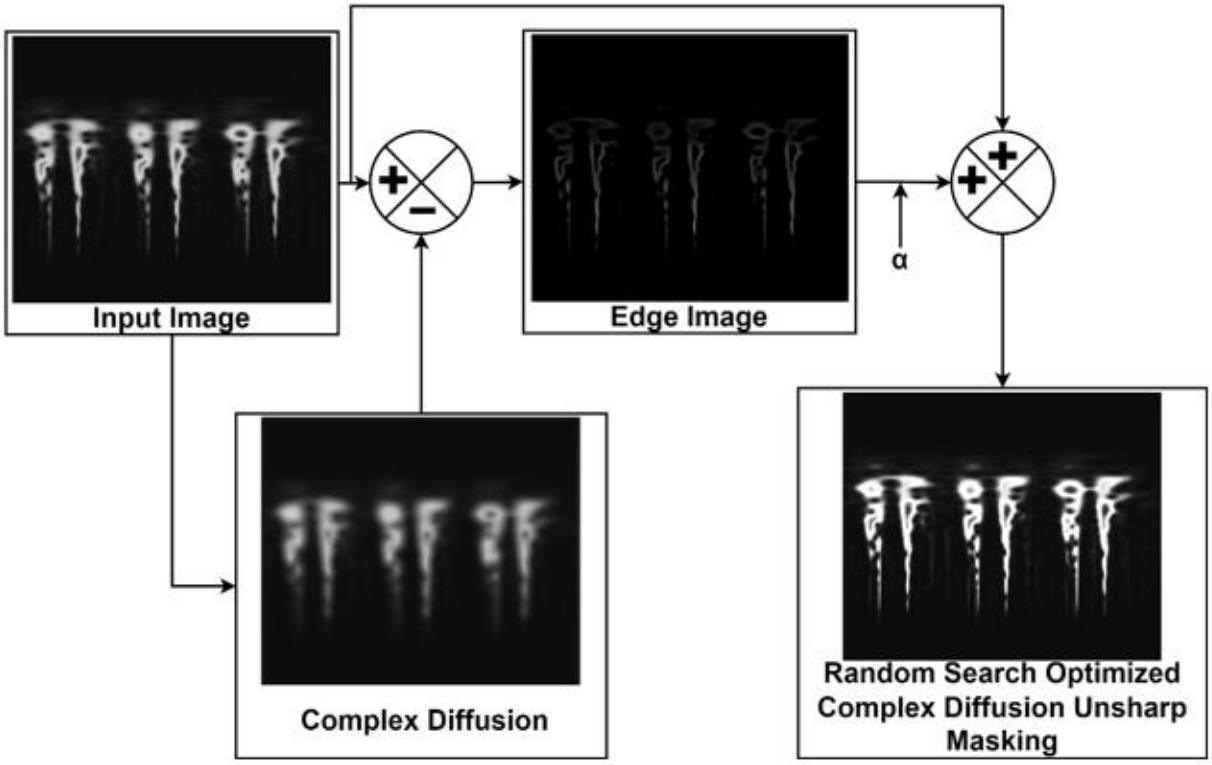

Fig 3: Proposed Random Search Optimized Complex Diffusion Unsharp Masking

$$x_f[n] = \sum_{k=0}^{M} b_k x_r[n-k] - \sum_{k=1}^{M} a_k x_f[n-k] \tag{3}$$

where, a_k and b_k Are filter coefficient. $x_f[n]$ It is a filtered PCG signal. Later, amplitude normalization is employed on the filtered signal. It mitigates the amplitude bias and variability caused by recording conditions. It is computed as:

$$x_n[n] = \frac{x_f[n]}{\max|x_f[n]|};\ n = 0, \ldots, N-1 \tag{4}$$

where, $x_n[n]$ It is a normalized signal. These preprocessing steps collectively enhance the signal quality. Afterward, SSWT is employed to transform normalized 1-D PCG signals into 2-D scalogram representations. The conventional continuous wavelet transform (CWT) produces a redundant time-scale representation. In contrast to this, SSWT reassigns wavelet coefficients from scale space to a sharpened frequency axis and generates an interpretable time-frequency map. The process begins with the computation of the CWT, followed by estimation of instantaneous frequencies from the phase information of the coefficients, and finally a synchro-squeezing step that reallocates the spectral energy to its true frequency bins, which offers a high-resolution scalogram. Eq. (5) illustrates the final mathematical expression of SSWT.

$$I_{SSWT} = S(w_k, b_m) = \sum_{n=0}^{N-1} x_n[n]\, \psi^*_{w_k}(n - b_m) \tag{5}$$

where, $S(w_k, b_m)$ is a 2-D scalogram image, ψ_{w_k}Is the wavelet function at frequency? w_k, b_m is a time shift, and $*$ Denotes the complex conjugation. It has been observed that the generated scalograms have low resolution and suffer from blurring. These issues affect the precise segmentation. Therefore, random search optimized complex diffusion unsharp masking (RSCDUM) has been proposed to enhance the quality of the scalogram.

The proposed RSCDUM integrates the attributes of a complex diffusion filter (CDF) [21] and unsharp masking (UM) [22] In a single framework. Fig. 3 exhibits the schematic outline of the proposed RSCDUM. The CDF offers the denoising of scalograms. It acts as a nonlinear low-pass operator. It suppresses noise while preserving important structural edges in the scalogram. CDF for the scalogram image I_{SSWT} Is computed as:

$$E(I_{SSWT}) = \nabla.\left(c\left(I_{SSWT_m}\right)\nabla I_{SSWT}\right) \tag{6}$$

Where, $c\left(I_{SSWT_m}\right)$, Is the diffusion coefficient [23]. It is estimated as:

$$c\left(I_{SSWT_m}\right) = \frac{e^{j\theta}.I_{SSWT_m}}{1+\left(\frac{I_{SSWT_m}}{k\theta}\right)^2} \tag{7}$$

where, k is a threshold constraint and θ Is the phase angle. Furthermore, Eq.(6) is minimized by Euler-Lagrange minimization and Gradient Descent. [24] To obtain the discretize form. It reads as:

$$I^{n+1}_{\mathrm{SSWT}} = I^{n}_{\mathrm{SSWT}} + \Delta t^n \nabla\left\{c\left(I_{SSWT_m}\right)^n \nabla I^{n}_{\mathrm{SSWT}}\right\} \tag{8}$$

where, n Is the number, and Δt Is the constant. The I_{SSWT}^{n+1} Eq. (8) offers a low-pass scalogram, and for simplicity, it may be expressed as I_{SSWTLP}.
To enhance the scalogram, the low-pass operator of UM is substituted with CDF. This modification improves the edge information of the scalogram. Furthermore, the high-pass or edge scalogram image can be obtained using the following scheme:

$$I_{SSWTHP} = I_{SSWT} - I_{SSWTLP} \tag{9}$$

The integrated framework of CDF and UM (CDFUM) for the scalogram image is governed as:

$$I_{CDFUM} = I_{SSWT} + \alpha * I_{SSWTHP} \tag{10}$$

where, I_{CDFUM} Is the unsharped image of the input scalogram image I_{SSWT}. α Is the edge scaling constant?

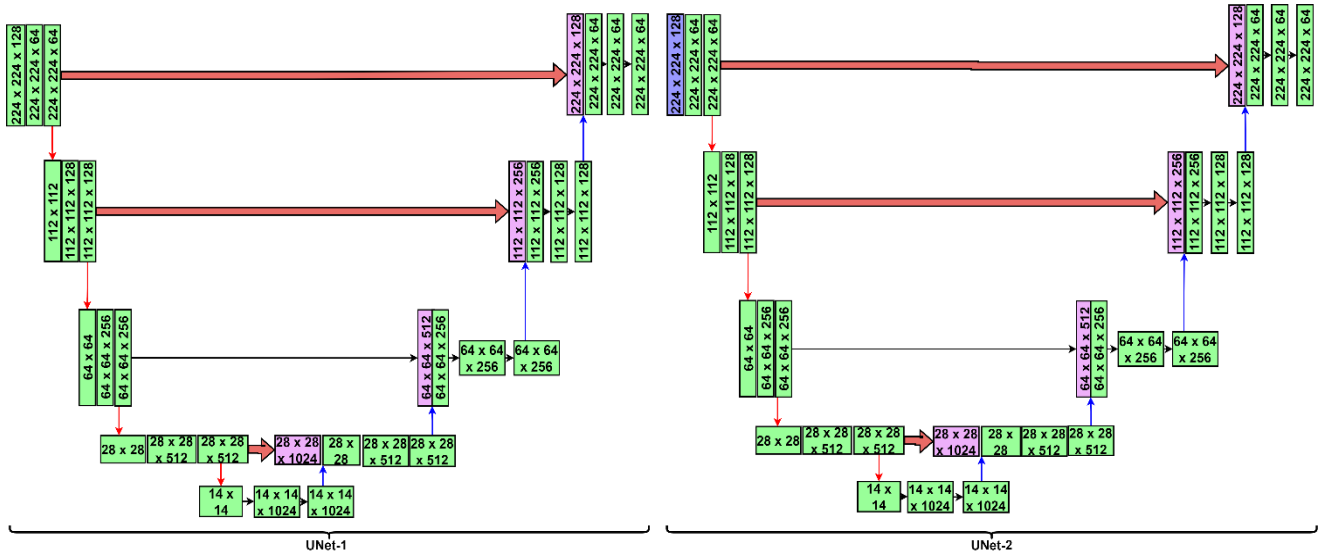

Fig 4: Detailed Architecture of the Proposed MW-Net

The quality of the scalogram image obtained via CDFUM is dependent on the selection of the controlling parameters, such as k, Δt, n, and α. Moreover, literature shows that these values are opted based on the user, which makes the overall process more iterative. Thus, random search (RS) optimization [25] Has been employed to tune the controlling parameter of CDFUM.
The structural similarity index map (SSIM) is applied as an objective function of RS-tuned CDUM (RSCDUM). The rationale behind opting for SSIM as an object function is to ensure that the enhanced scalogram maintains high perceptual similarity with the original while improving edge sharpness and resolution. The objective function reads as:

$$\{k^*, \Delta t^*, n^*, \alpha^*\} = \underset{k,\Delta t,n,\alpha}{\operatorname{argmax}}\, SSIM(I_{CDFUM}) \tag{11}$$

In the next step, the generated scalograms are augmented and have been classified into training, validation, and testing sets in a ratio of 70%:15%:15%. It is expressed as:

$$D_{RSCDUM} = (D_{train}, D_{val}, D_{test})_{RSCDUM} \tag{12}$$

Moreover, the training and validation scalogram images are annotated with the aid of the Visual Geometric Image Annotator (VIA) [26]. The process involved in it is expressed as:

$$D_{RSCDUM-train} = (I_i, M_i)_{i=1}^{N_1} \tag{13}$$

and,

$$D_{RSCDUM-val} = \left(I_j, M_j\right)_{j=1}^{N_2} \tag{14}$$

where, I_i, and I_j Are the scalograms? M_i, and M_j Are the corresponding annotated labels of both the training and validation scalogram data? N_1, and N_2 Signify the total number of annotated scalograms.
The proposed MW-Net is an advanced form of the conventional W-Net (CW-Net) [27] That offers the segmentation of instances. It employs two cascaded U-Nets [28] As demonstrated in Fig.3, the first U-Net serves as an autoencoder to learn compact latent representations of scalograms, while the second U-Net functions as a segmentation decoder to generate a pixel-wise map for heart sound segmentation.
Et $I \in \mathbb{R}^{H\times W\times C}$ is the input scalogram where H, W, and C denote the height, width, and number of channels, respectively. Furthermore, the input scalogram is processed with the encoder module. It extracts hierarchical features through successive convolutional operations and down-sampling. The literature shows that the CW-Net utilizes ReLU as an activation function. However, ReLU suffers from limitations such as non-smooth gradients and loss of negative information. These factors reduce the effectiveness of feature learning. To address these, Mish activation [29] is employed. It allows a small negative value and ensures a steady flow of gradients. Moreover, it offers a self-regularizing feedback learning, which improves feature learning. Thus, at level i The feature representation is obtained as:

$$F_i = Mish\left(W_i^{(2)} * Mish\left(W_i^{(1)} * I_i + b_i^{(1)}\right) + b_i^{(2)}\right) \quad (15)$$

where, $*$ Denotes the 2-D convolutional operator. $W_i^{(j)}$ and $b_i^{(j)}$ These are the weights and biases. The Mish activation function is computed as:

$$Mish(I) = I.\tanh(\ln(1 + e^I)) \quad (16)$$

The output is subsequently down-sampled using max pooling:

$$I_{i+1} = MaxPool(F_i) \quad (17)$$

Afterward, the feature extracted from the encoder module is processed through the bottleneck module, as shown in Fig.4. It applies convolutional operations without a pooling operation. Moreover, the ReLU activation is replaced with Mish. The revised expression reads as:

$$Z = Mish\left(W_B^{(2)} * Mish\left(W_B^{(1)} * X_n + b_B^{(1)}\right) + b_B^{(2)}\right) \quad (18)$$

In the next step, the bottleneck features are fed to the decoder module. It symmetrically up-samples the encoded features and combines them with corresponding encoder features through skip connections to preserve spatial details. At level j The up-sampled features are expressed as:

$$U_j = Up(Z_i) \quad (19)$$

and,

$$D_j = Mish\left(W_j^{(2)} * Mish\left(W_j^{(1)} * \left[U_j \parallel F_j\right] + b_i^{(1)}\right) + b_i^{(2)}\right) \quad (20)$$

where, $||$ denotes concatenation and $Up(.)$ Represents either transposed convolution or bilinear interpolation. The output segmentation mask of the first U-Net is expressed as:

$$S_1 = \sigma(W_o * D_1 + b_o) \quad (21)$$

where, $\sigma(.)$ Is the sigmoid function. The S_1 Offers a coarse segmentation and serves as input for the second cascade U-Net. It follows a similar encoder-decoder structure and generates a refined segmentation map. It corrects the coarse predictions, improves boundary sharpness, and reduces false positives. It is denoted as:

$$S_2 = UNet_2(S_1) \quad (22)$$

The segmentation performance is optimized by a hybrid loss function that integrates pixel-wise accuracy with boundary preservation. The overall loss function reads as:

$$L_{Total} = \lambda_1 L_{Seg} + \lambda_2 L_{Dice} \quad (23)$$

where, λ_1 and λ_2 These are trade-off hyperparameters. L_{Seg} and L_{Dice} They are computed in terms of binary cross-entropy and dice loss, respectively.

4. Results and Discussion

The proposed MW-Net method has been designed and developed for the publicly accessible PhysioNet/Computing in Cardiology Challenge 2016 [18] PCG dataset. It includes a total of 3153 PCG recordings. These PCG signals are initially cropped to a duration of 3 to 6 cardiac cycles. The cropped signals are resampled, denoised through a Butterworth bandpass, and the amplitude is normalized. Furthermore, the normalized signals are converted to a 2-D scalogram with the aid of SSWT. A total of 3000 scalograms are generated for the implementation of the proposed method. Moreover, to execute the proposed RSCDUM, initial hyperparameter values are obtained from the existing literature. [30]. The upper and lower boundary values of the parameters, such as k, Δt, n, and α are set to $[0.1, 0.25]$, $[0.1, 1]$, $[10, 170]$, and $[0.7, 2.5]$.

In the next step, statistical augmentation techniques [31] Such as rotation and vertical flip, are employed to increase the quantity of scalograms. Furthermore, a total of 9000 scalograms are generated that combine original and augmented scalograms. Afterward, the data are categorized as training, validation, and test data in a ratio of 70%:15%:15. Table 2 shows the complete information about the training, validation, and test before and after augmentation.

Table 2: Complete dataset description before and after augmentation

Number of recordings	Scalogram generated	Augmentation (Rotation, and vertical flip WITH ORIGINAL scalogram)	
3153	3000	9000	train (70%) = 6300 validation (15%) = 1350 test (15%) = 1350

4.1 Performance Evaluation Parameters

Table 3 presents the list of performance parameters. [32] That has been utilized for the quantitative assessment of the proposed MW-Net. These include accuracy (ACC), precision (PRC), recall (REC), F1-score (F1S), and intersection over union (IOU). These parameters are evaluated in terms of the confusion metrics component, which includes true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

Table 3: Performance evaluation parameters

Metrics	Formula
ACC	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$
PRC	$PR = \frac{TP}{TP + FP}$
REC	$REC = \frac{TP}{TP + FN}$
F1sc	$F1S = \frac{2 * PC * SEN}{PC + SEN}$
IoU	$\frac{\lvert t \cap t^* \rvert}{\lvert t \cup t^* \rvert'}$

4.2 Experimental Setup

The proposed MW-Net has been developed within a Python framework. It operates on an Intel Core Xeon (R) processor, equipped with 128 GB of RAM and an NVIDIA RTX A2000. Furthermore, the hyperparameters for the development of the proposed MW-Net are selected based on prior research. It includes the number of epochs (EPC), learning rate. (lr), batch size (BS), and optimizer. Research reveals that the majority of recent methodologies utilize Adam as an optimizer with a lr Ranging from 0.01 to 0.001. The majority of methodologies generally utilize a BS between 1 and 128, with EPC ranging from 15 to 1000. Table 4 summarizes such recent methodologies, which are adapted U-Net and W-Net. Furthermore, based on the observation, the selected MW-Net parameters for the training are as follows: epochs at 100, batch size at 4, learning rate at 0.0001, and optimizer as Adam.
The rationale for choosing the hyperparameters is as follows:
A batch size of 4 balances GPU memory constraints for high-resolution SSWT scalograms while providing stable gradients; a learning rate of 0.0001 with Adam is standard for segmentation tasks and gave stable convergence; 100 epochs with early stopping ensured convergence without overfitting. Moreover, the RSCDUM parameter bounds used in random search in the methods (RSCDUM) subsection are used to facilitate reproducibility.

Table 4: Performance Evaluation Parameters

Works	Hyperparameters			
	Epoch	bs	lr	optimizer
Park et al. [10]	30	128	0.0002	Adam
Renna et al. [13]	15	1	0.0001	Adam
Mukherjee et al. [16]	100	128	0.0005	Adam
Venegas et al. [17]	1000	4	0.0001	Adam
Singh et al. [33]	1000	-	0.01	SGD
This Work (MW-Net)	100	4	0.0001	Adam

4.3 Qualitative and Quantitative Analysis

The performance assessment of the proposed MW-Net has been presented in terms of qualitative and quantitative analysis. The comparative qualitative study has been demonstrated among existing HS segmentation approaches, such as U-Net and W-Net. Moreover, the performance study employs methods based on the energy envelope, such as 3rd and 4th order Shannon energy (SHE), along with time-frequency analysis methods, including HMM, to accomplish the comparative quantitative analysis.
Fig. 5 illustrates the complete steps involved in the development of the proposed MW-Net. The subfigure (a) shows the input PCG signal. Furthermore, the subfigure shows the scalogram, which is generated by SSWT. The subfigure (c) illustrates the outcome of the proposed RSCDUM, which enhances edge sharpness and suppresses irrelevant background textures. At the end, subfigure (d) demonstrates the final segmented heart sound obtained through the proposed MW-Net.

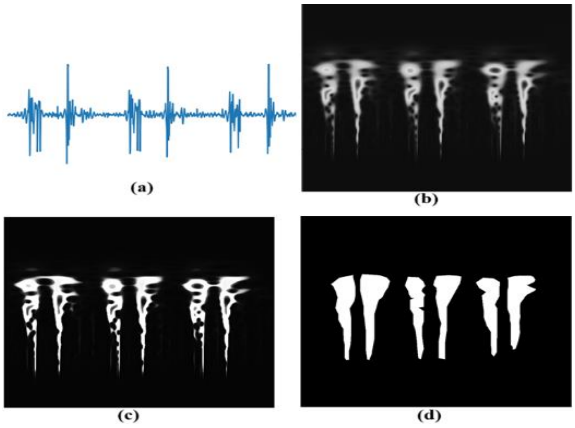

Fig 5: Results of steps involved where (a) is the input PCG signal; (b) is the scalogram via SSWT; (c) is the result of the proposed RSCDUM; and (d) is the segmented heart sound by the proposed MW-Net

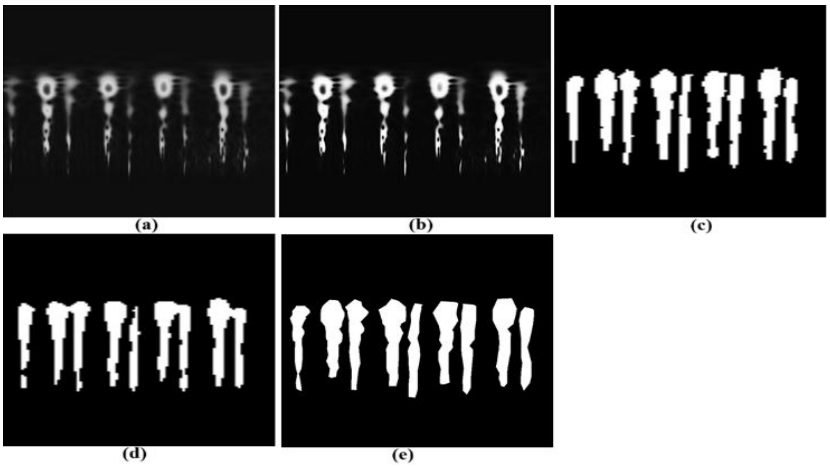

Fig 6: Comparative qualitative assessment where (a) is the input scalogram; (b) is the outcome of the proposed RSCDUM; (c), (d), and (e) are the segmented heart sound by the U-Net, W-Net, and the proposed MW-Net

Table 5: Comparative quantitative performance evaluation

Methods	Metrics				
	ACC (%)	PRC (%)	REC (%)	F1SC (%)	IOU (%)
3rd order SHE	91.42	94.19	96.88	95.53	91.42
4th order SHE	92.16	95.19	96.66	95.95	92.16
HMM	94.12	96.14	97.71	96.88	94.12
U-Net	92.89	95.46	97.18	96.33	92.89
Proposed	98.28	98.77	99.50	99.13	98.28

Table 6: Percentage Improvement of the Proposed Method in Accuracy

	ACC (%)	Δ Improvement Vs Proposed (%)
3rd order SHE	91.42	+7.50
4th order SHE	92.16	+6.64
HMM	94.12	+4.42
U-Net	92.89	+5.79
Proposed	98.28	-

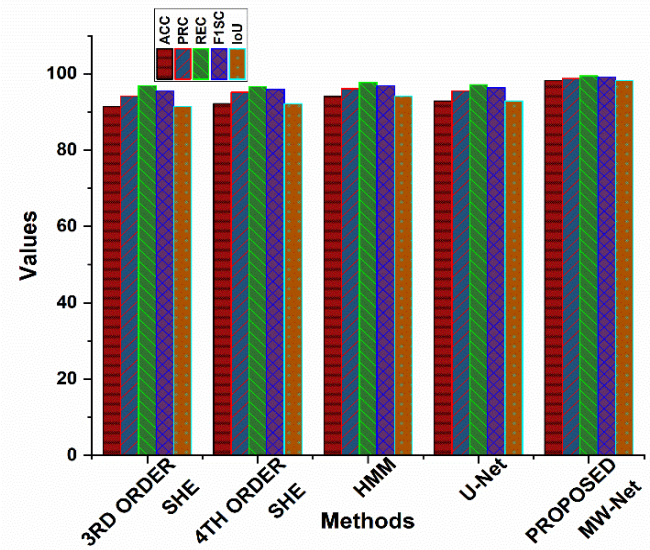

Fig 7: Comparative graphical bar plot of quantitative assessment for the data in Table 5

Fig. 6 presents a comparative qualitative evaluation of segmentation results obtained by U-Net, W-Net, and the proposed MW-Net. Sub-figures (a) and (b) of these figures show the input scalogram and the result of the proposed RSCDUM. Furthermore, the subfigures (c), (d), and (e) of these figures illustrate the segmentation mask obtained by U-Net, W-Net, and the proposed MW-Net. The observation reveals that the result of the U-Net suffers from over-segmentation and boundary irregularities. In contrast, the result of W-Net has limited inaccuracies at finer object contours. Apart from this, the segmentation mask obtained by the proposed MW-Net achieves precise boundary localization and structural consistency. This comparative analysis highlights the qualitative effectiveness of the proposed MW-Net, which ensures precise segmentation.

Table 5 exhibits a comparative quantitative evaluation among the existing segmentation approaches, such as 3rd-order SHE, 4th-order SHE, HMM, and U-Net, with the proposed MW-Net. The evaluation is assessed in terms of metrics presented in Table 3. The analysis illustrates that 3rd order SHE, 4th order SHE, HMM, and U-Net achieved an ACC of 91.42%, 92.16%, 94.12%, and 92.89%. In contrast to these, the proposed MW-Net significantly outperformed all baselines, achieving 98.28% ACC, 98.77% PRC, 99.50% REC, 99.13% F1SC, and 98.28% IoU. Fig. 7 shows the comparative graphical bar plot of Table 5. These results indicate that the proposed framework reduces false positives and false negatives. Moreover, it provides precise segmentation of heart sounds. Table 6 presents a comparative evaluation of performance improvement. It is evaluated with respect to accuracy. The proposed MW-Net exhibits a relative improvement

over 3rd-order SHE, 4th-order SHE, HMM, and U-Net by 7.50%, 6.64%, 4.42%, and 5.79%. This analysis entails the generalizability of the proposed method over existing segmentation approaches.

4.4 Generalizability and Robustness

The validation on a single public dataset limits the claim of broad generalizability. To mitigate this, we performed the following analysis and discussion:

a) **Noise robustness:** RSCDUM enhances edge contrast and reduces background texture, which empirically improves segmentation under moderate noise.
b) **Cross-domain considerations:** PCG recordings vary by device, auscultation site, and patient population. Therefore, it is recommended for cross-dataset validation (e.g., ARCA23K, MITHS datasets) as the next step; MW-Net's image-based approach is amenable to transfer learning and fine-tuning on small target datasets.
c) **Model adaptation:** For domain shift, simple strategies such as fine-tuning only decoder layers, domain-specific normalization, or augmenting SSWT parameters can be effective and are discussed as practical steps.

4.5 Computational complexity and deployment considerations

The computational scaling of RSCDUM and MW-Net is as follows:

a) **RSCDUM cost:** RSCDUM involves iterative diffusion updates (Eq. (8)); its runtime scales approximately with $O(n \cdot H \cdot W)$per image where n is the diffusion iteration count. Random search adds a multiplicative cost proportional to the number of sampled parameter sets; it has been constrained search budget to make RSCDUM feasible in pre-processing (e.g., tens to low hundreds of trials).
b) **MW-Net cost:** MW-Net is two cascaded U-Nets; inference cost is roughly twice that of a single U-Net of comparable width. Memory and compute scale with the number of channels and the scalogram resolution. To enable deployment, we discuss practical strategies: model pruning, knowledge distillation, mixed-precision inference (FP16), and quantization. For edge deployment, we suggest candidate hardware (NVIDIA Jetson Orin/NX, Google Coral TPU, or small industrial GPUs) and outline expected trade-offs between latency and accuracy.

5. Conclusion and Future Work

This paper presented a deep MW-Net framework for the segmentation and detection of heart sounds. The proposed framework employed SSWT for 1-D PCG signal conversion to 2-D scalograms. Furthermore, an RSCDUM was proposed and applied to redress challenges such as noise and low contrast in the generated scalograms. Apart from this, the Mish activation function was integrated at the architecture level to refine the efficacy of the proposed MW-Net. The effectiveness of the proposed MW-Net was assessed visually and statistically.
The performance was evaluated for the openly accessible PhysioNet/Computing in Cardiology Challenge 2016 PCG dataset. The proposed MW-Net achieved ACC, PRC, REC, F1SC, and IoU values of 98.28%, 98.77%, 99.50%, 99.13%, and 98.28%, respectively. The performance analysis revealed that the proposed deep framework exhibits effectiveness against signal variability and noise, along with a precise heart sound segmentation of S_1 and S_2.

5.1 Limitations and Future Directions:

There are some limitations of the proposed MN-Net model regarding cross-dataset validation, robustness benchmarking, real-time, and edge deployment, which will be focused on in future work as follows:

a) **Cross-dataset validation:** MW-Net will be evaluated on additional datasets (ARCA23K, other institutional collections) and report transfer learning protocols to quantify domain generalization.
b) **Robustness benchmarking:** A systematic noise-robustness experiment (additive white noise, device-specific artifact modeling) will be planned to report SNR-vs-performance curves.
c) **Real-time and edge deployment:** For real-time inference, a compressed MW-Net variant via pruning/quantization and test it on candidate edge devices (e.g., NVIDIA Jetson Xavier NX / Orin NX or Coral Edge TPU) will be developed to report latency and memory footprints.
d) **Multimodal fusion:** Integration with ECG (R-peak alignment) and other Biomedical signals will be explored. Moreover, an experiment will be done with late fusion and attention-based fusion modules to assess classification gains for abnormal sounds.
e) **Explainability and clinician evaluation:** Visual saliency or attention maps will be incorporated, and clinician feedback will be sought on failure cases to guide model refinement.

Moreover, future work will also focus on extending this framework for abnormal sound classification, real-time inference on edge hardware, and integration with multimodal data such as ECG for enhanced clinical utility.

References

[1] 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, 2017, doi: 10.1109/TBME.2016.2559800.
[2] P. Joseph et al., "cardiovascular disease in the Americas: the epidemiology of cardiovascular disease and its risk factors," Lancet Reg. Heal. - Am., vol. 42, p. 100960, 2025, doi: 10.1016/j.lana.2024.100960.
[3] M. D. Teja and G. M. Rayalu, "Hybrid time series and machine learning models for forecasting cardiovascular mortality in India: an age specific analysis," BMC Public Health, vol. 25, no. 1, 2025, doi: 10.1186/s12889-025-23318-7.
[4] T. H. Chowdhury, K. N. Poudel, and Y. Hu, "Time-Frequency Analysis, Denoising, Compression, Segmentation, and Classification of PCG Signals," IEEE Access, vol. 8, pp. 160882–160890, 2020, doi: 10.1109/ACCESS.2020.3020806.
[5] S. A. Alali et al., "Optimized CNN-based denoising strategy for enhancing longitudinal monitoring of heart failure," Comput. Biol. Med., vol. 184, no. November 2024, p. 109430, 2025, doi: 10.1016/j.compbiomed.2024.109430.
[6] M. Nath, S. Srivastava, N. Kulshrestha, and D. Singh, "Detection and localization of S1 and S2 heart sounds by 3rd order normalized average Shannon energy envelope algorithm," Proc. Inst. Mech. Eng. Part H J. Eng. Med., vol. 235, no. 6, pp. 615–624, 2021.
[7] Y. Jang et al., "Fully Convolutional Hybrid Fusion Network with Heterogeneous Representations for Identification of S1 and S2 from Phonocardiogram," IEEE J. Biomed. Heal. Informatics, vol. 28, no. 12, pp. 7151–7163, 2024, doi: 10.1109/JBHI.2024.3431028.

[8] M. Nath and S. Srivastava, "4th Order Shannon Energy Envelope Approach for Localization of S1 and S2 for Early-Stage Detection of Heart Valve Dysfunction," Trait. du Signal, vol. 40, no. 2, pp. 479–490, 2023, doi: 10.18280/ts.400207.

[9] Y. Chen, Y. Sun, J. Lv, B. Jia, and X. Huang, "End-to-end heart sound segmentation using deep convolutional recurrent network," Complex Intell. Syst., vol. 7, no. 4, pp. 2103–2117, 2021, doi: 10.1007/s40747-021-00325-w.

[10] C. Park et al., "Enhancement of phonocardiogram segmentation using convolutional neural networks with Fourier transform module," Biomed. Eng. Lett., vol. 15, no. 2, pp. 401–413, 2025, doi: 10.1007/s13534-025-00458-8.

[11] Q. Liu, X. Wu, and X. Ma, "An automatic segmentation method for heart sounds," Biomed. Eng. Online, vol. 17, no. 1, pp. 1–22, 2018, doi: 10.1186/s12938-018-0538-9.

[12] P. Xiao and K. Wang, "Segmentation of Heart Sound Signals Using Improved Hilbert Transform and Wavelet Packet Transform," Circuits, Syst. Signal Process., vol. 44, no. 7, pp. 4752–4773, 2025, doi: 10.1007/s00034-025-03000-4.

[13] F. Renna, J. Oliveira, and M. T. Coimbra, "Deep Convolutional Neural Networks for Heart Sound Segmentation," IEEE J. Biomed. Heal. Informatics, vol. 23, no. 6, pp. 2435–2445, 2019, doi: 10.1109/JBHI.2019.2894222.

[14] 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, 2018, doi: 10.1109/TBME.2018.2843258.

[15] T. Fernando, H. Ghaemmaghami, S. Denman, S. Sridharan, N. Hussain, and C. Fookes, "Heart Sound Segmentation Using Bidirectional LSTMs with Attention," IEEE J. Biomed. Heal. Informatics, vol. 24, no. 6, pp. 1601–1609, 2020, doi: 10.1109/JBHI.2019.2949516.

[16] A. Mukherjee, R. Banerjee, and A. Ghose, "A Novel U-Net Architecture for Denoising of Real-world Noise Corrupted Phonocardiogram Signal," 2023, [Online]. Available: http://arxiv.org/abs/2310.00216.

[17] V. M. Venegas et al., "Automated Phonocardiogram Segmentation a 1D U-Net Convolutional Neural Network : A Binary Approach," pp. 0–12, 2025, doi: 10.20944/preprints202501.2087.v1.

[18] G. D. Clifford et al., "Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016," Comput. Cardiol. (2010)., vol. 43, pp. 609–612, 2016, doi: 10.22489/cinc.2016.179-154.

[19] I. Daubechies, J. Lu, and H.-T. Wu, "Synchrosqueezed Wavelet Transforms: a Tool for Empirical Mode Decomposition," pp. 1–23, 2009, [Online]. Available: http://arxiv.org/abs/0912.2437.

[20] I. W. Selesnick and C. Sidney Burrus, "Generalized digital butterworth filter design," IEEE Trans. Signal Process., vol. 46, no. 6, pp. 1688–1694, 1998, doi: 10.1109/78.678493.

[21] R. Srivastava and S. Srivastava, "Restoration of Poisson noise corrupted digital images with nonlinear PDE based filters along with the choice of regularization parameter estimation," Pattern Recognit. Lett., vol. 34, no. 10, pp. 1175–1185, 2013, doi: 10.1016/j.patrec.2013.03.026.

[22] A. Kumar and S. Srivastava, "Restoration and enhancement of breast ultrasound images using extended complex diffusion based unsharp masking," Proc. Inst. Mech. Eng. Part H J. Eng. Med., p. 09544119211039317, 2021.

[23] G. Gilboa, N. Sochen, and Y. Y. Zeevi, "Image enhancement and denoising by complex diffusion processes," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 8, pp. 1020–1036, 2004, doi: 10.1109/TPAMI.2004.47.

[24] Y. F. Pu, "Fractional-Order Euler-Lagrange Equation for Fractional-Order Variational Method: A Necessary Condition for Fractional-Order Fixed Boundary Optimization Problems in Signal Processing and Image Processing," IEEE Access, vol. 4, pp. 10110–10135, 2016, doi: 10.1109/ACCESS.2016.2636159.

[25] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," J. Mach. Learn. Res., vol. 13, pp. 281–305, 2012.

[26] A. Dutta and A. Zisserman, "The {VIA} Annotation Software for Images, Audio and Video," 2019, doi: 10.1145/3343031.3350535.

[27] X. Xia and B. Kulis, "W-Net: A Deep Model for Fully Unsupervised Image Segmentation," 2017, [Online]. Available: http://arxiv.org/abs/1711.08506.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, pp. 234–241.

[29] D. Misra, "Mish: A Self Regularized Non-Monotonic Activation Function," 31st Br. Mach. Vis. Conf. BMVC 2020, 2020, doi: 10.5244/c.34.191.

[30] A. Kumar, P. Kumar, and S. Srivastava, "A skewness reformed complex diffusion based unsharp masking for the restoration and enhancement of Poisson noise corrupted mammograms," Biomed. Signal Process. Control, vol. 73, no. August 2021, p. 103421, 2022, doi: 10.1016/j.bspc.2021.103421.

[31] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[32] P. Kumar, A. Kumar, S. Srivastava, and Y. Padma Sai, "A novel bi-modal extended Huber loss function based refined mask RCNN approach for automatic multi instance detection and localization of breast cancer," Proc. Inst. Mech. Eng. Part H J. Eng. Med., p. 09544119221095416.

[33] K. R. Singh, A. Sharma, and G. K. Singh, "W-Net: Novel Deep Supervision for Deep Learning-based Cardiac Magnetic Resonance Imaging Segmentation," IETE J. Res., vol. 69, no. 12, pp. 8960–8976, 2023, doi: 10.1080/03772063.2022.2098836.