IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 May 2023, accepted 26 June 2023, date of publication 5 July 2023, date of current version 22 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3292551

RESEARCH ARTICLE

An End-to-End Deep Learning Framework for Real-Time Denoising of Heart Sounds for Cardiac Disease Detection in Unseen Noise

SHAMS NAFISA ALI[1], SAMIUL BASED SHUVO[1], MUHAMMAD ISHTIAQUE SAYEED AL-MANZO[2], ANWARUL HASAN[3,4], (Member, IEEE), AND TAUFIQ HASAN[1,5], (Senior Member, IEEE)

[1]mHealth Laboratory, Department of Biomedical Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka 1205, Bangladesh
[2]Department of Cardiac Surgery, National Heart Foundation Hospital and Research Institute, Mirpur, Dhaka 1216, Bangladesh
[3]Department of Mechanical and Industrial Engineering, Qatar University, Doha, Qatar
[4]Biomedical Research Center, Qatar University, Doha, Qatar
[5]Center for Bioengineering Innovation and Design, Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

Corresponding authors: Taufiq Hasan (taufiq@bme.buet.ac.bd) and Anwarul Hasan (ahasan@qu.edu.qa).

This work was supported by the Open Access funding provided by the Qatar National Library.

ABSTRACT The heart sound signals captured via a digital stethoscope are often distorted by environmental and physiological noise, altering their salient and critical properties. The problem is exacerbated in crowded low-resource hospital settings with high noise levels which degrades the diagnostic performance. In this study, we present a novel deep encoder-decoder-based denoising architecture (LU-Net) to suppress ambient and internal lung sound noises. Training is done using a large benchmark PCG dataset mixed with physiological noise, i.e., breathing sounds. Two different noisy datasets were prepared for experimental evaluation by mixing unseen lung sounds and hospital ambient noises with the clean heart sound recordings. We also used the inherently noisy portion of the PASCAL heart sound dataset for evaluation. The proposed framework showed effective suppression of background noises in both unseen real-world data and synthetically generated noisy heart sound recordings, improving the signal-to-noise ratio (SNR) level by 5.575 dB on an average using only 1.32 M parameters. The proposed model outperforms the current state-of-the-art U-Net model with an average SNR improvement of 5.613 dB and 5.537 dB in the presence of lung sound and unseen hospital noise, respectively. LU-Net also outperformed the state-of-the-art Fully Convolutional Network (FCN) by 1.750 dB and 1.748 dB for lung sound and unseen hospital noise conditions, respectively. In addition, the proposed denoising method model improves classification accuracy by 38.93% in the noisy portion of the PASCAL heart sound dataset. The results presented in the paper indicate that our proposed architecture demonstrated a robust denoising performance on different datasets with diverse levels and characteristics of noise. The proposed deep learning-based PCG denoising approach is a pioneering study that can significantly improve the accuracy of computer-aided auscultation systems for detecting cardiac diseases in noisy, low-resource hospitals and underserved communities.

INDEX TERMS Heart sound, real time denoising, deep learning, denoising autoencoder, cardiovascular diseases, cardiac disease detection.

I. INTRODUCTION

Cardiovascular diseases (CVDs) continue to be one of the leading causes of morbidity and mortality, claiming approximately 17.9 million lives every year [1]. The CVD-related death toll is ever on the rise, reaching about 32% of global deaths [1]. With inadequate facilities, insufficiency of trained physicians, and a lack of proper diagnostic equipment, the residents of the developing and under-developed countries are more vulnerable to CVD-related casualties [2]. Early diagnosis and treatment can be helpful in alleviating the adverse outcomes of CVDs. Heart sounds, generated by the mechanical activity of the heart valves during blood flow, reveal many diagnostically important information regarding

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Huang.

the physiological condition of the heart [3]. Thus, the phonocardiogram (PCG) is of paramount importance, being an effective, non-invasive, and comparatively inexpensive way for preliminary screening of CVDs [4]. Nevertheless, inter-listener variability and subjectivity in the interpretation limit its applicability.

The recent advancements in embedded systems and smartphone technologies have enabled the deployment of artificial intelligence-based computer-aided cardiac auscultation frameworks in point-of-care locations and, thereby, have the potential to improve healthcare delivery substantially. However, reliable and objective assessment of CVDs through these computerized methods is still a challenge. High susceptibility to motion artifact during signal acquisition, intrinsic spectral overlap among heart sound and lung sound (breathing sound), noise due to power supply interference, ambient noise in a hospital (door opening/closing/knocking, phone ringing, movement and speech of other patients in the outpatient department (OPD)), distortions introduced by the variation in device functionality (diaphragm, sensor, amplifier) affect the quality of PCGs and may even mask the presence of abnormalities in the perceived signal [3]. These factors can further degrade the performance of the AI-based automated diagnostic methods in real-life scenario [5]. Therefore, these factors must be considered crucial when designing a robust computerized auscultation tool for aiding the healthcare professionals in the CVD-screening process.

Over the past few decades, denoising noisy auscultation sounds have been thoroughly explored. Several statistical, time-domain and frequency-domain techniques have been investigated to denoise heart sounds [5], [6], [7] and lung sounds [8], [9]. Discrete wavelet transform (DWT) [10], empirical mode decomposition (EMD) [11], variational mode decomposition (VMD) [6], combination of singular value decomposition (SVD) and compressed sensing [12], non-negative matrix factorization (NMF) with adaptive contour representation computation (ACRC) from corresponding spectrogram [7] are some of the best performing approaches reported so far. Nevertheless, these techniques are often computationally intensive, time-consuming, and highly dependent on the data, predefined basis functions, the number of decomposition levels, thresholding parameters, and types [5], [10]. Moreover, most of these methods are evaluated using controlled clinical settings or simulated noisy conditions that oversimplify the irregular, unpredictable non-additive transient distortions from multiple sources [10]. All these data dependency factors and low protocol versatility concertedly make the existing methods of heart sound denoising a herculean task for the non-homogeneous real-world data.

Deep learning-based methods have garnered much attention and become the mainstream in a variety of applications and diverse branches of biomedical engineering [20], [23], [25], [29], [30]. In the same vein, a considerable amount of research work has been directed towards automated CVD screening through AI-assisted PCG segmentation [31], [32] and classification frameworks [4], [33], [34], [35], [36] due to the availability of multiple publicly available large PCG datasets. The inhomogeneity introduced by the inherent noises for automated PCG classification has been taken into consideration by a very few recent studies [36] however, development of dedicated deep learning-based denoising algorithms aiming to improve the diagnostic performance still remains unexplored. To the best of the knowledge of the authors, only a single work exists in the literature that has utilized deep learning-based architectures, i.e., 2D U-Net and denoising convolutional neural network (DnCNN) for denoising heart sound signals [37]. However, the work [37] lacks relevance from a clinical perspective since only random Gaussian noise was considered for synthetically corrupting heart sounds. The framework was not validated for any real-life noisy data. Additionally, the network performs in the time-frequency domain i.e., the signal sequence needs to be reshaped into a 2D vector for using the model weight on it and, again, reshaped into a 1D vector for final auditory inference. Thus, there is a significant buffering time between the input and the inference, which makes it unsuitable for applying in real-life scenarios. Furthermore, the denoising method was not evaluated to determine its impact on PCG classification performance. Practical denoising algorithms may introduce distortions that can degrade classification performance, and thus, such evaluations are crucial. Irrespective of these shortcomings, considering this work as the stepping stone and being motivated by the successful implementation of deep learning based methods for denoising speech and other non-stationary time-series signals (see a summary in Table 1), especially physiological signals like electrocardiogram (ECG), electroencephalogram (EEG), photoplethysmography (PPG), this paper aims to address the challenges of designing a robust computerized heart auscultation tool by proposing a comprehensive, resilient, end-to-end deep learning framework.

In this work, we propose an end-to-end deep learning framework for the real-time denoising of noisy heart sound recordings corrupted by real-world noises, i.e., lung sounds and hospital ambient noise. We aim to ensure that our framework retains the signal morphology, especially the murmurs, the primary indicators of cardiac abnormality, despite severe distortions in the input. Our network design is inspired by the fundamental nature of heart sounds and their applicability in a real-world setting. The heart sound signal is quasi-periodic as it is generated at regular intervals by sequential opening and closing of heart valves as blood flows through heart chambers [38]. Therefore, we hypothesize that a recurrent module in the network will improve denoising performance compared to standalone convolutional models because of its capability to identify temporal relationships. A flow diagram of the proposed framework is shown in Fig. 1. The main contributions of this paper are summarized below:

- To the best of our knowledge, we are the first group to propose a robust deep encoder-decoder based real-time PCG denoising framework named LU-Net, which

TABLE 1. Summary of the recent deep learning based works on speech and physiological signal denoising.

Speech Denoising			
Author (Year)		Noise Type	Network
Kong et al. [13] (2022)		Several kinds of environmental noise, music from 3 benchmark dataset	CleanUNet (encoder-decoder architecture with self-attention blocks)
Alamdari et al. [14] (2021)		Babble, wind, engine, driving car	FCN with Noisy2Noisy signal mapping
Liu et al. [15] (2020)		10 types of noise (2 artificial and 8 from a benchmark databaset)	CP-GAN (Context Pyramid Generative Adversarial Network)
Pandey et al. [16] (2019)		Babble, cafeteria noise with other different non-speech sounds	Temporal convolutional neural network
Rethage et al. [17] (2018)		Environmental noise (conditions in a park, a bus or a cafe)	Wavenet (CNN with non-causal, dilated convolutions)
Physiological Signal Denoising			
Author (Year)	Signal Type	Noise Type	Network
Pouyani et al. [18] (2022)	Lung sound	Gaussian white noise	DWT-ANN
Aghaomidi et al. [19] (2022)	ECG	MA, BW, EM (I, II), RN	DeepRTSNet (2D Encoder-Decoder)
Kiranyaz et al. [20] (2022)	ECG	Clean and corrupted samples from the CPSC-2020 dataset	1D Cycle-GAN
Chuang et al. [21] (2022)	EEG	Eye, muscle, heart, channel noise,	IC-U-Net (1D U-Net-based DAE) other artifacts (during driving, walking)
Bing et al. [22] (2021)	ECG	MA, BW, EM (II)	DeepCEDNet (2D Encoder-Decoder)
Guan et al. [23] (2021)	ECG	MA, BW, EM (II)	LDTF (Low-dimensional denoising embedding transformer)
Yi et al. [24] (2021)	EEG	Ocular & muscle artifact	EEGDnet (2D Transformer)
Sawangjai et al. [25] (2021)	EEG	Ocular & muscle artifact with different movements using a single upper extremity	EEGANet (GAN)
Singh et al. [26] (2020)	ECG	MA, BW, EM (II), RN	CNN-GAN
Sun et al. [27] (2020)	EEG	Ocular & muscle artifact, ECG noise	1D-ResCNN (CNN with parallel residual blocks)
Chiang et al. [28] (2019)	ECG	MA, BW, EM (II)	FCN-based DAE (1D)
Lee et al. [29] (2018)	PPG	RN, Saturation noise, Sloping noise	Bidirectional Recurrent Auto-Encoder

* MA - Muscle artifacts, BW - Baseline wander, EM - electrode motion, RN - Random noise, DAE - Denoising autoencoder, ANN - Artificial neural network, FCN - Fully convolutional network, I- Single noise types applied separately, II - Three kinds of noise (BW, EM, MA) are mixed with equal weight to form a complex noise.

utilizes U-Net as a backbone with LSTM modules to capture the spatial and temporal features, precisely the quasi-periodic information embedded in the PCG waveform. The model enhances the perceptual quality by attenuating real-life noise components present in PCG recordings during cardiac auscultation of any raw PCG recording without preprocessing or hand-crafted feature extraction. An expert cardiologist conducted a blind informal auditory test to confirm the model's effectiveness.

- We also proposed a dedicated SNR estimation scheme for real-life PCG signals contaminated with irregular, unpredictable, multi-source distortions in low-resource setting hospitals since obtaining the clean ground truth of the corresponding noise signal is not feasible in such scenarios. The proposed scheme was verified

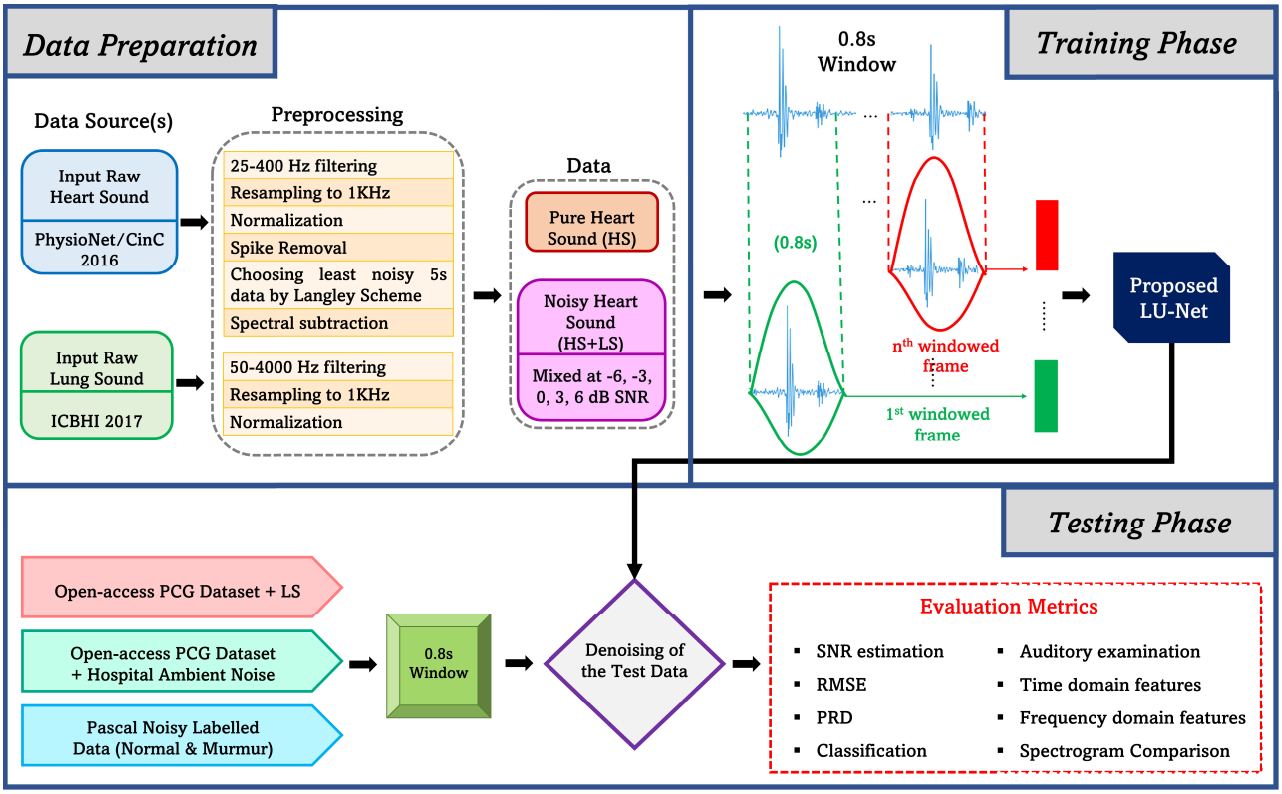

FIGURE 1. A graphical overview of the end-to-end denoising workflow. After several generic pre-processing steps, pure Heart sound (HS) and Lung sound (LS) have been obtained which are mixed at particular SNRs to synthetically generate noisy HS data (distorted with LS). Frames of 0.8s from each of the recordings are successively passed to the proposed architecture. This sums up the training phase. In the testing phase, using the training weight, three completely unseen dataset(s) (synthetically generated noisy data with LS and hospital ambient noise as well as real-life noisy data) have been denoised and evaluated using multiple relevant metrics.

using the signal quality label annotations found in the literature.

- We conducted extensive experiments to assess the robustness of our proposed model on multiple blind test datasets with varying degrees and types of noise. To challenge the model on multiple levels, we used two types of synthetically generated noisy PCGs with different SNR levels, including lung sound noise and hospital ambience noise (which was not encountered during training). Additionally, we used ground truth PCGs from a different dataset to evaluate the model's ability to generalize and restore salient PCG features from a noisy input, regardless of the PCG source. Finally, we evaluated the model's performance on a benchmark dataset containing real-life noisy PCGs from both normal and diseased (murmur) conditions.
- We conducted a study to assess whether our proposed model affects the AI-based state-of-the-art CVD classification model to evaluate whether the characteristics properties of heart sound required for differential diagnosis persist after denoising it using the proposed model. A deep encoder-decoder-based real-time PCG denoising framework is designed to attenuate the noise components present in PCG recordings during cardiac auscultation.

The remainder of this paper is organized as follows. In Section II, we highlight the types of noises that majorly degrade the quality of PCG, Section III-B contains the detailed description of our proposed method. The data resources used are described in Section IV. Section V provides a detailed overview of the experimentation schemes, i.e., training and test data preparation, baseline systems, and evaluation metrics used. Section VI contains the qualitative and quantitative analysis of the results obtained by the proposed method as well as the baseline systems. The effect of the loss function and computational efficiency on the performance of the proposed system is also discussed in this section. In Section VII, we discuss the implications and limitations of our method, highlighting the future directions, and finally, in Section VIII we summarize our findings with a conclusive remark.

II. BACKGROUND

In the case of heart sounds, one of the primary causes of signal deterioration is the impact of noise from several sources, such as lung sounds caused by breathing, hospital ambient noise, human conversation, intestinal activity, stethoscope movements, sensor variability, and so forth [3]. Susceptibility to noise is a matter of concern in the automated evaluation of cardiac disorders, especially in low-resource settings. This section discusses the origin and

properties of the different distortions present in heart sound recordings.

A. LUNG SOUND NOISE

Lung and cardiac auscultation are two critical bio-signals for cardiorespiratory diagnosis. Lung sounds are produced throughout the respiratory cycle when air enters and exits the airways, while heart sounds are produced by the heart valve, leaflets, and blood movement in cardiac chambers [38].

The frequency ranges of the two main components of normal heart sound recordings, namely the first (S1) and second (S2) heart sounds, are between 20 and 150 Hz [39], while components of murmurs may be heard between 30Hz and 700Hz [40]. On the other hand, typical lung sounds span between 50 and 1000 Hz, while tracheal sounds vary between 850 and 1000 Hz. Abnormal lung sounds, such as wheeze and crackle, may vary in frequency from 400-1600Hz and 100-500Hz, respectively [41], [42]. Pure cardiac auscultation signals are usually not accessible due to the fact that both signals originate from proximal anatomical sites [38]. Thus, the measured signals are often a mixture of heart and lung sounds, and they are characterized by intrusive quasi-periodic interference in both the temporal and spectral domains [43]. Even experienced cardiologists are sometimes perplexed by the diastolic murmur-shaped breathing sounds present in a typical HS recording, coupled with low pitched wheeze and bronchial breathing [36]. Therefore, the diagnostic quality of heart sound auscultation may be greatly improved by suppressing respiration signals from heart sound signals.

B. ENVIRONMENTAL AND STETHOSCOPE MOTION NOISE

Due to the transitory nature of heart sound signals, they are susceptible to irregular ambient noise components that are randomly distributed in the time and frequency domain, such as children wailing in the background, people talking, hospital activity sounds, street announcement sounds, etc. [43]. Another potential source of noise is the sliding motions of the stethoscope diaphragms as a result of the physician relocating the recording spot, the patient becoming agitated, or accidental displacement [44]. These noise components are often of high amplitude, last for a short duration, and introduce substantial time-frequency overlap between heart sound signals [43]. These effects add to the complexity of the analysis and make it more cumbersome for the physicians to extract relevant diagnostic information. Therefore, noise reduction would improve the reliability of quantitative and qualitative assessments using heart sound signals.

III. PROPOSED APPROACH

A. PROBLEM FORMULATION

When the pure, noise-free PCG signal is corrupted with several irrelevant components coming from the environment or system, a noisy signal is formed as follows:

$$y = x + n \tag{1}$$

where $y \in \mathbb{R}^{1\times N}$ (N is the number of samples in the signal sequence) represents an acquired noisy PCG signal, $x \in \mathbb{R}^{1\times N}$ denotes a noise-free PCG signal (theoretical) and $n \in \mathbb{R}^{1\times N}$ denotes the noisy signal components that are additively integrated with the noise-free PCG signal and degrade their quality.

Algorithm 1 LU-Net

Input : $X_t = \{x_t^i\}_{i=1,\ldots,n}$, $Y_t = \{y_t^i\}_{i=1,\ldots,n}$ - ground-truth and noisy PCG signal in the training dataset, respectively;
$X_v = \{x_v^i\}_{i=1,\ldots,m}$, $Y_v = \{y_v^i\}_{i=1,\ldots,m}$ - ground-truth and noisy signal in the validation dataset, respectively.
Output: $F(y;\theta)$ - an optimized network
Loss: $L(x,\hat{x})$ - mean squared error

Initialize θ weights
repeat
  Acquire noise-free prediction $\hat{x}_t = F(y_t;\theta)$
  Calculate training loss between prediction and ground-truth = $L(x_t,\hat{x}_t)$
  Update θ using Adam optimizer with respect to the loss $L(x_t,\hat{x}_t)$
  Acquire noise-free prediction $\hat{x}_v = F(y_v;\theta)$
  Calculate validation loss between prediction and ground-truth = $L(x_v,\hat{x}_v)$
  Preserve θ if $L(x_v,\hat{x}_v)$ improves, ignore if Step=1
until *loss* $L(x_t,\hat{x}_t)$ *converges*
return $F(y;\theta)$

$$\hat{x} = F(y;\theta) \tag{2}$$

A deep learning-based end-to-end signal denoising model implies that giving a noisy PCG signal y as the input will provide the corresponding noise-free one $\hat{x}$ as the output. This is achieved by constructing a highly complex nonlinear mapping function $F(\cdot)$ i.e., neural network and training its learnable parameter sets θ to minimize the disparity between the estimated noise-free PCG signal $\hat{x}$ and corresponding noise-free one x using a suitable objective function (reconstruction loss). This phenomenon can be formulated as shown in (2):

In this paper, we leverage the widely-adopted Mean Square Error (MSE) as the loss function, $L(x,\hat{x})$ to train the proposed architecture by small batch gradient descent method and gradually minimize the loss by Adam optimization method. The training pipeline and the associated optimization process is shown in Algorithm 1.

B. PROPOSED MODEL ARCHITECTURE

The proposed network is a convolutional encoder-decoder-based architecture with bi-directional long short term memory (Bi-LSTM) modules in the skip connections, as illustrated in Fig. 2. It is characterized by the encoder and decoder layers (each of the encoder or decoder layer is represented by the subscript, $i = 1$ to 5), the bottleneck and the output

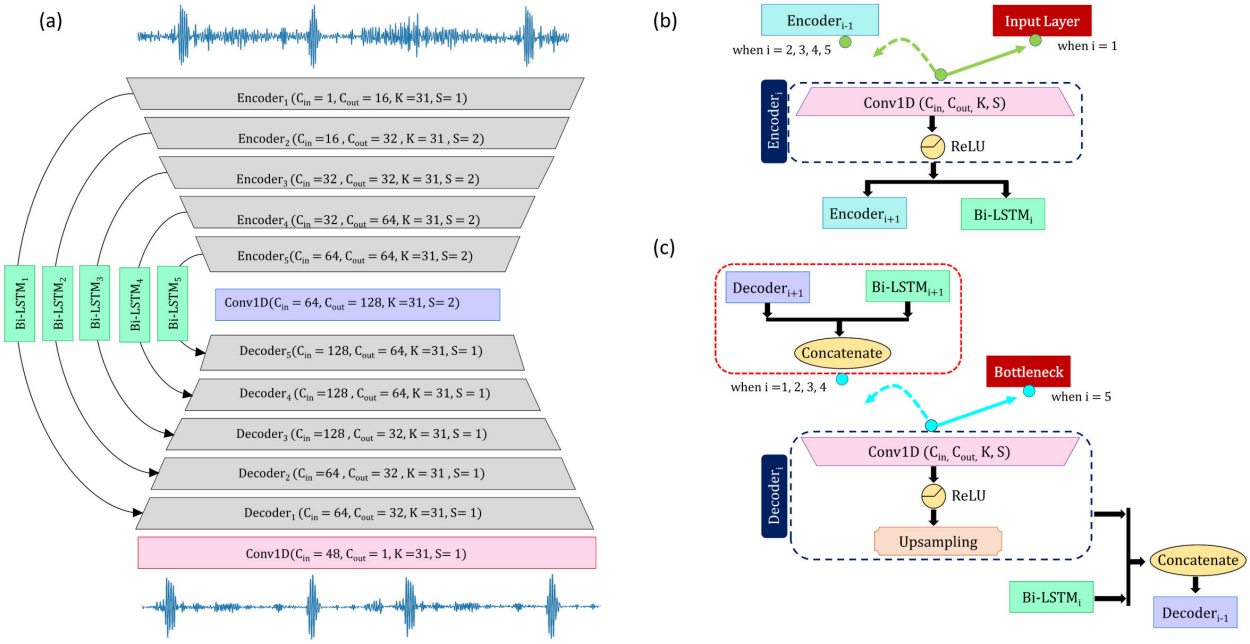

FIGURE 2. **(a) Proposed LU-Net architecture with the noisy PCG as input on the top and the denoised PCG as output on the bottom. (b) Detailed representation of the encoder and decoder layers in conjunction with the corresponding Bi-LSTMs in the skip connections.**

convolutional layer. Each encoder and decoder layer is defined by the input (C_{in}) and output (C_{out}) dimensions along the frame length, kernel size (K) and stride (S).

As input, the first encoder layer ($i = 1$), $Encoder_1$ receives a noisy input frame, x_t while the rest of the encoder layers, $Encoder_i$ for $i = 2$ to 5, receive the output from the previous encoder layer, $Encoder_{i-1}$ as input. Each encoder layer contains a 1D convolution layer with a kernel size of K and stride of S with a defined number output channels, C_{out} followed by a ReLU activation. The output feature maps resulting from $Encoder_i$ is fed into Bi-$LSTM_i$ and $Encoder_{i+1}$ (passed to bottleneck i.e., latent representation for $Encoder_5$). Since $Encoder_{i=2\text{-}5}$ contain convolution layer with a stride of 2, they successively create lower dimensional representation of the noisy input frame along the compression path. The $Decoder_i$ consists of a 1D convolution layer followed by a ReLU non-linearity activation and an UpSampling1D layer. $Decoder_5$ directly inputs the feature maps generated by the bottleneck. The rest of the decoder layers, $Decoder_{i=1\text{-}4}$ receive the concatenated features from $Decoder_{i+1}$ and $LSTM_{i+1}$ as input. Thereafter, the output from $Decoder_i$ is concatenated with Bi-$LSTM_i$ and fed into the $Decoder_{i-1}$. Because of the presence of UpSampling layer, the receptive field gradually expands while propagating through the expansion path. Finally, the output from $Decoder_1$ is passed through a convolution layer, where $C_{out} = 1$ which provides the corresponding denoised output sequence, $\hat{y}_t$. For a detailed visual description of the encoder and decoder, see Fig. 2(b).

In the decoder phase of a typical U-Net, the bottleneck is gradually expanded by upsampling or transposed convolution. The receptive field of the bottleneck is small; thus, when this approach is directly used, the network tends to lose salient low-level information. To minimize this loss, we incorporate the Bi-LSTM module in the proposed framework as it can internally concatenate the forward and backward vectors to a single vector that captures all the hidden attributes present in a PCG frame. Its inherent non-causality also enables the model to learn the long-term dependencies with fewer parameters. In the proposed architecture, Bi-$LSTM_i$ has 8, 16, 16, 32, 32 units, respectively for $i = 1$ to 5. For all of them, `he_Normal` kernel initializer [45] and hyperbolic tangent (tanh) activation have been employed, and `return_sequence` have been set to be TRUE so that consistent and homogeneous output of the same length can be obtained. The proposed model will be referred to as LSTM U-Net (LU-Net) in the rest of the paper.

IV. DATA RESOURCES

A. 2016 PhysioNet/CinC HEART SOUND (PHS) DATASET

The 2016 PhysioNet/CinC challenge dataset [3] is an openly accessible cross-corpus archive of PCG recordings collected by seven different research groups from a total number of 764 subjects in either clinical or non-clinical settings. It contains 3240 PCG recordings with 84,425 cardiac cycles ranging from 35 to 159 bpm. The recordings are of varying duration (5s-120s) and are collected from six different clinical settings. Apart from domain variance, i.e., in terms of acquisition device and recording set-up, the presence of several noises (e.g., breathing, stethoscope movement, intestinal activity, peripheral talking, etc.) in the recordings make the dataset suitable for designing noise-robust algorithms.

B. PASCAL HEART SOUND CHALLENGE DATASET

The PASCAL dataset [46] contains data from 2 sources having varying duration, between 1-30s. Dataset-A consists of 176 recordings collected via the iStethoscope Pro iPhone app at a frequency of 44.1kHz, while Dataset-B contains 656 files collected in a clinical setting using the DigiScope stethoscope at 4kHz. In Dataset-A, there are four categories: Normal, Murmur, Extra Heart Sound, and Artifact. Dataset-B consists of 3 classes: Normal, Murmur, and Extrasystole. However, in the training set of Dataset-B, there are sub-directories containing noisy data of normal (120) and murmur (29). Apart from these classes, both datasets contain some unlabelled data.

C. OPEN-ACCESS HEART SOUND (OAHS) DATASET

The open-access heart sound dataset (OAHS dataset) [47] is a publicly available noise-free PCG dataset containing a total number of 1000 recordings. Five classes, i.e., Normal (N), Aortic stenosis (AS), Mitral regurgitation (MR), Mitral stenosis (MS), Mitral valve prolapse (MVP), are annotated, each class containing 200 recordings. The recordings are sampled at 8kHz and have varying duration.

D. ICBHI 2017 DATASET

The International Conference on Biomedical Health Informatics (ICBHI) 2017 dataset is the largest publicly available respiratory sound database [48]. Two independent research teams from Portugal and Greece have collected 920 audio samples from 126 subjects at different sampling frequencies (4kHz, 10kHz, and 44.1kHz). The total recording duration is 5.5 hours, while each data length varies between 10-90s. The dataset contains 6898 respiratory cycles annotated by respiratory experts either as normal or having respiratory anomalies, namely, wheeze, crackle, and wheeze and crackle. The dataset also includes labels regarding the subject's pathological condition, i.e., healthy, and seven distinct disease classes, namely Bronchiectasis, Bronchiolitis, Chronic Obstructive Pulmonary Disease (COPD), Asthma, Pneumonia, Upper Respiratory Tract Infection (URTI), and Lower Respiratory Tract Infection (LRTI), along with their corresponding collection site. Further details about the dataset can be found in [49].

E. HOSPITAL AMBIENT NOISE (HAN) DATASET

This dataset has been prepared using a non-copyrighted YouTube video[1] of 68 minutes where the audio occurrences were recorded from different places (corridor, waiting room, etc.) of a busy hospital. By manually selecting the noisy portions, 562 segments of audio, each of 5s duration, have been prepared, which are made available in Kaggle [50]. Since the major fequency components of the chunks are found within 500 Hz, the data samples are filtered by applying a 3rd order Butterworth low pass filter having cut off at 500 Hz. Then, it is resampled to 1000 Hz, following the Nyquist criteria so that it can be used as a noise signal with the heart sound while testing the robustness of the denoising framework.

V. EXPERIMENTAL EVALUATION

We trained our model using the PHS dataset contaminated with lung sound and tested it on an entirely distinct heart sound dataset, the OAHS dataset, which was also corrupted with lung sound data that differed significantly from the ones used for training. Moreover, we evaluated the performance of our model on hospital ambient noise, which was entirely unfamiliar to the model. We also tested the model on real-life noise-contaminated heart sounds from the PASCAL dataset. The experimental design, data preprocessing, implementation details and evaluation metrics are described in the following subsections. Codes and representative samples are available at `Github`.[2]

A. TRAINING DATA PREPARATION

Training of the denoising model is done using heart sound recordings from the PHS dataset. All samples have been resampled at 1kHz following the application of a 3rd order Butterworth bandpass filter with a passband of 25Hz to 400Hz. A spike removal algorithm [51] is used to remove impulsive components from the signals. All signals are normalized to the range $[-1, 1]$ to reduce amplitude variability. Due to the presence of significant noise in the initial portion of some signals, the cleanest 5s segment is extracted from each recording using a wavelet entropy-based automatic cleanest segment selection algorithm introduced by Langley et al. [52]. Next, six non-overlapping frames of 0.8s is extracted from each 5s segment (explained in VI-D). These frames are not taken at any specific event, such as S1 or S2, but rather at random points along a PCG cycle, continuing until six samples are collected.

We used lung sounds from the ICBHI 2017 dataset as the noise source to create synthetic noisy PCG recordings. Lung sounds are filtered with a 6th order Butterworth bandpass filter with upper and lower cut-off frequencies of 50 and 2500 Hz [53], respectively, followed by a 1kHz resampling step and min-max normalization. Among the various lung sound auscultation sites, only sounds from the Anterior right (Ar), Posterior right (Pr), and Lateral right (Lr) positions were used since these locations are least contaminated with heart sound components.

In each of the 0.8s PCG segments, a lung sound is synthetically added with SNR values of −6dB, −3dB, 0dB, 3dB, 6dB. A total of 93480 noisy frames and corresponding noise-free PCG frames are used to train the denoising model as inputs and outputs, respectively, by retaining the train-validation split of the original PHS dataset [3].

B. TEST DATA PREPARATION

A two-way testing protocol is considered for synthetic and real-world noisy conditions. The relatively clean OAHS

[1] https://youtu.be/3LUuyDdWOy4

[2] https://github.com/mHealthBuet/Heart-Sound-Denoising

dataset recordings are mixed with lung sound and hospital ambient noise to generate two synthetic noisy test sets, OAHS-LS and OAHS-HAN, respectively. To represent the real-world test scenario, we used the noisy recordings of the PASCAL dataset, which were corrupted by different sources during data collection. We ensure that none of the test samples or their corresponding noise recordings are used during training.

All the HS datasets have been resampled to 1kHz, followed by a min-max normalization to the range $[-1, 1]$. PCG signals of the PASCAL dataset (only 149 samples labeled as noisy) are truncated from the start of the recording up to 2.4s. On the other hand, audio signals from the OAHS dataset have been padded to a length of 3.5s to account for the irregular length [54]. Each PCG recording is divided into 0.8s segments for processing. Lung sounds from the ICBHI 2017 are processed in the same manner as in training.

C. DATA PREPARATION FOR CLASSIFICATION

To examine the influence of denoising on classification performance, we utilized PASCAL and OAHS datasets. First, we trained and validated the classification model using clean PCG signals from the normal and murmur categories in the PASCAL dataset. We then tested the model's classification performance using only the noisy samples from the same dataset and their corresponding denoised version. On the other hand, we have partitioned the OAHS dataset into three distinct sets: training, validation, and test, with a ratio of 70 : 10 : 20. The test portion has been mixed with lung sound and hospital ambient noise to generate the test OAHS-LS and OAHS-HAN datasets, respectively. These two PCG datasets are processed in the same manner as done in the test data preparation for enhancement V-B.

D. BASELINE SYSTEMS

The Fully Convolutional Network (FCN) [55] and U-Net [56] are popular deep learning architectures typically utilized for image-to-image transformation. Nevertheless, their 1D versions are frequently employed in audio-to-audio transformation, such as denoising, enhancement, and suppression tasks in other fields [21], [28], [57]. Since audio denoising is essentially a task of audio-to-audio transformation, 1D variants of U-Net and FCN are selected as baseline systems.

In this work, inspired by [28], an FCN-based denoising autoencoder has been constructed as a baseline system (baseline-1). In addition, a 1D U-Net is constructed as a second baseline (baseline-2), where the encoder maps the input data into a lower-dimensional representation while the decoder reconstructs the input data from this representation. The encoder is composed of repetitions of a convolutional layer followed by a rectified linear unit (ReLU) activation layer that imposes non-linearity to the feature maps extracted by the filters of the convolutional layers. The decoder path is formed by mirroring the encoder layers in the reverse order. Each layer in the decoder is followed by a ReLU layer. As the audio is gradually downsampled followed by upsampling, it results in a rapid increase in the receptive field that is convenient for the propagation of global time and frequency information stored in the audio sequence. For ensuring homogeneity and a fair basis of comparison, input frame length, kernel size, stride size, the number layers and corresponding hyper-parameters of LU-Net are retained in the baselines.

TABLE 2. Hyper-parameters of the proposed denoising framework.

Hyper-parameters	Values
Batch size	128
Learning rate	0.0001
Epoch	100
Optimizer	Adam
Loss function	Mean square error

E. EXPERIMENTAL SETUP

The deep learning architectures are implemented using TensorFlow and Keras, while all the models are trained and tested on Intel(R) Xeon(R) CPU and NVidia K80 GPUs provided by Kaggle notebooks. Mean Square Error (MSE) and Sparse categorical cross-entropy are used as the loss function for training the denoising and classification models, respectively.

The adaptive learning rate optimizer (Adam) with an initial learning rate (lr) of 10^{-4} and batch size of 128 are utilized for training both models. Due to the better efficacy of Adam compared to the other optimizers, this stochastic momentum based approach was chosen to accelerate the model training [58].

A mini Batch balancing scheme [4] is employed to pass an equal number of samples from each class on all the batches during classification model training. A summary of the considered hyperparameters are listed in Table 2.

F. EVALUATION METRICS FOR DENOISING

The denoising performance is evaluated using true SNR, estimated SNR, percent root mean square difference (PRD), and root-mean squared error (RMSE), as detailed below.

1) TRUE SNR METRIC

SNR is the primary metric for assessing the noise reduction performance. It is defined as follows:

$$SNR(dB) = 10\log_{10}\left(\frac{P_s}{P_n}\right) = 10\log_{10}\left(\frac{\Sigma_{t=1}^{N} x[t]^2}{\Sigma_{t=1}^{N} n[t]^2}\right) \quad (3)$$

where, P_s and P_n represent signal and noise power, respectively. Also, $x[t]$ and $n[t]$ indicate the t^{th} sample of the signal and noise, respectively, while N denotes the total number of samples in the signal and noise.

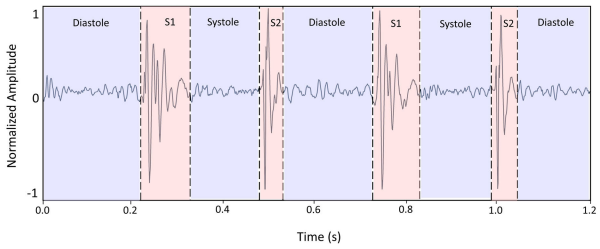

FIGURE 3. A typical PCG signal with four states of the cardiac cycle (S1, systole, S2, diastole). The pink and blue shaded areas contribute to the noisy signal power and the noise power, respectively.

2) ESTIMATED SNR METRIC

To calculate true SNR using (3), the clean and the noise signal must be known. It is thus impossible to calculate the true SNR in real-life noisy PCG signals as we only have access to the noisy signal. We propose an SNR estimation algorithm designed particularly for heart sounds in such cases. The method is described below.

A typical heart sound signal is first segmented into four main regions: S1, systole, S2, and diastole (Fig. 3). We assume that the systole and diastole regions only contain background noise, while the S1 and S2 regions contain both signal and background noise. Using these assumptions, we estimate the noise power, noisy signal power and signal power as follows.

$$P_n = \frac{\sum_{t=1}^{N_{\text{sys}}} x_{\text{sys}}[t]^2}{N_{\text{sys}}} + \frac{\sum_{t=1}^{N_{\text{dias}}} x_{\text{dias}}[t]^2}{N_{\text{dias}}} \tag{4}$$

$$P_{ns} = \frac{\sum_{t=1}^{N_{\text{S1}}} x_{\text{S1}}[t]^2}{N_{\text{S1}}} + \frac{\sum_{t=1}^{N_{\text{S2}}} x_{\text{S2}}[t]^2}{N_{\text{S2}}} \tag{5}$$

$$P_s = P_{ns} - P_n \tag{6}$$

Here, P_n, P_{ns} and P_s represent the estimated noise power, noisy signal power, and signal power, respectively. The signals $x_{\text{sys}}[t]$, $x_{\text{dias}}[t]$, $x_{\text{S1}}[t]$ and $x_{\text{S2}}[t]$ indicate the PCG signal segments for the systole, diastole, S1 and S2 regions, respectively. The corresponding lengths of these signals are indicated by the variables N_{sys}, N_{dias}, N_{S1}, and N_{S2}, respectively.

To validate the estimated SNR metric, we perform an experiment using the signal quality labels provided in [59] for the PHS dataset. We hypothesize that if the estimated SNR metric is a reliable estimate of PCG signal quality, it will strongly correlate with the manually annotated quality measures. In [59], the quality of a subset of PCG signals was manually annotated with 5 labels: very bad (1), bad (2), borderline (3), good (4) and excellent (5). We calculated the mean and standard deviation of the proposed estimated SNR for each quality label, while the Pearson coefficient and coefficient of determination were calculated between quality labels and the average estimated SNRs. The results illustrated in Fig. 4 show that the proposed SNR estimate indeed has a strong correlation with the subjective quality measure of PCG signals. Thus, we justify using this estimated SNR metric for the quality evaluation of intrinsically noisy PCG signals.

G. ROOT-MEAN-SQUARED ERROR (RMSE)

RMSE is generally used to assess the deviation between intended and actual signals. A lower RMSE indicates a smaller difference. We use the following formulation to calculate the RMSE between the clean and denoised HS recordings:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(x[t] - \hat{x}[t])^2} \tag{7}$$

where, $x[t]$ and $\hat{x}[t]$ denote the clean and denoised PCG signals, respectively, while N denotes the signal length.

H. PERCENT ROOT-MEAN-SQUARED DIFFERENCE (PRD)

The PRD metric denotes recovery efficiency by comparing the input signal to the reconstructed signal. Lower PRD indicates a superior reconstruction. PRD is calculated as:

$$PRD = \sqrt{\frac{\sum_{t=1}^{N}(x[t] - \hat{x}[t])^2}{(x[t])^2}} \tag{8}$$

where, $x[t]$ and $\hat{x}[t]$ represent the clean and denoised PCG signals, respectively, while the signal length is given by N.

I. EVALUATION METRICS FOR CLASSIFICATION

We hypothesize that improving the PCG signal quality through denoising will also improve its automatic classification performance. To evaluate this hypothesis, we perform classification experiments using the OAHS-LS, OAHS-HAN and PASCAL datasets using the model proposed in [60]. The classification performance is assessed using the accuracy metric, which now acts as a secondary quantitative evaluation metric for the proposed PCG denoising method.

VI. EXPERIMENTAL RESULTS

A. OBJECTIVE PERFORMANCE EVALUATION FOR DENOISING

This section describes the evaluation of the proposed denoising scheme compared to the baseline models primarily using the SNR metric. In the case of the synthetically generated noisy PCG datasets (OAHS-LS, OAHS-HAN), we also use the PRD, and RMSE metrics for comparison. The results are summarized in Tables 3 and 4, and Fig. 5.

The left column of Fig. 5 demonstrates evaluated performance on the OAHS-LS dataset where the PCG signals are mixed with lung sound noise. Here, we observe that LU-Net consistently outperforms FCN and U-Net across all evaluated metrics. The improvement in output SNR, PRD, and RMSE, at all SNR levels indicates that the proposed network reconstructs noise-free PCG signals more accurately compared to the baseline methods. While FCN outperforms the U-Net model in terms of SNR improvement (Fig. 5(a)) and RMSE (Fig. 5(e)) at low SNRs, U-Net performs better at higher SNR levels. However, this trend is not evident in PRD, where FCN performs inferior to U-Net for all the input SNR levels (Fig. 5(c)). Overall, compared to the baseline

TABLE 3. Denoising Performance Comparison between FCN, U-Net and LU-Net architectures on OAHS for Lung sound (OAHS-LS).

	Output SNR (dB)			PRD			RMSE		
Input SNR (dB)	FCN	U-Net	Proposed LU-Net	FCN	U-Net	Proposed LU-Net	FCN	U-Net	Proposed LU-Net
-6	1.608	0.084	**2.095**	0.917	0.859	**0.816**	0.145	0.159	**0.137**
-3	2.910	2.724	**3.984**	0.869	0.803	**0.752**	0.126	0.129	**0.111**
0	4.107	4.710	**5.846**	0.805	0.721	**0.669**	0.110	0.103	**0.091**
3	5.042	6.423	**7.447**	0.730	0.620	**0.577**	0.099	0.085	**0.077**
6	5.649	7.710	**8.695**	0.661	0.527	**0.493**	0.094	0.074	**0.067**
Avg.	3.863	4.330	**5.613**	0.796	0.706	**0.661**	0.115	0.110	**0.097**

TABLE 4. Denoising Performance Comparison between FCN, U-Net and LU-Net architectures on OAHS for Hospital noise (OAHS-HAN).

	Output SNR (dB)			PRD			RMSE		
Input SNR (dB)	FCN	U-Net	Proposed LU-Net	FCN	U-Net	Proposed LU-Net	FCN	U-Net	Proposed LU-Net
-6	1.044	-1.374	**1.069**	0.926	0.737	**0.633**	0.145	0.200	**0.151**
-3	2.988	1.131	**3.540**	0.879	0.680	**0.572**	0.123	0.151	**0.115**
0	4.207	4.060	**5.991**	0.815	0.620	**0.532**	0.108	0.108	**0.088**
3	5.069	6.769	**7.896**	0.742	0.550	**0.485**	0.099	0.080	**0.072**
6	5.641	8.671	**9.194**	0.674	0.480	**0.431**	0.094	0.065	**0.063**
Avg.	3.789	3.851	**5.537**	0.807	0.613	**0.530**	0.114	0.121	**0.098**

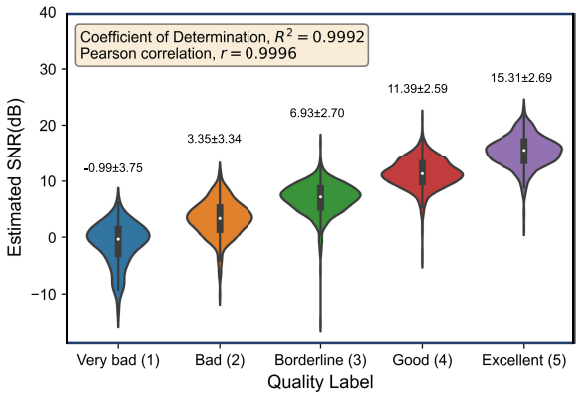

FIGURE 4. Validation of the proposed SNR estimation scheme for real-life PCG signals with embedded noise. The obtained SNRs justify the corresponding signal quality. A highly qualified physician and two senior researchers with combined knowledge in heart sound signal processing conducted these high-quality label annotations. Each recording was given a quality label value between 1 and 5. The final label was created by averaging the annotations provided in [59].

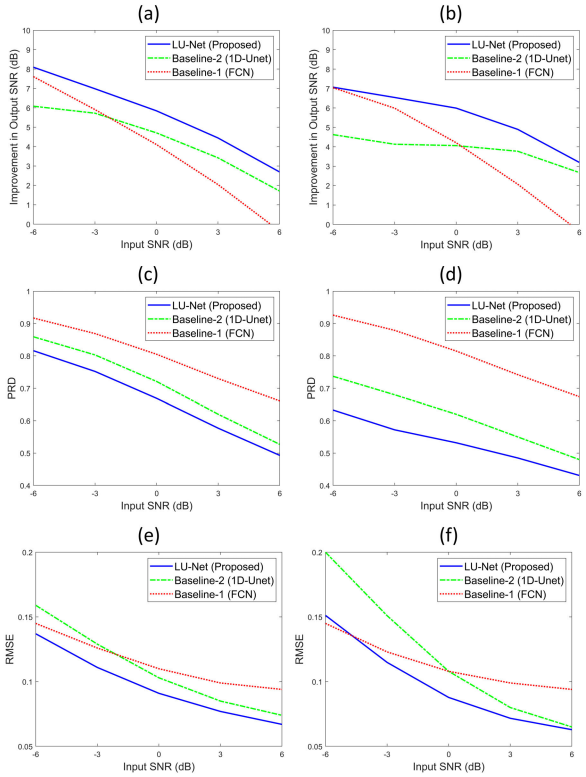

FIGURE 5. Comparison of the denoising performance of all the evaluated methods for OAHS-LS (left) and OAHS-HAN (right) dataset. (a) & (b) Improvement in output SNR at varying input SNR levels, (c) & (d) PRD at varying input SNR levels, (e) & (f) RMSE at varying input SNR levels.

models across all input SNRs, the proposed LU-Net improves SNR by an average of 5.613dB, which is superior to FCN and U-Net models by 1.750dB and 1.283dB, respectively. In addition, LU-Net achieved an average reduction in RMSE by 2.138% and 1.356%, and an average reduction in PRD by 3.112% and 4.474%, when compared to FCN and U-Net, respectively. The performance metrics for different input SNR levels are summarized in Table 3. These experimental results demonstrate the effectiveness of the proposed denoising method in suppressing lung sound noise.

In the case of the real-life noisy PCG recordings from the PASCAL dataset, since it is impossible to calculate the true SNR, we use the proposed heart sound SNR estimation scheme to evaluate the denoising performance. The results are reported in the first column of Table 6. In this case, the proposed LU-Net improves the estimated SNR by 6.517 dB, which is 26.175% and 2.725% superior relative to U-Net and FCN, respectively. We have already shown in the previous section that the estimated SNR closely correlates with PCG signal quality. Therefore, we conclude that the proposed method is effective in noise suppression, even in the case of intrinsically noisy PCG recordings.

In the right column of Fig. 5, the performance of the denoising methods is illustrated on the OAHS-HAN dataset, where the PCG signals were corrupted by unseen hospital noise. The results depict once again that the proposed

TABLE 5. Classification performance open-access heart sound dataset (OAHS dataset).

Applied Actual SNR (dB	Classification Performance (%Acc)			
	Lung noise		Hospital noise	
	Noisy	Denoised	Noisy	Denoised
-6	48.35	72.13	55.06	71.28
-3	58.31	83.11	70.98	83.17
0	73.04	90.32	83.55	88.62
3	80.91	92.09	83.33	87.92
6	84.76	92.66	87.27	89.40

TABLE 6. Denoising and classification performance on PASCAL dataset for real-world noise.

Performance metric	Noisy Signal	Denoised signal		
		FCN	U-Net	LU-Net
Estimated SNR (dB)	7.456	13.635	10.348	14.017
% Accuracy	26.17	60.40	54.34	**65.10**

LU-Net outperforms FCN and U-Net on all metrics studied, including output SNR, PRD, and RMSE for all input SNR levels, except for input noise −6dB, where FCN shows marginally improved performance compared to LU-Net with respect to RMSE (Fig. 5(f)). Compared to the two baseline models, the proposed method improves the SNR by 5.537 dB on average across all input SNRs, which is 1.748 dB and 1.686 dB higher than the FCN and U-Net models, respectively. In addition, when compared to FCN and U-Net models, the proposed LU-Net yields an average reduction in RMSE of 1.918% and 2.295%, and a reduction in PRD of 2.600% and 8.300%, respectively. Interestingly, for this type of noise, FCN outperforms U-Net at low SNRs in terms of SNR improvement (Fig. 5(b)) and RMSE (Fig. 5(f)) but significantly fall short of U-Net and the proposed LU-Net at higher SNRs. Finally, in terms of the PRD performance metric, the proposed LU-Net method provided the best performance over the entire range of input SNR values (Fig. 5(d)). Detailed results for the hospital noise experiments are also provided in Table 4. Therefore, we may conclude that the proposed LU-Net model is effective in suppressing previously unseen hospital noise present in PCG recordings.

B. SUBJECTIVE PERFORMANCE EVALUATION FOR DENOISING

Using 15 randomly chosen noisy samples from PASCAL dataset and their corresponding denoised data, a blind informal auditory test is performed with the help of an expert cardiologist. The original and the corresponding denoised recordings were randomly ordered in a list. The cardiologist provided each recording a score between 1-5 depending on its subjective quality, with a 1 being described as "diagnostically poor quality HS with very intrusive background noise" and a 5 being "diagnostically excellent quality HS with unnoticeable background noise". To ensure unbiased assessment, the cardiologist was not informed of the list having noisy and denoised version of the same data. From the mean scores for original raw recordings (1.60) and the corresponding denoised recordings (3.53), it is confirmed that the derived denoised signals are improved in terms of subjective quality. This, in turn, proves that, our network is well-capable of suppressing noise components and it is significantly beneficial in terms of diagnosis.

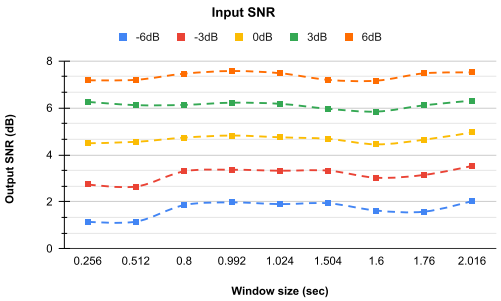

FIGURE 6. SNR improvement at different input noise levels with respect to different window lengths using PHS validation dataset. A window size of 0.8s demonstrates decent denoising performance with minimal latency.

C. CLASSIFICATION RESULTS

The classification performance has been evaluated for both the noisy and the denoised recordings to assess the impact of denoising for AI-based CVD detection. Experiments have been conducted using the OAHS-LS, OAHS-HAN, and PASCAL datasets. From the results summarized in Table 6, it can be observed that the denoised signals attained an average classification accuracy improvement of 17.00% and 6.67% over noisy signals with lung sound and hospital noise, respectively. In the case of the noisy PASCAL dataset, the denoised signals outperformed the raw PCG signal in classification accuracy by a large margin of about 38% (see Table 6), demonstrating the significance of the proposed LU-Net in terms of diagnostic interpretation. This dataset consists of real-world noisy heart sound signals from various sources, and thus, the performance improvement is noteworthy [46].

D. OPTIMIZATION OF WINDOW LENGTH

The selection of the processing window lengths is a vital system parameter for a real-time heart sound denoising framework. A larger window size might cause latency, whereas a shorter window size, on the other hand, may considerably reduce the model's denoising capabilities. Various denoising experiments with window lengths ranging from 0.032s to 2.016s are carried out to optimize this parameter. All of the experiments are done on the validation set of the PHS dataset using the baseline U-Net architecture. After comparing the SNR improvement, we have found that a window length of 0.8s provides a reasonable trade-off between latency and decent denoising performance.

E. INFERENCE SPEED AND EFFICIENCY

The proposed LU-Net contains 1.32 M trainable parameters, 206 M floating-point operations per second (FLOPS), and requires only 15.3 MB of memory. Thus, it achieves substantial real-time PCG signals denoising and a great reduction in storage and processing power requirements. Because the suggested paradigm can immediately improve

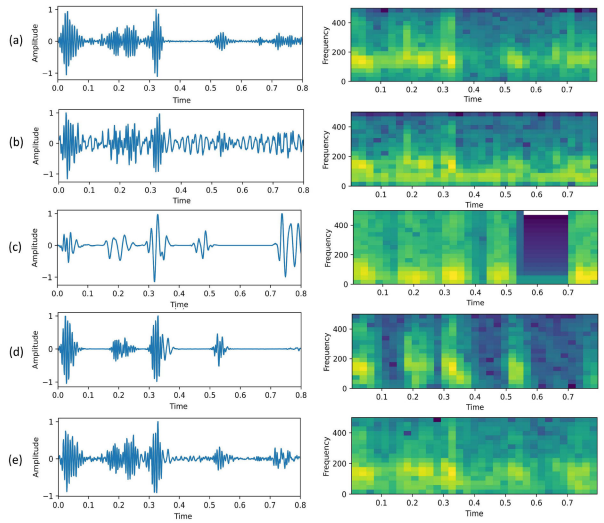

FIGURE 7. Waveforms and spectrograms of a typical abnormal PCG signal. (a) original signal, (b) noisy signal (HS+LS mixed at 3 dB SNR), denoised signals using (c) FCN (d) U-Net (e) LU-Net, respectively. LU-Net reconstructed the PCG with high-fidelity, preserving the S1, S2 and the systolic and diastolic murmurs.

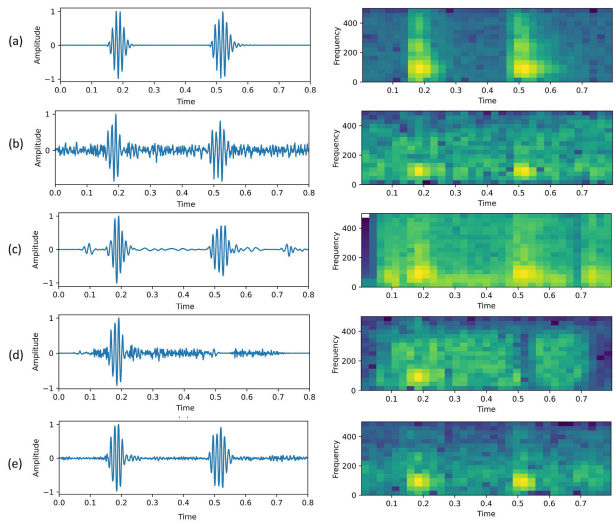

FIGURE 8. Waveforms and spectrograms of a typical normal PCG signal. (a) original signal, (b) noisy signal (HS+HAN mixed at 6 dB SNR), denoised signals using (c) FCN (d) U-Net (e) LU-Net, respectively. The proposed LU-Net shows superior performance in retaining the signal morphology.

raw PCG recordings without the need for pre-processing and a hand-crafted feature extraction approach, it proves the robustness of the model and makes it suitable for IoT integrated mobile devices and wearable sensors.

F. COMPARISON OF LOSS FUNCTIONS

While training the proposed network, we investigated three different loss functions: mean square error ($\mathcal{L}_{\mathbf{MSE}}$), negative signal-to-noise ratio ($\mathcal{L}_{\mathbf{SNR}}$) [61], and scale-invariant signal-to-distortion ratio ($\mathcal{L}_{\mathbf{SI\text{-}SDR}}$) [62]. Table 7 shows the SNR improvement, RMSE, and PRD scores over OAHS-HAN dataset for SNRs of 6, 3, 0, 3, and 6 dB. As we can see,$\mathcal{L}_{\mathbf{SI\text{-}SDR}}$ loss show promising results for low SNR values but degrades significantly at higher SNR values. A similar pattern is also found for $\mathcal{L}_{\mathbf{SNR}}$. On the other hand, $\mathcal{L}_{\mathbf{MSE}}$ exhibits consistent performance, SNR by 3.154 dB and 1.121 dB, while exhibiting 3.358% and 1.112% reduction in RMSE and 27.286% and 21.986% reduction in PRD, respectively, when compared to loss functions $\mathcal{L}_{\mathbf{SI\text{-}SDR}}$ and $\mathcal{L}_{\mathbf{SNR}}$.

VII. DISCUSSION

As demonstrated in Section VI, the proposed end-to-end framework has superseded the standalone baseline models in terms of all the quantitative metrics. Compared to the baselines, the proposed model consists of skip connections with bi-LSTM modules between the encoder and decoder, which are absent in both the baselines. We may attribute the improved results to these added modules. The incorporation of bi-LSTMs enables the model to better learn the rhythmic pattern of the quasi-periodic heart sound as it employs both the backward and forward information of the PCG sequence at every time instance. Preservation of salient rhythmic information from both past and future makes the reconstruction task comparatively easier for the model. The superiority of LU-Net over the baselines can be visually observed in Fig. 7 and 8. While the baseline enhancement schemes fail to reproduce the morphology of the signals accurately, the proposed LU-Net reconstructed them with high fidelity, preserving the S1, S2, and the characteristic murmurs (present only in Fig. 7).

Further interesting insights were revealed during the evaluation phase. The resulting improvement in SNR is found to be consistent for three completely unseen noisy PCG datasets. Despite being trained with only LS noise, the proposed LU-Net exhibited significant noise removal performance for both LS noise and unseen HAN. For lower SNRs (i.e., −6dB, −3dB), the improvement in SNR is more promising (see Table 3 & 4). A noticeable improvement is also observed for higher SNRs (i.e.,3dB, 6dB) in the case of both noise types. Nevertheless, the slightly better performance obtained while using HAN is possibly due to its aperiodic nature and the dissimilar spectral distribution compared to HS.

The experiments on synthetically generated noisy PCG establish the framework's effectiveness concerning the scenario of additive noise. The denoising experiment on the real-life PaHS dataset further enhances its relevance from the clinical perspective. At the same time, all the classification experiments reveal the proposed method's aptness to be used in conjunction with the existing computer-aided CVD diagnosis systems. The proposed framework is evaluated for a wide range of varieties in PCG, and the results obtained indicate its robustness and superiority compared to the existing methods; these concertedly justify its applicability, especially in challenging low-resource hospital settings.

However, there are few limitations to our study that should be taken into consideration when interpreting the results. One major limitation is the absence of a heart sound dataset

TABLE 7. Performance of different loss functions in PCG denoising system measured using various performance metrics.

Evaluation metrics	SNR imrovement			PRD			RMSE		
Loss Function	$\mathcal{L}_{\text{SI-SDR}}$	$\mathcal{L}_{\text{SNR}}$	$\mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{SI-SDR}}$	$\mathcal{L}_{\text{SNR}}$	$\mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{SI-SDR}}$	$\mathcal{L}_{\text{SNR}}$	$\mathcal{L}_{\text{MSE}}$
-6	7.186	**7.710**	7.069	0.879	0.860	**0.633**	0.147	**0.140**	0.151
-3	4.970	6.460	**6.540**	0.844	0.820	**0.572**	0.137	0.119	**0.115**
0	2.584	4.810	**5.991**	0.802	0.760	**0.532**	0.129	0.1	**0.088**
3	-0.025	2.740	**4.896**	0.762	0.690	**0.485**	0.123	0.095	**0.072**
6	-2.799	0.360	**3.194**	0.730	0.620	**0.431**	0.121	0.090	**0.063**

that contains both clean and noisy signals. As a result, we had to synthetically mix noise to train our encoder and decoder network. While we attempted to create a realistic noise mixture, it is possible that our synthetic dataset does not fully capture the complexity and variability of real-life noise. Another limitation is related to the dataset used in our study. We used the PHS dataset, which is currently the largest publicly available dataset of heart sounds. However, this dataset is highly imbalanced, with a large number of the samples containing only normal heart sounds. As a result, the encoder and decoder network may have been biased towards normal sounds and could potentially distort abnormal murmurs.

Although the proposed denoising framework has demonstrated satisfactory performance on multiple well-known PCG datasets representing real-life scenarios, we plan to further improve and generalize the denoising network, LU-Net's training by using a class-balanced in-house dataset. This dataset will be composed of a larger number of PCG recordings, with varying noise types and levels, to ensure better generalization of the proposed approach. Additionally, several advanced network strategies, such as Generative Adversarial Networks (GAN) and Transformers, can be attempted to obtain further optimized performances with more efficiency and non-latency. In the future, we aim to integrate the proposed LU-Net model into digital stethoscopes or wearable devices to perform automatic PCG denoising. This integration will enable real-time processing of PCG recordings and help clinicians in their diagnostic decision-making process, particularly in low-resource, noisy environments. Another interesting avenue for future research is the exploration of the effectiveness of the proposed approach to other medical signal denoising tasks, such as ECG, EEG or PPG denoising.

VIII. CONCLUSION

In this work, we have proposed a generalized, robust deep learning framework for real-time denoising of noisy PCG recordings, which is crucial in automatic heart sound abnormality detection. An SNR estimation scheme has also been proposed for quality assessment of real-world PCG data contaminated with irregular, multi-source, transient distortions. Upon experimentation with multiple unseen datasets with diverse levels and characteristics of noise, we have demonstrated that the proposed method is robust to different types of input noise. Compared to two state-of-the-art systems, FCN and U-Net, the proposed method has provided a relative SNR increase of 31.178% and 22.851%, respectively, on noisy cardiac sounds distorted by respiratory sounds. Moreover, 31.569% and 30.454% relative improvement in SNR was obtained in the case of unseen hospital noise conditions compared to FCN and U-Net systems, respectively. The proposed network has considerably fewer trainable parameters compared to existing models, which eventually elevates the potential of the model to be deployed in memory-constrained platforms for real-time applications. There are certain limitations in this study, such as the use of synthetic noise mixture training and the possibility of bias towards normal sounds caused by an imbalanced dataset. However, in the future, we aim to address these issues by enhancing and expanding the training of LU-Net through a class-balanced in-house dataset. Additionally, we intend to integrate this technology into digital stethoscopes or wearable devices to perform real-time automatic PCG denoising. Moreover, we plan to evaluate its effectiveness for denoising other medical signals in the future.

ACKNOWLEDGMENT

The Open Access funding provided by the Qatar National Library. *(Shams Nafisa Ali and Samiul Based Shuvo are co-first authors.)*

REFERENCES

[1] World Health Organization. (2021). *Cardiovascular Diseases (CVDs) Fact Sheet*. Accessed: Feb. 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] A. Rosengren, A. Smyth, and S. Rangarajan, "Socioeconomic status and risk of cardiovascular disease in 20 low-income, middle-income, and high-income countries: The prospective urban rural epidemiologic (PURE) study," *Lancet Global Health*, vol. 7, no. 6, pp. e748–e760, Jun. 2019.

[3] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, and A. E. Johnson, "An open access database for the evaluation of heart sound algorithms," *Physiological Meas.*, vol. 37, no. 12, pp. 2181–2213, Dec. 2016.

[4] A. I. Humayun, S. Ghaffarzadegan, Md. I. Ansari, Z. Feng, and T. Hasan, "Towards domain invariant heart sound abnormality detection using learnable filterbanks," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 8, pp. 2189–2198, Aug. 2020.

[5] D. Gradolewski and G. Redlarski, "Wavelet-based denoising method for real phonocardiography signal recorded by mobile devices in noisy environment," *Comput. Biol. Med.*, vol. 52, pp. 119–129, Sep. 2014.

[6] V. G. Sujadevi, N. Mohan, S. Sachin Kumar, S. Akshay, and K. P. Soman, "A hybrid method for fundamental heart sound segmentation using group-sparsity denoising and variational mode decomposition," *Biomed. Eng. Lett.*, vol. 9, no. 4, pp. 413–424, Nov. 2019.

[7] D. Pham, S. Meignen, N. Dia, J. Fontecave-Jallon, and B. Rivet, "Phonocardiogram signal denoising based on nonnegative matrix factorization and adaptive contour representation computation," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1475–1479, Oct. 2018.
[8] N. S. Haider, "Respiratory sound denoising using empirical mode decomposition, Hurst analysis and spectral subtraction," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102313.
[9] C. S. Lee, M. Li, Y. Lou, and R. Dahiya, "Restoration of lung sound signals using a hybrid wavelet-based approach," *IEEE Sensors J.*, vol. 22, no. 20, pp. 19700–19712, Oct. 2022.
[10] P. K. Jain and A. K. Tiwari, "An adaptive thresholding method for the wavelet based denoising of phonocardiogram signal," *Biomed. Signal Process. Control*, vol. 38, pp. 388–399, Sep. 2017.
[11] A. Gavrovska, M. Slavkovic, I. Reljin, and B. Reljin, "Application of wavelet and EMD-based denoising to phonocardiograms," in *Proc. Int. Symp. Signals, Circuits Syst. (ISSCS)*, Jul. 2013, pp. 1–4.
[12] Y. Zheng, X. Guo, H. Jiang, and B. Zhou, "An innovative multi-level singular value decomposition and compressed sensing based framework for noise removal from heart sounds," *Biomed. Signal Process. Control*, vol. 38, pp. 34–43, Sep. 2017.
[13] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7867–7871.
[14] N. Alamdari, A. Azarang, and N. Kehtarnavaz, "Improving deep speech denoising by Noisy2Noisy signal mapping," *Appl. Acoust.*, vol. 172, Jan. 2021, Art. no. 107631.
[15] G. Liu, K. Gong, X. Liang, and Z. Chen, "CP-GAN: Context pyramid generative adversarial network for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6624–6628.
[16] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6875–6879.
[17] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5069–5073.
[18] M. F. Pouyani, M. Vali, and M. A. Ghasemi, "Lung sound signal denoising using discrete wavelet transform and artificial neural network," *Biomed. Signal Process. Control*, vol. 72, Feb. 2022, Art. no. 103329.
[19] P. Aghaomidi, A. Mohammadisarab, J. Mazloum, M. A. Akbarzadeh, M. Orooji, N. Mokari, and H. Yanikomeroglu, "DeepRTSNet: Deep robust two-stage networks for ECG denoising in practical use case," *IEEE Access*, vol. 10, pp. 128232–128249, 2022.
[20] S. Kiranyaz, O. C. Devecioglu, T. Ince, J. Malik, M. Chowdhury, T. Hamid, R. Mazhar, A. Khandakar, A. Tahir, T. Rahman, and M. Gabbouj, "Blind ECG restoration by operational cycle-GANs," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 12, pp. 3572–3581, Dec. 2022.
[21] C.-H. Chuang, K.-Y. Chang, C.-S. Huang, and T.-P. Jung, "IC-U-Net: A U-Net-based denoising autoencoder using mixtures of independent components for automatic EEG artifact removal," *NeuroImage*, vol. 263, Nov. 2022, Art. no. 119586.
[22] P. Bing, W. Liu, and Z. Zhang, "DeepCEDNet: An efficient deep convolutional encoder–decoder networks for ECG signal enhancement," *IEEE Access*, vol. 9, pp. 56699–56708, 2021.
[23] J. Guan, W. Wang, P. Feng, X. Wang, and W. Wang, "Low-dimensional denoising embedding transformer for ECG classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1285–1289.
[24] P. Yi, K. Chen, Z. Ma, D. Zhao, X. Pu, and Y. Ren, "EEGDNet: Fusing non-local and local self-similarity for 1-D EEG signal denoising with 2-D transformer," 2021, *arXiv:2109.04235*.
[25] P. Sawangjai, M. Trakulruangroj, C. Boonnag, M. Piriyajitakonkij, R. K. Tripathy, T. Sudhawiyangkul, and T. Wilaiprasitporn, "EEGANet: Removal of ocular artifacts from the EEG signal using generative adversarial networks," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 4913–4924, Oct. 2022.
[26] P. Singh and G. Pradhan, "A new ECG denoising framework using generative adversarial network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 2, pp. 759–764, Mar. 2021.
[27] W. Sun, Y. Su, X. Wu, and X. Wu, "A novel end-to-end 1D-ResCNN model to remove artifact from EEG signals," *Neurocomputing*, vol. 404, pp. 108–121, Sep. 2020.
[28] H. Chiang, Y. Hsieh, S. Fu, K. Hung, Y. Tsao, and S. Chien, "Noise reduction in ECG signals using fully convolutional denoising autoencoders," *IEEE Access*, vol. 7, pp. 60806–60813, 2019.
[29] J. Lee, S. Sun, S. M. Yang, J. J. Sohn, J. Park, S. Lee, and H. C. Kim, "Bidirectional recurrent auto-encoder for photoplethysmogram denoising," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2375–2385, Nov. 2019.
[30] S. B. Shuvo, S. N. Ali, S. I. Swapnil, T. Hasan, and M. I. H. Bhuiyan, "A lightweight CNN model for detecting respiratory diseases from lung auscultation sounds using EMD-CWT-based hybrid scalogram," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2595–2603, Jul. 2021.
[31] X. Wang, C. Liu, Y. Li, X. Cheng, J. Li, and G. D. Clifford, "Temporal-framing adaptive network for heart sound segmentation without prior knowledge of state duration," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 2, pp. 650–663, Feb. 2021.
[32] S. Das, S. Pal, and M. Mitra, "Acoustic feature based unsupervised approach of heart sound event detection," *Comput. Biol. Med.*, vol. 126, Nov. 2020, Art. no. 103990.
[33] A. K. Dwivedi, S. A. Imtiaz, and E. Rodriguez-Villegas, "Algorithms for automatic analysis and classification of heart sounds—A systematic review," *IEEE Access*, vol. 7, pp. 8316–8345, 2019.
[34] S. Based Shuvo, S. Samiul Alam, S. Umme Ayman, A. Chakma, P. D. Barua, and U. Rajendra Acharya, "NRC-Net: Automated noise robust cardio net for detecting valvular cardiac diseases using optimum transformation method with heart sound signals," 2023, *arXiv:2305.00141*.
[35] S. K. Ghosh, R. N. Ponnalagu, R. K. Tripathy, and U. R. Acharya, "Automated detection of heart valve diseases using chirplet transform and multiclass composite classifier with PCG signals," *Comput. Biol. Med.*, vol. 118, Mar. 2020, Art. no. 103632.
[36] F. B. Azam, M. I. Ansari, S. I. S. K. Nuhash, I. McLane, and T. Hasan, "Cardiac anomaly detection considering an additive noise and convolutional distortion model of heart sound recordings," *Artif. Intell. Med.*, vol. 133, Nov. 2022, Art. no. 102417.
[37] T. S. Sharan, R. Bhattacharjee, S. Sharma, and N. Sharma, "Evaluation of deep learning methods (DnCNN and U-Net) for denoising of heart auscultation signals," in *Proc. 3rd Int. Conf. Commun. Syst., Comput. IT Appl. (CSCITA)*, Apr. 2020, pp. 151–155.
[38] R. Nersisson and M. M. Noel, "Heart sound and lung sound separation algorithms: A review," *J. Med. Eng. Technol.*, vol. 41, no. 1, pp. 13–21, Jan. 2017.
[39] H. Ren, H. Jin, C. Chen, H. Ghayvat, and W. Chen, "A novel cardiac auscultation monitoring system based on wireless sensing for healthcare," *IEEE J. Translational Eng. Health Med.*, vol. 6, pp. 1–12, 2018.
[40] K. Tsai, W. Wang, C. Cheng, C. Tsai, J. Wang, T. Lin, S. Fang, L. Chen, and Y. Tsao, "Blind monaural source separation on heart and lung sounds based on periodic-coded deep autoencoder," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 11, pp. 3203–3214, Nov. 2020.
[41] Y. Kim, Y. Hyon, S. S. Jung, S. Lee, G. Yoo, C. Chung, and T. Ha, "Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning," *Sci. Rep.*, vol. 11, no. 1, p. 17186, Aug. 2021.
[42] D. Emmanouilidou, E. D. McCollum, D. E. Park, and M. Elhilali, "Computerized lung sound screening for pediatric auscultation in noisy field environments," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 7, pp. 1564–1574, Jul. 2018.
[43] H. Tang, T. Li, Y. Park, and T. Qiu, "Separation of heart sound signal from noise in joint cycle frequency–time–frequency domains based on fuzzy detection," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 10, pp. 2438–2447, Oct. 2010.
[44] S. K. Ghosh, R. K. Tripathy, and P. R N, "Evaluation of performance metrics and denoising of PCG signal using wavelet based decomposition," in *Proc. IEEE 17th India Council Int. Conf. (INDICON)*, Dec. 2020, pp. 1–6.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE ICCV*, Dec. 2015, pp. 1026–1034.
[46] P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor. (2011). *The PASCAL Classifying Heart Sounds Challenge 2011*. [Online]. Available: http://www.peterjbentley.com/heartchallenge/index.html
[47] *Classification of Heart Sound Signal Using Multiple Features*. Accessed: Jun. 2022. [Online]. Available: https://github.com/yaseen21khan/Classification-of-Heart-Sound-Signal-Using-Multiple-Features-

[48] (2017). *ICBHI 2017 Challenge*. [Online]. Available: https://bhichallenge.med.auth.gr/
[49] B. Rocha, "A respiratory sound database for the development of automated classification," in *Precision Medicine Powered by pHealth and Connected Health*. Singapore: Springer, 2017, pp. 33–37.
[50] S. N. Ali and S. B. Shuvo. (2021). *Hospital Ambient Noise Dataset*. [Online]. Available: https://www.kaggle.com/nafin59/hospital-ambient-noise
[51] S. E. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent hidden Markov model," *Physiological Meas.*, vol. 31, no. 4, pp. 513–529, Apr. 2010.
[52] P. Langley and A. Murray, "Heart sound classification from unsegmented phonocardiograms," *Physiological Meas.*, vol. 38, no. 8, pp. 1658–1670, Jul. 2017.
[53] S. Reichert, R. Gass, C. Brandt, and E. Andrés, "Analysis of respiratory sounds: State of the art," *Clin. Med. Circulatory, Respiratory Pulmonary Med.*, vol. 2, pp. 45–58, May 2008.
[54] A. Raza, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B.-W. On, "Heartbeat sound signal classification using deep learning," *Sensors*, vol. 19, no. 21, p. 4819, Nov. 2019.
[55] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[56] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.
[57] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," 2018, *arXiv:1806.03185*.
[58] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[59] H. Tang, M. Wang, Y. Hu, B. Guo, and T. Li, "Automated signal quality assessment for heart sound signal by novel features and evaluation in open public datasets," *BioMed Res. Int.*, vol. 2021, pp. 1–15, Feb. 2021.
[60] S. B. Shuvo, S. N. Ali, S. I. Swapnil, M. S. Al-Rakhami, and A. Gumaei, "CardioXNet: A novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings," *IEEE Access*, vol. 9, pp. 36955–36967, 2021.
[61] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2019, pp. 175–179.
[62] S. Li, H. Liu, Y. Zhou, and Z. Luo, "A Si-SDR loss function based monaural source separation," in *Proc. 15th IEEE Int. Conf. Signal Process. (ICSP)*, vol. 1, Dec. 2020, pp. 356–360.

SHAMS NAFISA ALI received the B.Sc. degree in biomedical engineering from the Department of Biomedical Engineering (BME), Bangladesh University of Engineering and Technology (BUET), in 2022, where she is currently pursuing the M.Sc. degree. She is a Lecturer with BME, BUET. As an active member of the mHealth Research Group, her research works primarily focuses on biomedical signal and image processing, computer-aided diagnosis and AI in healthcare, and wearable devices. She is also interested to pursue the aspects of biophotonics, bioinstrumentation, soft-intelligent materials, and biohybrid robots.

SAMIUL BASED SHUVO received the B.Sc. degree from the Department of Biomedical Engineering (BME), Bangladesh University of Engineering (BUET), where he is currently pursuing the M.Sc. degree. He is a Lecturer with the BME, BUET. His research focuses on biomedical signal processing and deep learning-based health informatics, and he is an active member of the mHealth laboratory. His research interests include bio-simulations, bio-instrumentation, and biomechanics systems.

MUHAMMAD ISHTIAQUE SAYEED AL-MANZO received the M.B.B.S. degree. He is currently a Registrar of cardiac surgery with the Department of Cardiac Surgery, National Heart Foundation Hospital and Research Institute, Dhaka. His research interest includes various aspects of cardiovascular surgery, including coronary artery bypass graft and cardiac abnormalities of neonates and children. He is an FCPS.

ANWARUL HASAN (Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of Alberta, Canada, in 2010. He was an Assistant Professor of biomedical and mechanical engineering with the American University of Beirut, Lebanon, and a Visiting Assistant Professor and an NSERC Postdoctoral Fellow with Harvard University and the Massachusetts Institute of Technology, USA. He is currently an Associate Professor with the Department of Mechanical and Industrial Engineering and the Biomedical Research Center, Qatar University. His current research interests include biomaterials, tissue engineering, 3D bioprinting, diabetic wound healing, cancer biochips, and machine learning and artificial intelligence in health care applications.

TAUFIQ HASAN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical and electronic engineering (EEE) from the Bangladesh University of Engineering and Technology (BUET) and the Ph.D. degree in electrical engineering from The University of Texas at Dallas. He was a member of the Center of Robust Speech Systems (CRSS), The University of Texas at Dallas. He was a Research Scientist with Robert Bosch Research and Technology Center, Palo Alto, CA, USA. He is currently with the Department of Biomedical Engineering, BUET, as an Associate Professor, where he leads the mHealth Research Group. He is also with the Center for Bioengineering Innovation and Design (CBID), Department of Biomedical Engineering, Johns Hopkins University. His research interests include biomedical signal/image analysis and medical device design.