CardioXNet: A Novel Lightweight CRNN Framework for Classifying Cardiovascular Diseases from Phonocardiogram Recordings

Samiul Based Shuvo*, Shams Nafisa Ali*, Soham Irtiza Swapnil*

Abstract—The alarmingly high mortality rate and increasing global prevalence of cardiovascular diseases (CVDs) signify the crucial need for early detection schemes. Phonocardiogram (PCG) signals have been historically applied in this domain owing to its simplicity and cost-effectiveness. However, insufficiency of expert physicians and human subjectivity affect the applicability of this technique, especially in the low-resource settings. For resolving this issue, in this paper, we introduce CardioXNet, a novel lightweight CRNN architecture for automatic detection of five classes of cardiac auscultation namely normal, aortic stenosis, mitral stenosis, mitral regurgitation and mitral valve prolapse using raw PCG signal. The process has been automated by the involvement of two learning phases namely, representation learning and sequence residual learning. The first phase mainly focuses on automated feature extraction and it has been implemented in a modular way with three parallel CNN pathways i.e., frequency feature extractor (FFE), pattern extractor (PE) and adaptive feature extractor (AFE). ID-CNN based FFE and PE respectively learn the coarse and fine-grained features from the PCG while AFE explores the salient features from variable receptive fields involving 2D-CNN based squeeze-expansion. Thus, in the representation learning phase, the network extracts efficient time-invariant features and converges with great rapidity. In the sequential residual learning phase, because of the bidirectional-LSTMs and the skip connection, the network can proficiently extract temporal features. The obtained results demonstrate that the proposed end-to-end architecture yields outstanding performance in all the evaluation metrics compared to the previous state-of-the-art methods with up to 99.6% accuracy, 99.6% precision, 99.6% recall and 99.4% F1-score on an average while being computationally comparable.

Index Terms—Phonocardiogram analysis, Unsegmented heart sound, Cardiovascular disease, Lightweight CRNN architecture, Deep learning, SqueezeNet.

I. INTRODUCTION

Cardiovascular diseases (CVDs), taking away millions of human lives every year, are inciting major concerns in the global healthcare landscape. In 2016 alone, 17.9 million people, almost 31% of entire global deaths occurred from CVDs [1]. In the same vein, lately an alarming upward trend of CVD-oriented deaths has been observed in the developing and under-developed countries [2]. Therefore, with a view to slackening the current pace of death toll and minimizing the adverse impacts of these diseases, early diagnosis and treatment of CVDs are of paramount importance.

Heart, one of the most important organs of the human body, produces distinct sounds during its course of mechanical activity. Since the occurrence of a specific type of disease alters the heart functionality in a definite manner, the auscultations also change accordingly and thus, they have been historically utilized for screening CVDs [3]. Although Lub (S1) and Dub (S2) play pivotal role for the detection of cardiac anomalies [3], sometimes the irregular variants, third heart sound (S3), fourth heart sound (S4) and murmurs are also taken into account while making pathological inference using the stethoscope [4]. The simple, non-invasive nature of this auscultation based diagnosis strategy have made it the most popular and attractive choice of the cardiologists for performing the initial investigation. [5]. Besides auscultation, multiple other cardiac signals involving a wide range of advanced methods such as, electrocardiogram (ECG), angiography, echocardiography, myocardial perfusion imaging (MPI), cardiac computed tomography (CCT), cardiovascular magnetic resonance (CMR), carotid pulse graph, apex cardiogram etc. are being utilized as modern diagnostic tools for effective screening of CVDs as they vividly reflect the overall transthoracic physiological conditions of the cardiovascular system [6], [7]. However, in situations where these advanced methods are scarce, lung auscultation stands out as a simple and reliable medium for detecting CVDs. Nevertheless, even for an expert physician, it is quite challenging to readily detect CVDs just by listening to the auscultation. This situation is further exacerbated owing to subjective dissimilar interpretation of the same auscultation by the physicians [8]. The disproportionate number of skilled medical professionals with respect to the total population adds up to the situation and further retard the initial diagnostic speed. In this scenario, artificial intelligence-empowered automated cardiac screening systems on the basis of PCG classification can play a pivotal role to assist the physicians in their decision-making process.

Since phonocardiogram (PCG), the visualization of heart sound on graphical waveform, can be easily processed to extract essential discriminatory features for the identification of cardiac anomalies [9], monitoring heart condition via PCG is becoming an increasingly popular clinical practice. With the amalgamation of machine learning (ML), state-of-the-art networks and advanced audio processing techniques, the unprofound manual screening can be replaced with automated classification frameworks for prompt large-scale predictions.

*These authors share first authorship on, and contributed equally to, this work.

Manuscript received October XX, 20XX; revised November XX, 20XX.
Fig. 1: An overview architecture of the proposed CardioXNet consisting of 2 phase learning i.e., representation learning and sequence residual learning. The microstructure of the internal modules (AFE; FE and PE) under representation learning phase are shown in figures 3, 4.

Therefore, numerous research works have explored the domain of automated PCG classification over the years [10]–[27]. In the literature, a distinct trend of performing 2-step PCG classification can be conspicuously observed. A predominantly trend of performing a segmentation step (localizing lub (S1) and dub(S2) from cardiac auscultation) prior to conducting the classification of the recordings is seen in these works. For materializing the latter step, initially the researchers in the past have utilized statistical analysis and several ML-based algorithms like, Support Vector Machines (SVM) [3], [13], k-Nearest Neighbor (k-NN) [18], Random Forest [23], Hidden Markov Models [25], Multilayer Perceptron (MLP) [3], [16] etc. in conjunction with many signal processing and feature extraction techniques based on time-domain, frequency-domain and time-frequency domain. Among the feature extraction techniques, fast Fourier transform (FFT), short Fourier transform (STFT), discrete wavelet transform (DWT), continuous wavelet transform (CWT), Q wavelet transform (TQWT), chirplet transform (CT), S-transform, Hilbert–Huang transform (HHT), TF decomposition, Empirical mode decomposition (EMD), Mel-Frequency Cepstral Coefficients (MFCCs), Variance Fractal Dimension (VFD) are observed in many works [9]. Despite being quite promising, most of these ML-based algorithm rely on hand-engineered features in the classifier and thus, are not generalized [19]. Advent of deep learning (DL) architectures namely Convolutional Neural Networks (CNNs), Unsupervised Pre-trained networks (UPNs), Recurrent and Recursive neural networks (RNNs) and their application in the the domain of PCG classification have intuitively resolved both the generalization and the accuracy concern by utilizing the inherent self-learning competency of these networks [14], [15]. However, these deep networks attain the automated feature extraction capability only after going through a computationally complex extensive training phase with a significantly large dataset [28]. The resource-intense requirement of DL-based frameworks makes them unbefitting to be deployed in the low-resource point-of-care locations of the developing and under-developed countries. Therefore, to implement in mobile platforms and wearable devices, reduction of network parameters along with optimized memory footprint is of paramount importance. A handful of cross-domain studies have investigated the issue and introduced a few advanced strategies such as lightweight networks [29], [30], weight quantization [31] and low precision computation techniques [32]. Recently, some studies have successfully implemented these concepts in various biomedical applications including ECG classification [33], respiratory disease classification [28] and most importantly, PCG classification [15].

In this paper, we propose CardioXNet, a novel lightweight CRNN framework with two learning scheme i.e., representation learning and sequence residual learning for automatic detection of CVDs from PCG signal. Representation learning includes adaptive feature extractor, frequency feature extractor and pattern extractor which combinedly extract time-invariant salient features from the raw audio of the Github PCG database [34]. Sequence residual learning contains bidirectional connections and is used for extracting temporal features. The use of these schemes enables the network to classify efficiently and converge rapidly on the raw PCG signals. The performance of the proposed end-to-end lightweight CRNN network has been evaluated considering the standard evaluation metrics and meticulously compared with the existing state-of-the-art works in terms of classifying the given 5 class of CVDs using the PCG signal.

The rest of the paper is organized as follows: In Section II, we have discussed the physiological origin of cardiac auscultation and presented an overview of the previous studies on detection of CVDs using PCG signal. Deep learning network fundamentals are discussed in Section III; a detailed description of our proposed architecture, along with the intuitions behind designing it is presented in Section IV; the implementation settings, evaluation criteria and experimental results are discussed with reference to the literature in Section V; the future prospects of the work along with the existing limitations are mentioned in Section VI and finally, we conclude our study in Section VII.

II. BACKGROUND

In the literature, numerous works have been reported on multi-faceted PCG classification. Among them, a vast majority
of the existing works have focused on performing physiologic (normal) vs. pathologic (abnormal) prediction of the PCG [3], [5], [12]–[15]. Again, some have explored the domain of multi-class (normal, artifact, extrasystole, extrahls, murmurs) cardiac anomaly prediction [6], [10], [11] while some have done multi-class (innocent, systolic, diastolic, pansystolic, ejection) classification of murmurs, the most efficient indicator of CVD incidence [26], [27]. Being motivated by the unparalleled success of the aforementioned schemes, several research works from all over the world are being geared towards direct classification of cardiovascular diseases utilizing cardiac auscultation. By elucidating PCG data, to date researchers have successfully classified many CVDs such as, coronary artery disease (CAD), atrial fibrillation (AF), aortic stenosis (AS), dilated cardiomyopathy (DC), mitral stenosis (MS), papillary muscle dysfunction (PMD), mitral regurgitation (MR), atrial septal defect (ASD), ventricular septal defect (VSD), mitral valve prolapse (MVP) etc. [16]–[24].

A. Physiological Origin of Cardiac Auscultation

Cardiac auscultation is a complex sound signal generated by the mechanical activities of the heart. It contains the impact of vibration of the valvular opening and closing, systolic and diastolic movement of muscles and tendons and the internal blood perfusion [35]. However, there are two types of fundamental heart sounds (FHSs). At the onset of the systolic phase, when the atrioventricular (AV) mitral and tricuspid valves close, the first FHS, “lub” (S1) can be heard [36]. The definitive abnormality in S1 might indicate the occurrence of MS, AF and ASD [37]. S1 is followed by the second FHS, “dub” (S2) at the beginning of the diastolic phase from the closure of the pulmonic and aortic valves. This relatively high-pitched sound demonstrates higher frequency [38] in case hypertension and AS during an event of cardiac abnormality.

In normal cardiac condition, distinct pattern of systolic and diastolic period can be observed. Nevertheless, 3rd and 4th heart sound, respectively S3 and S4, murmurs, gallops, clicks might occur because of CVD incidence [35]. While S3 mostly indicates reduced systolic function and congestive heart failure (except for pregnant ladies, athletes and young people); S4 is an clear indicator of diastolic dysfunction. The audible whooshing sounds i.e., murmurs might be generated because of the stiffening of the heart valves, narrowing of the blood vessels and turbulent flow [38]. Murmurs are very significant for classifying CVDs, as their diverse pitch and frequency pattern represent many CVDs including MR, AR, AS, MS etc. [37]. However, innocent murmurs are sometimes produced depending on the physiological condition of the individual.

B. Literature Review

As mentioned earlier, the automatic detection of CVDs and heart valve disorders (HVDs) have been explored in multiple previous works. Some of these works have utilized publicly available data while some implemented the task with in-house datasets. In this subsection, we will provide a detailed overview of some of the prominent works in this domain.

Employing divergence analysis on 50 wavelet features with MLP classifier, Dokur et. al [16] has achieved up to 99% accuracy in determining discriminatory features of VSD, MR, AR, AS, MS, diastolic rumble, summation gallop, aortic insufficiency and some other cardiac abnormalities using a self-constructed multi-source data. In [17], septal defects were detected along with other heart defects like valvular defects, ventricular hypertrophy, constrictive pericarditis etc from publicly available PCG database using TQWT based subbands decomposed from segmented heart disease. The classification was performed using least squares support vector machine (LS-SVM) algorithm and achieved overall accuracy of 99.03%. Using the Github PCG database [34], a multivariate comprehensive analysis for detecting 5 classes, i.e., MS, AS, MR, MVP, N has been carried out in [18]. Discrete Wavelets Transform (DWT) and Mel Frequency Cepstral Coefficient (MFCCs) have been employed as feature extraction technique while SVM, deep neural network (DNN) and centroid displacement based K-NN classifier have been adopted as classifier. The features and classifiers are attempted both single-handily and in an ensemble manner through rigorous experimentation. Merging the DWT and MFCC features, centroid displacement based K-NN, DNN and SVM respectively yields an accuracy of 97.4%, 92.1% and 97.9% [18]. In the same vein, using the same dataset, a recent study has proposed WaveNet, a novel 1D CNN network having residual blocks with dilated 1D CNN and multiple skip connections [19]. Upon 10-fold cross-validation, they have achieved training accuracy of 97% and approximately 90% validation accuracy. Along with sensitivity 92.5% and specificity 98.1%, the model exhibits satisfactory performance. A comprehensive basis for CAD diagnosis has been provided in [20] employing a novel multi-domain feature fusion framework. For the in-house dataset used, the concerted feeding of the selected MFCC features with the deep features into a MLP classifier has obtained a promising accuracy of 90.43%.

Several studies have utilized PCG signals for classifying heart valve disorders (HVDs). In [21], an in-house dataset of 120 PCG signals was used for HVD classification using wavelet packet(WP) decomposition and SVM. The classification task was performed on 2 classes (Normal and HVDs) achieving specificity of 96.67%. Another work [22] involving in-house dataset, has classified normal, pulmonary and mitral stenosis heart valve diseases utilized discrete fourier transform (DFT) and Burg autoregressive (AR) spectrum analysis for feature extraction. Using Principal Component Analysis (PCA) for dimensionality reduction and ANN for classification, the work has achieved maximum accuracy score of 95%. Maintaining the legacy, Ghoshet. et al [23] has classified the HVD classes i.e., MR, AR, AS, N using magnitude and phase features, wavelet synchrosqueezing transform (WSST) based time-frequency matrix with a Random Forest classifier on the Github PCG database [34] dropping the MVP class. The proposed paradigm has demonstrated the mean accuracy to be 95.13%. In a later work by the same author [24], using Chirplet transform (CT), local energy (LEN) and local entropy (LENT) features in conjunction with a composite classifier, an overall accuracy of 98.33% was obtained on the same dataset.
III. NETWORK FUNDAMENTALS

A. Convolutional Neural Network (CNN)

Convolutional neural networks, a variant of state-of-the-art artificial neural networks, mainly combine three architectural generalization ideas by ensuring different invariance techniques such as, shared weights, local receptive fields and the subsequent presence of sub-sampling of spatial or temporal channels with a view to reducing the complexity of the network model. Each of the CNN layer consists of a number of neural units and receives a set of extracted features as input from the previous layer. CNNs for images, speech, and time-series signals have different sets of kernel formation and they map the input into different types of features from each location [39].

1) 1D CNN: In spite of being analogous to the regular neural network (ANN), 1D CNN generally takes raw sequential data i.e., time-series data like PCG, EEG as input. The main advantage of 1D CNN over ANN is that it can show excellent performance on time-series data without engaging any time or resource intensive feature extraction steps.

Assuming the input of 1D CNN is a tensor representing the audio waveform denoted as X, feature extraction of 1D convolution can be defined as,

\[ T = F(X; \Theta) = f_L(\ldots f_1(f_1(X; \Theta_1)|\Theta_2)|\Theta_L) \]  

Here, \( L \) is the number of hidden layers, \( \Theta \) is a set of parameters that maps the input to the prediction \( T \).

The 1-th convolutional layer’s operation can be defined as:

\[ T_l = f_l(X_l; \Theta_l) = h(W \otimes X_l + b), \Theta_l = [W, b] \]  

Here, \( \otimes \) denotes the convolution operation, \( X_l \) is a two-dimensional input matrix of \( N \) feature maps, \( W \) is a set of \( N \) one dimensional kernels (receptive fields) used for extracting a new set of features from the input array, \( b \) is the bias vector, and \( h(\cdot) \) is the activation function. The shapes of \( X_l, W \) and \( T_l \) are respectively (\( N, d \)), (\( N, m \)) and (\( N, d - m + 1 \)) [40].

2) 2D CNN: 2D CNN is generally used for extracting spatial features using its kernel. It has multiple hidden layers, each layer having a number of neural units. The neural unit is basically a function connecting inputs and outputs. The input function can be denoted as [41]:

\[ x_k = \sum_{n=1}^{N_n-1} conv2D(w_{nk}^n, s_{i}^{n-1}) + b_k^n \]  

Here, \( s_{i}^{n-1} \) represents the output of the ith neuron of (n-1)th layer and \( w_{nk}^n \) and \( b_k^n \) are the bias and kernel of the kth neuron in the nth layer, accordingly, conv2D represents a two-dimensional regular convolution with valid padding. The input function of the kth neuron can be denoted as [41]:

\[ y_k^n = f \left( \sum_{n=1}^{N_n-1} conv2D(w_{nk}^n, s_{i}^{n-1}) + b_k^n \right) \]  

B. Recurrent Neural Network (RNN)

RNN is a special kind of ANN model where the input samples contain more inter-dependencies. RNN has a significant specialisation at storing the information of the past time steps. In other words, the parameters available at time \( t+1 \) will be affected by output produced at time \( t \) [42]. RNN is mainly used for sequential and dynamic information processing and thus, it is very useful for multiple types of audio signal processing i.e. music and speech processing, pathological auscultation processing etc. The Generalized delta rule [43] and energy minimization function are two of the major algorithms used for RNN training [44].

1) Long-Short Term Memory: The Long-Short Term Memory (LSTM), proposed by Hochreiter and Schmidhuber in 1997 is a primary building block of RNN. It has been substantially explored for sequence learning. A standard LSTM cell (see Figure 2) consists of three distinct gates and three main cell-memory networks. The respective equations for each of these gates and networks are presented below:

Input gate:

\[ i_t = Sigmoid(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \]  

Forget gate:

\[ f_t = Sigmoid(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \]  

Output gate:

\[ o_t = Sigmoid(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \]  

Input cell:

\[ g(t) = tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g) \]  

Memory cell:

\[ c(t) = f_t \odot c_{t-1} + i_t \odot g_t \]  

Output cell:

\[ h(t) = o_t \odot tanh(c_t) \]  

Here, sigmoid and tanh are the activation functions which map the non-linearity of the features [45].
2) **Bidirectional LSTM:** The main disadvantage of one-directional LSTM is that it only encompasses the past information. To overcome this limitation of the one-directional LSTM, a bi-directional LSTM (Bi-LSTM) approach has been introduced by Schuster and Paliwal [46] for efficiently investigating both the past and the future context. The Bi-LSTM is defined as,

\[
\begin{align*}
\vec{h} &= \sigma(x_t U + \vec{h}_{t-1} W + b_t) \\
\overrightarrow{h} &= \sigma(x_t U + \overrightarrow{h}_{t-1} W + b_t)
\end{align*}
\]

Here, \(\vec{h}\) is the forward hidden state and \(\overrightarrow{h}\) is the backward hidden state. Concatenation of both of these states formed the hidden state at time \(t\).

### C. Micro-architecture of Fire module

The Fire module, composed of two layers i.e., Squeeze and Expansion layer, is a dedicated parameter reduction module introduced in the SqueezeNet architecture [47]. In the Squeeze layer, \(3 \times 3\) kernels have been replaced with \(1 \times 1\) kernels which results in 9x reduction in the number of parameters. Then, the squeezed features are fed into an expansion layer. The expansion layer consists of a layer with \(1 \times 1\) filters and a layer with \(3 \times 3\) filters and the extracted features of these layers concatenated together in the channel dimension.

By maintaining 8x fewer output channels in the squeeze layer compared to the expansion layer, squeeze ratio (SR) of 0.125 has been achieved in the original Fire module [47].

### IV. Network Architecture

The proposed network namely CardioXNet is constructed in a modular way, employing two major learning paradigms i.e., representation learning and sequence residual learning. This architecture is designed for effectively classifying short unsegmented PCG recordings. Figure 1 portrays the high-level network architecture.

![Fig. 3: Micro structure of (a) Frequency Feature Extractor (FFE); (b) Pattern Extractor (PE). Fs represents sampling frequency](image)

---

**A. Representation learning**

This part is trained with a view to extracting time-invariant features from each of the raw PCG signals. It consists of three parallel CNN pathways namely, Frequency Feature Extractor (FFE), Pattern Extractor (PE) and Adaptive Feature Enhancer (AFE). The micro-structures of FFE and PE have been constructed being inspired from Deep SleepNet [42] while the concept of AFE block is influenced by the Fire module of SqueezeNet [47].

1) **Frequency Feature Extractor (FFE):** FFE consists of four 1D convolutional layers and two max-pooling layers with the primary filter size of sampling frequency (Fs) \(\times 4\) and stride size set to Fs/2 for the 1D convolutional (conv1) layer to capture the frequency components. The subsequent filter and stride sizes are chosen by rigorous hyperparameter tuning (see Figure 3).

2) **Pattern Extractor (PE):** Similar to FFE, PE also consists of four 1D convolutional layers and two max-pooling layers. However, fine-grained convolution with filter and stride size set to Fs/2, and Fs/16 respectively are used which helps to recognise the appearance of various PCG pattern. With the help of hyperparameter tuning, the filter and stride sizes are chosen (see Figure 3).

3) **Adaptive Feature Enhancer (AFE):** The input sequences are reshaped into a 2D tensor and fed to AFE. AFE consists of 2D convolutional layers, batch-normalization, max-pooling layers and squeeze-expansion layers, inspired by the Fire module of SqueezeNet architecture [47]. Batch normalization layers helps to stabilize and speed up the training process [48].
while abstract feature maps generated from the initial convolutional layers are then passed to the squeeze-expansion layers which results in lowering of parameter count by manifolds (see Figure 4). Apart from efficient feature extraction, it also ensures faster convergence in the training phase, a highly appreciated task from the perspective of training deep networks in mobile platforms.

Outputs from three CNN paths are concatenated together and forwarded to the sequence residual learning part.

$$X_{\text{concat}} = X_{\text{AFE}} \oplus X_{\text{FFE}} \oplus X_{\text{PE}}$$  \hspace{1cm} (13)

Here, $X_{\text{concat}}$ is the concatenated feature vector and $X_{\text{AFE}}, X_{\text{FFE}}, X_{\text{PE}}$ are the extracted feature vectors from three CNN paths.

B. Sequence residual learning

Sequence residual learning is trained to extract the temporal information from the sequence of extracted features in the representation learning part. Two layers of bi-LSTMs have been employed to learn temporal information which enables the encoding of both past and future information by processing with two independent LSTMs. Since PCG is a time-series data, the application of Bi-:STM is best suited for capturing the sequential information embedded in them.

A skip connection has been employed to implement residual functionality and enabling the addition of temporal information and previously extracted features from the CNNs. The concatenated feature vector is fed into a prediction layer with probability nodes, calculated by the softmax function.

In all the convolutional layers and LSTM layers, Rectified Linear Unit (ReLU) [49] activation function has been used, which is defined as:

$$f(x) = \max(0, x),$$  \hspace{1cm} (14)

It has been employed to diminish the vanishing gradient issue and also contributes to the faster convergence. For minimizing the over-fitting phenomenon, dropout layers have been employed.

V. EXPERIMENTS

A. Experimental Dataset

In this study, a publicly available dataset of PCG recordings collected from the following Github repository [34] has been utilized for training and validating the PCG classification scheme. The recordings were collected from various sources like books and websites and contained a total number of 1000 PCG recordings in .wav format in five different classes i.e., Normal (N), Aortic stenosis (AS), Mitral regurgitation (MR), Mitral stenosis (MS), Mitral valve prolapse (MVP). Each of the classes has 200 recordings. All the recordings are sampled at 8 kHz. Figure. 5 illustrates the waveform for each of the disease classes.

B. Evaluation Metrics

The dataset was divided into training and validation parts with 80% of the PCG signal used for training the model and 20% of the signal used for validation. Well-known and important performance evaluation metrics such as, Accuracy, Precision, Recall and F1- score are chosen to evaluate the proposed framework quantitatively.

$$\text{Acc} = \frac{TP + TN}{TP + FP + TN + FN}$$  \hspace{1cm} (15)

$$\text{Precision} = \frac{TP}{TP + FP}$$  \hspace{1cm} (16)

$$\text{Recall} = \frac{TP}{TP + FN}$$  \hspace{1cm} (17)

$$F1 \text{ score} = \frac{2 * TP}{2 * TP + FP + FN}$$  \hspace{1cm} (18)

C. Experimental Setup

The CardioXNet is designed using Keras and TensorFlow backend and training work has been performed using NVidia K80 GPUs provided by Kaggle notebooks.

Sparse categorical crossentropy loss and adaptive learning rate optimizer (Adam) with the learning rate of 0.00001 have been engaged in the training process and a batch size of 16 is selected for training and validation purpose.

TABLE I: Classification Report for different classes using CardioXNet

| Class | Accuracy | Precision | Recall | F-1 score |
|-------|----------|-----------|--------|-----------|
| AS    | 1.00     | 0.98      | 1.00   | 0.99      |
| MR    | 1.00     | 1.00      | 1.00   | 1.00      |
| MS    | 1.00     | 1.00      | 0.98   | 0.99      |
| MVP   | 0.98     | 1.00      | 1.00   | 1.00      |
| N     | 1.00     | 1.00      | 1.00   | 1.00      |
| Weighted Avg. | 0.996 | 0.996 | 0.996 | 0.994 |

D. Performance of the Proposed Framework

In this work, the PCG dataset has been classified into 5 different classes using CardioXNet. The proposed model achieved near perfect validation accuracy on the given dataset showing very high precision and recall scores for all the classes (see Table. I). Furthermore, Figure. 6 (a) depicts that the model converged with great rapidity reaching 100% training accuracy, while validation accuracy reached 99.6% within 100 epochs. Additionally, Figure. 6 (b) provides an insight into the minimization of the categorical cross entropy loss. The incorporation of adaptive feature extractor block with the squeeze and expansion layers mainly attributed to this faster convergence of the model. The confusion matrix shown in Figure 7, bears proof of the superior performance of the CardioXNet in classification of all the classes.
Fig. 5: Waveform of the existing CVD classes on PCG dataset. (a) Aortic stenosis (AS), (b) Mitral regurgitation (MR), (c) Mitral stenosis (MS), (d) Mitral valve prolapse (MVP) and (e) Normal.

Fig. 6: (a) Epoch vs accuracy of the CRNN during training and validation (b) Epoch vs loss of the CRNN during training and validation.
E. Comparison with the Existing Works

A detailed comparative analysis among the proposed CardioXNet and the previous works have been presented in Table II. In [50], manually crafted MFCC, DWT and MFCC-DWT features were extracted for feeding into the classifiers i.e., KNN, SVM and DNN. Although multiple analysis were performed adopting several feature extraction and classifier combinations, MFCC-DWT based conjoined features demonstrated the best performance, giving 97.4%, 97.9% and 92.1% accuracy respectively, on KNN, SVM and DNN classifiers. Another work proposing a novel deep WaveNet architecture has achieved a training accuracy of 97% and validation accuracy of 90% for 5 class CVD classification [19]. Furthermore, two recent studies on the classification of HVDs have utilized this dataset dropping the MVP class. The first work has achieved the mean accuracy of 95.13% using wavelet synchrosqueezing transform function (WSST) for extracting the magnitude and phase features from the PCG and classified the signals with the Random Forest (RF) classifier [23] while the second one contains several complex processing steps. After performing filtering and Shannon energy based cardiac cycle segmentation, local energy and local entropy features were extracted in conjunction with Chirplet transform (CT). Upon the evaluation of the features using multi-class composite classifier (MCC), the framework has obtained overall accuracy of 98.33% [24]. Therefore, is is clearly evident that the proposed CardioXNet architecture outperforms the previous works by a relatively large margin, achieving an accuracy of 99.5%. Individual class-wise evaluation metrics also show similar trend of superior performance in comparison with all the existing works.

The strategies adopted in [23], [24], [50] require several computationally complex and time-consuming hand-engineered feature extraction techniques on the PCG dataset, while the CardioXNet is an end-to-end model which is capable of directly predicting CVDs from the raw PCG signal. Thus, the proposed scheme is well-performing and highly efficient with reduced complexity and latency in terms of making automated predictions. Although the WaveNet model can optimally classify the raw PCG recordings with relatively lower training parameters, it fails to perform as well as the proposed CardioXNet despite having advanced network and deep feature extraction components. Rather, it lagged by around 9.6% accuracy in the validation phase in comparison to our proposed model. This clearly marks the lacking of robustness of the WaveNet model. In addition, the model is ambiguously represented with no clear depiction of kernels and layers; thereby, making the network irreproducible. From this exhaustive analysis, it can be inferred that the proposed CardioXNet manifests significant improvement over all the existing state-of-the-art works using the GitHub PCG database.

F. Computational Efficiency of the proposed model

In this part, a detailed benchmarking is performed on the CardioXNet: considering trainable parameters, required end-to-end time and size after training the parameters. The proposed lightweight CNN model has extremely low end to end time of 63.687(±0.06) ms. Since the proposed paradigm can directly classify the raw PCG recordings requiring no preprocessing and hand crafted feature extraction technique which evidently demonstrates the robustness of the model. Moreover, the proposed model has 1.9 M trainable parameters and a smaller memory requirement of only 23 MB. Therefore, it can achieve higher accuracy than the considered networks while necessitating significantly low storage space and processing power. This makes the CardioXNet more suitable for real-time wearable and mobile applications [51].

VI. LIMITATION AND SCOPEs FOR FUTURE WORKS

The current work is performed only on a specific dataset which is collected from various sources in ideal clinical settings. However, in real world scenarios, the PCG data contain noise, various environmental sounds and might be susceptible to variation in acquisition device. Additionally, different disease classes in the given dataset is balanced, but in reality, according to the disease prevalence and patient

| Feature Extraction | Model       | Accuracy | F1-score |
|--------------------|-------------|----------|----------|
| MFCC-DWT [50]      | KNN         | 97.4     | 99.2     |
|                    | DNN         | 97.9     | 99.7     |
|                    | SVM         | 92.1     | 98.3     |
|                    | WaveNet [19]| 90.0     | -        |
| WSST [23]          | Random Forest| 95.13   | -        |
| Chirplet Transform [24] | CardioXNet (Proposed) | 99.6 | 99.4 |
availability, the training data might be susceptible to sparse unbalanced data for each of the disease class. These factors might require additional preprocessing and augmentation procedures. The work can be further improved if multiple larger PCG datasets are available with diverse CVD annotations. Several advanced network strategies like Generative Adversarial Networks (GAN), Variational Autoencoder (VAE) can be attempted to obtain further optimized performances with more efficiency and non-latency.

In future, we aim to integrate our CardioXNet in digital stethoscopes or wearable devices with a cloud server connection for performing automatic PCG classification and predicting on different CVDs using our pre-trained model in real time to assist the clinicians in their diagnostic decision.

VII. CONCLUSION

In this work, a novel lightweight CRNN network, CardioXNet, has been proposed for automatic detection of different cardiovascular disease classes without performing any preprocessing steps on the PCG signal. The model involves both representation learning and sequence residual learning by utilizing both the CNN and bi-LSTM layers to extract invariant and temporal features. The framework demonstrated state-of-the-art performance with an accuracy score of 99.6% outperforming all the previous works on the given PCG dataset by a considerably large margin. Besides showing superior results, the use of the proposed end-to-end network with significantly low number of parameters make this model well-qualified for usage in any embedded system applications. We believe that this work can contribute in the advancement of real-time, automated CVD classification from PCG recordings and can have significant real life impacts in the domain of clinical diagnostics.

REFERENCES

[1] “CVD Causes One-Third of Deaths Worldwide: Study Examines Global Burden of CVD From 1990 to 2015 - American College of Cardiology,” [Online]. Available: https://www.acc.org/latest-in-cardiology/articles/2017/05/17/1602/cvd-causes-one-third-of-deaths-worldwide.

[2] A. D. Bowry, J. Lewey, S. B. Dagani, and N. K. Choudhry, “The burden of cardiovascular disease in low- and middle-income countries: epidemiology and management,” Canadian Journal of Cardiology, vol. 31, no. 9, pp. 1151–1159, 2015.

[3] A. Bourouhou, A. Jilbab, A. Hammouch, “Heart sounds classification for a medical diagnostic assistance,” International Journal of Online and Biomedical Engineering (iJOE), vol. 15, no. 11, pp. 88–103, 2019.

[4] R. V. Vargheese and K. Ramachandran, “A novel heart sound activity detection framework for automated heart sound analysis,” Biomedical Signal Processing and Control, vol. 13, pp. 174–188, 2014.

[5] H. Alaskar, N. Alzhrani, A. Hussain, and F. Almarshed, “The implementation of pretrained alexnet on ppg classification,” in International Conference on Intelligent Computing. Springer, 2019, pp. 784–794.

[6] F. Chakir, A. Jilbab, C. Nacir, and A. Hammouch, “Phonocardiogram signals processing approach for pascal classifying heart sounds challenge,” Signal, Image and Video Processing, vol. 12, no. 6, pp. 1149–1155, 2018.

[7] M. U. Akram, A. Shaukat, A. Ramazan, et al., “Quality assessment and classification of heart sounds using ppg signals,” in Applications of Intelligent Technologies in Healthcare. Springer, 2019, pp. 1–11.

[8] U. Alam, O. Asghar, S. Q. Khan, S. Hayat, and R. A. Malik, “Cardiac auscultation: an essential clinical skill in decline,” British Journal of Cardiology, vol. 17, no. 1, p. 8, 2010.

[9] S. Ismail, I. Siddiqui, and U. Akram, “Localization and classification of heart beats in phonocardiography signals—a comprehensive review,” EURASIP Journal on Advances in Signal Processing, vol. 2018, no. 1, p. 26, 2018.

[10] S.-W. Deng and J.-Q. Han, “Towards heart sound classification without segmentation via autocorrelation feature and diffusion maps,” Future Generation Computer Systems, vol. 60, pp. 13–21, 2016.

[11] O. Deperlioglu, U. Kose, D. Gupta, A. Khanna, and A. K. Sangaiah, “Diagnosis of heart diseases by a secure internet of health things system based on autoencoder deep neural network,” Computer Communications, vol. 162, pp. 31–50, 2020.

[12] A. I. Humayun, S. Ghaifazadegan, M. I. Ansari, Z. Feng, and T. Hasan, “Towards domain invariant heart sound abnormality detection using learnable filterbanks,” IEEE Journal of Biomedical and Health Informatics, 2020.

[13] W. Zhang, J. Han, and S. Deng, “Heart sound classification based on scaled spectrogram and partial least squares regression,” Biomedical Signal Processing and Control, vol. 32, pp. 20–28, 2017.

[14] S. Latif, M. Usman, R. Rana, and J. Qadir, “Phonocardiographic sensing using deep learning for abnormal heartbeat detection,” IEEE Sensors Journal, vol. 18, no. 22, pp. 9393–9400, 2018.

[15] B. Xiao, Y. Xu, X. Bi, J. Zhang, and X. Ma, “Heart sounds classification using a novel 1-d convolutional neural network with extremely low parameter consumption,” Neurocomputing, vol. 392, pp. 153–159, 2020.

[16] Z. Doiği and T. Ogmi, “Feature determination for heart sounds based on divergence analysis,” Digital Signal Processing, vol. 19, no. 3, pp. 521–531, 2009.

[17] S. Patidar, R. B. Pachori, and N. Garg, “Automatic diagnosis of septal defects based on tunable-q wavelet transform of cardiac sound signals,” Expert Systems with Applications, vol. 42, no. 7, pp. 3315–3326, 2015.

[18] A.-Y. Son, S. Kwon, et al., “Classification of heart sound signal using multiple features,” Applied Sciences, vol. 8, no. 12, p. 2344, 2018.

[19] S. L. Oh, V. Jahnunnen, C. P. Ooi, R.–S. Tan, E. J. Ciaccio, T. Yamakawa, M. Tanabe, M. Kobayashi, and U. R. Acharya, “Classification of heart sound signals using a novel deep wavenet model,” Computer Methods and Programs in Biomedicine, p. 105604, 2020.

[20] H. Li, X. Wang, C. Liu, Q. Zeng, Y. Zheng, X. Chu, L. Yao, J. Wang, Y. Liao, and C. Karmakar, “A fusion framework based on multi-domain features and deep learning features of phonocardiogram for coronary artery disease detection,” Computers in Biology and Medicine, p. 103733, 2020.

[21] S. Choi, “Detection of valvular heart disorders using wavelet packet decomposition and support vector machine,” Expert Systems with Applications, vol. 35, no. 4, pp. 1679–1687, 2008.

[22] H. Uğuz, “A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases,” Journal of medical systems, vol. 36, no. 1, pp. 61–72, 2012.

[23] S. K. Ghosh, R. K. Tripathy, R. Ponnalagu, and R. B. Pachori, “Automated detection of heart valve disorders from the ppg signal using time-frequency magnitude and phase features,” IEEE Sensors Letters, vol. 3, no. 12, pp. 1–4, 2019.

[24] S. K. Ghosh, R. Ponnalagu, R. Tripathy, and U. R. Acharya, “Automated detection of heart valve diseases using chirplet transform and multiclass composite classifier with ppg signals,” vol. 118, p. 103632, 2020.

[25] H. Fahad, M. U. Ghani Khan, T. Saba, A. Rehman, and S. Iqbal, “Microscopic abnormality classification of cardiac murmurs using anfis and hmm,” Microscopy Research and Technique, vol. 81, no. 5, pp. 449–457, 2018.

[26] P. Wang, C. S. Lim, S. Chauhan, J. Y. A. Foo, and V. Anantharaman, “Phonocardiographic signal analysis method using a modified hidden markov model,” Annals of Biomedical Engineering, vol. 35, no. 3, pp. 367–374, 2007.

[27] V. N. Vargheese and K. Ramachandran, “Effectiveness of heart sound segmentation and murmur classification using empirical wavelet transform and instantaneous phase for electronic stethoscope,” IEEE Sensors Journal, vol. 17, no. 12, pp. 3861–3872, 2017.

[28] S. B. Shuvo, S. N. Ali, S. I. Swapnil, T. Hasan, and M. I. H. Bhuian, “A feature extraction cnn method for detecting respiratory diseases from lung auscultation sounds using cnn-based hybrid scalogram,” arXiv preprint arXiv:2009.04402, 2020.

[29] M. Hussain, M. A. Haque, et al., “Swishnet: a fast convolutional network for speech, music and noise classification and segmentation,” arXiv preprint arXiv:1812.00149, 2018.

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.
[31] S. Seo and J. Kim, “Efficient weights quantization of convolutional neural networks using kernel density estimation based non-uniform quantizer,” *Applied Sciences*, vol. 9, no. 12, p. 2559, 2019.

[32] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Quantized neural networks: Training neural networks with low precision weights and activations,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6869–6898, 2017.

[33] Q. Wu, Y. Sun, H. Yan, and X. Wu, “Eeg signal classification with binarized convolutional neural network,” p. 103800, 2020.

[34] “Classification of Heart Sound Signal Using Multiple Features,” [Online]. Available: https://github.com/yaseen21khan/Classification-of-Heart-Sound-Signal-Using-Multiple-Features.

[35] 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, 2018.

[36] D. Mason, *Listening to the heart: a comprehensive collection of heart sounds and murmurs*. FA Davis Company, 2000.

[37] “Abnormal heart sounds and murmurs - 62 rationale | medical council of canada,” [Online]. Available: https://mcc.ca/objectives/expert/key/62/.

[38] S. Barma, B.-W. Chen, K. L. Man, and J.-F. Wang, “Quantitative measurement of split of the second heart sound (s2),” *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 12, no. 4, pp. 851–860, 2014.

[39] Y. LeCun, Y. Bengio, et al., “Convolutional networks for images, speech, and time series,” *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[40] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[41] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size,” arXiv preprint arXiv:1602.07360, 2016.

[42] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network,” arXiv preprint arXiv:1502.01381, 2015.

[43] G.-Y. Son, S. Kwon, et al., “Classification of heart sound signal using multiple features,” *Applied Sciences*, vol. 8, no. 12, p. 2344, 2018.

[44] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, 2015.

[45] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network,” arXiv preprint arXiv:1505.00853, 2015.

[46] G.-Y. Son, S. Kwon, et al., “Classification of heart sound signal using multiple features,” Applied Sciences, vol. 8, no. 12, p. 2344, 2018.

[47] S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B.-Y. Choi, and T. R. Faughnan, “Real-time human detection as an edge service enabled by a lightweight cnn,” in *2018 IEEE International Conference on Edge Computing (EDGE)*. IEEE, 2018, pp. 125–129.