Interspeech 2018
2-6 September 2018, Hyderabad, India

An Ensemble of Transfer, Semi-supervised and Supervised Learning Methods for Pathological Heart Sound Classification

Ahmed Imtiaz Humayun[1], *Md. Tauhiduzzaman Khan*[1], *Shabnam Ghaffarzadegan*[2], *Zhe Feng*[2] *and Taufiq Hasan*[1]

[1]mHealth Lab, Dept. of Biomedical Engineering, Bangladesh University of Engineering and Technology (BUET), Bangladesh.
[2]Human Machine Interaction Group-2, Robert Bosch Research and Technology Center (RTC), Sunnyvale, CA.

taufiq@bme.buet.ac.bd, shabnam.ghaffarzadegan@us.bosch.com

Abstract

In this work, we propose an ensemble of classifiers to distinguish between various degrees of abnormalities of the heart using Phonocardiogram (PCG) signals acquired using digital stethoscopes in a clinical setting, for the INTERSPEECH 2018 Computational Paralinguistics (ComParE) Heart Beats Sub-Challenge. Our primary classification framework constitutes a convolutional neural network with 1D-CNN time-convolution (tConv) layers, which uses features transferred from a model trained on the 2016 Physionet Heart Sound Database. We also employ a Representation Learning (RL) approach to generate features in an unsupervised manner using Deep Recurrent Autoencoders and use Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) classifiers. Finally, we utilize an SVM classifier on a high-dimensional segment-level feature extracted using various functionals on short-term acoustic features, i.e., Low-Level Descriptors (LLD). An ensemble of the three different approaches provides a relative improvement of 11.13% compared to our best single sub-system in terms of the Unweighted Average Recall (UAR) performance metric on the evaluation dataset.

Index Terms: Representation learning, Heart Sound Classification, Time-convolutional Layers.

1. Introduction

Cardiac auscultation is the most practiced non-invasive and cost-effective procedure for the early diagnosis of various heart diseases. Effective cardiac auscultation requires trained physicians, a resource which is limited especially in low-income countries of the world [1]. This lack of skilled doctors opens up opportunities for the development of machine learning based assistive technologies for point-of-care diagnosis of heart diseases. With the advent of smartphones and their increased computational capabilities, machine learning based automated heart sound classification systems implemented with a smart-phone attachable digital stethoscope in the point-of-care locations can be of significant impact for early diagnosis of cardiac diseases.

Automated classification of the PCG, i.e., the heart sound, have been extensively studied and researched in the past few decades. Previous research on automatic classification of heart sounds can be broadly classified into two areas: (i) PCG segmentation, i.e., detection of the first and second heart sounds (S1 and S2), and (ii) detection of recordings as pathologic or physiologic. For the latter application, researchers in the past have utilized Artificial Neural Networks (ANN) [2], Support Vector Machines (SVM) [3] and Hidden Markov Models (HMM) [4]. In, the 2016 Physionet/CinC Challenge was organized and an archive of 4430 PCG recordings were released for binary classification of normal and abnormal heart sounds. This particular challenge encouraged new methods being utilized for this task. Notable features used for this dataset included, time, frequency and statistical features [5], Mel-frequency Cepstral Coefficients (MFCC) [6], and Continuous Wavelet Transform (CWT). Most of the systems adopted the segmentation algorithm developed by Springer et al. [7]. Among the top scoring systems, Maknickas et al. [8] extracted Mel-frequency Spectral Coefficients (MFSC) from unsegmented signals and used a 2D CNN. Plesinger et al. [9] proposed a novel segmentation method, a histogram based feature selection method and parameterized sigmoid functions per feature, to discriminate between classes. Various machine learning algorithms including SVM [10], k-Nearest Neighbor (k-NN) [6], Multilayer Perceptron (MLP) [11, 12], Random Forest [5], 1D [13] and 2D CNNs [8], and Recurrent Neural Network (RNN) [14] were employed in the challenge. A good number of submissions used an ensemble of classifiers with a voting algorithm [5, 11, 12, 13]. The best performing system was presented by Potes et al. [13] that combined a 1D-CNN model with an Adaboost-Abstain classifier using a threshold based voting algorithm.

In audio signal processing, filter-banks are commonly employed as a standard pre-processing step during feature extraction. This was done in [13] before the 1D-CNN model. We propose a CNN based Finite Impulse Response (FIR) filterbank front-end, that automatically learns frequency characteristics of the FIR filterbank utilizing time-convolution (tConv) layers. The INTERSPEECH ComParE Heart Sound Shenzhen (HSS) Dataset is a relatively smaller corpus, with three class labels according to the degree of the disease; while the Physionet Heart Sounds Dataset has binary annotations. We train our model on the Physionet Challenge Dataset and transfer the learned weights for the three class classification task. We also avail unsupervised/semi-supervised learning to find latent representations of PCG.

2. Data Preparation

2.1. Datasets

2.1.1. The INTERSPEECH 2018 ComParE HSS Dataset

The INTERSPEECH 2018 ComParE Challenge [15] released the Heart Sounds Shenzhen PCG signal corpus containing 845 recordings from 170 different subjects. The recordings were collected from patients with coronary heart disease, arrhythmia, valvular heart disease, congenital heart disease, etc. The PCG recordings are sampled at 4 KHz and annotated with three class labels: (i) *Normal*, (ii) *Mild*, and (iii) *Moderate/Severe* (heart disease).

2.1.2. PhysioNet/CinC Challenge Dataset

The 2016 PhysioNet/CinC Challenge dataset [16] contains PCG recordings from seven different research groups. The train-

10.21437/Interspeech.2018-2413

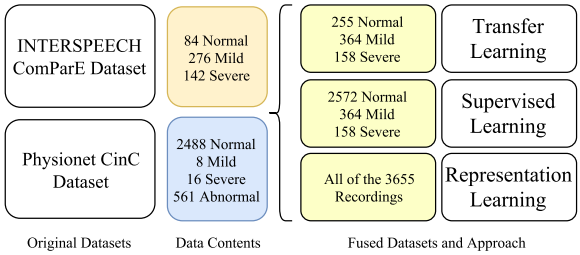

Figure 1: *Dataset preparation for transfer Learning, supervised Learning and representation learning using Physionet and ComParE corpus.*

ing data contains 3153 heart sound recordings collected from 764 patients with a total number of $84,425$ annotated cardiac cycles ranging from 35 to 159 bpm. Cardiac Anomalies range from coronary heart disease, arrhythmia, valvular stenosis/regurgitation, etc. The dataset has 2488 and 665 PCG signals annotated as *Normal* and *Abnormal*, respectively. The Aristotle University of Thessaloniki heart sounds database (AUTHHSDB) [17], a subset of the Physionet corpus (training-c), contains additional metadata based on the severity of the heart diseases. The recordings are sampled at 2000 Hz.

2.2. Data Imbalance Problem

The INTERSPEECH ComParE HSS Dataset suffers from significant class imbalance in its training set, which could introduce performance reduction for both classical machine learning and deep learning based classifiers. The training set is divided in a ratio of 16.7/55.0/28.3 percent between the *Normal/Mild/Severe* classes, with more than half of the training data comprising of PCG signals annotated as *Mild*". The result of the imbalance was evident in our recall metrics which are discussed later in Sec. 7.

2.3. Fused Training Sets

To cope with the class imbalance and increase the volume of the training data, we created 3 new fused training corpora out of the INTERSPEECH ComParE HSS Dataset and the Physionet/CinC Challenge Dataset training partitions. The AUTHHSDB (training-c) partition of the dataset was relabeled using the metadata files provided to have 7 *Normal*, 8 *Mild* and 16 *Severe* annotated recordings. The dataset distributions are depicted in Fig. 1. The fused datasets prepared for Transfer Learning (TL), Supervised Learning (SL) and Representation Learning (RL) will be referred to as TL-Data, SL-Data and RL-Data respectively.

3. Proposed Transfer Learning Framework

3.1. 1D-CNN Model for Abnormal Heart Sound Detection

The Physionet/CinC Challenge PCG database is a larger corpus with Normal and Abnormal labels designed for a binary classification task. We propose a 1D-CNN Neural Network improving the top scoring model [13] of the Physionet/CinC 2016 challenge. First, the signal is re-sampled to 1000 Hz (after an anti-aliasing filter) and decomposed into four frequency bands ($25-45$, $45-80$, $80-200$, $200-500$ Hz). Next, spikes in the recordings are removed [18] and PCG segmentation is performed to extract cardiac cycles [7]. Taking into account the longest cardiac cycle in the corpus, each cardiac cycle is zero padded to be 2.5s in length. Four different frequency bands of extracted from each cardiac cycle are fed into four different input branches of the 1D-CNN architecture. Each branch has two convolutional layers of kernel size 5, followed by a Rectified Linear Unit (ReLU) activation and a max-pooling of 2. The first convolutional layer has 8 filters while the second has 4. The outputs of the four branches are fed to an MLP network after being flattened and concatenated. The MLP network has a hidden layer of 20 neurons with ReLU activation and two output neurons with softmax activation. The resulting model provides predictions on every heart sound segment (cardiac cycle), which are averaged over the entire recording and rounded for inference.

3.2. Filter-bank Learning using Time-Convolutional (tConv) Layers

For a causal discrete-time FIR filter of order N with filter coefficients $b_0, b_1, \ldots b_N$, the output signal samples $y[n]$ is obtained by a weighted sum of the most recent samples of the input signal $x[n]$. This can be expressed as:

$$\begin{aligned} y[n] &= b_0 x[n] + b_1 x[n-1] + ... + b_N x[n-N] \\ &= \sum_{i=0}^{N} b_i x[n-i]. \end{aligned} \tag{1}$$

A 1D-CNN performs cross-correlation between its input and its kernel using a spatially contiguous receptive field of kernel neurons. The output of a convolutional layer, with a kernel of odd length $N+1$, can be expressed as:

$$\begin{aligned} y[n] = b_0 x[n + \tfrac{N}{2}] + b_1 x[n + \tfrac{N}{2} - 1] + ... + b_{\frac{N}{2}} x[n] + ... \\ + b_{N-1} x[n - \tfrac{N}{2} + 1] + b_N x[n - \tfrac{N}{2}] \\ = \sum_{i=0}^{N} b_i \, x[n + \tfrac{N}{2} - i] \end{aligned} \tag{2}$$

where $b_0, b_1, ...b_N$ are the kernel weights. Considering a causal system the output of the convolutional layer becomes:

$$y[n - \tfrac{N}{2}] = \sigma \left(\beta + \sum_{i=0}^{N} b_i x[n-i] \right) \tag{3}$$

where $\sigma(\cdot)$ is the activation function and β is the bias term. Therefore, a 1D convolutional layer with linear activation and zero bias, acts as an FIR filter with an added delay of $N/2$ [19]. We denote such layers as time-convolutional (tConv) layers [20]. Naturally, the kernels of these layers (similar to filter-bank coefficients) can be updated with Stochastic Gradient Descent (SGD). These layers therefore replace the static filters that decompose the pre-processed signal into four bands (Sec. 3.1). We use a special variant of the tConv layer that learns coefficients with a linear phase (LP) response.

3.3. Transfer Learning from Physionet Model

Our proposed tConv Neural Network is trained on the Physionet CinC Challenge Dataset with four-fold in house cross validation [21]. The model achieves a mean cross-validation accuracy of 87.10% and Recall of 90.91%. The weights up-to the flatten layer are transferred [22] to a new convolutional neural network architecture with a fully connected layer with two hidden layers of 239 and 20 neurons and 3 output neurons for *Normal*,

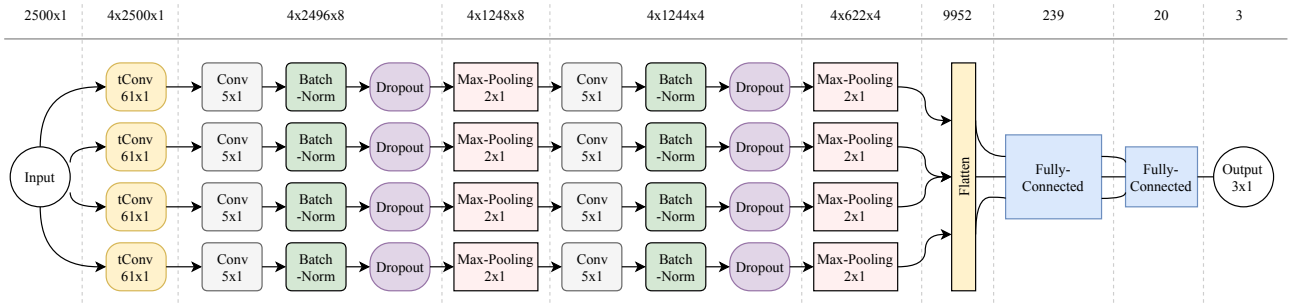

Figure 2: *Proposed architecture incorporating tConv layers for Transfer Learning.*

Mild and *Severe* classes (Fig. 2). The model weights are fine-tuned on TL-Data. TL-Data comprises of all of the samples from the INTERSPEECH ComParE Dataset and the *Normal* signals from the Physionet in house validation fold, from which the trained weights are transferred. We chose the weights of a model trained on Fold 1 for better per cardiac cycle validation accuracy. The cross-entropy loss is optimized with a stochastic gradient descent optimizer with a learning rate of $4.5 * 10^{-05}$. Dropout of 0.5 is applied to all of the layers except for the output layer. The model hyperparameters were *not* optimized while fine-tuning with TL-Data. The cost function was weighted to account for the class imbalance.

4. Representation Learning (RL) with Recurrent Autoencoders

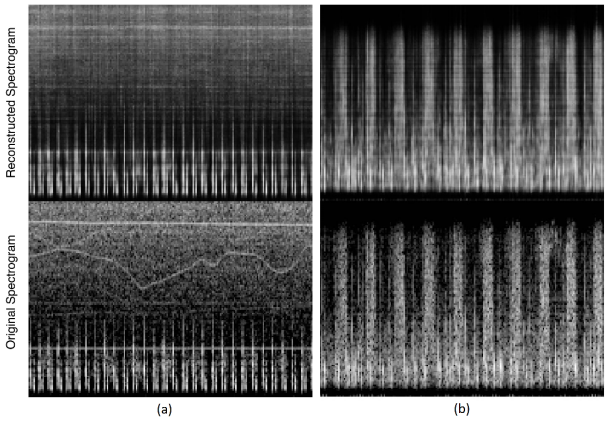

Figure 3: *Reconstructed Mel-spectrogram of recording thresholded to reduce background noise a) below -30 dB b) below -45 dB*

Representation learning is particularly of interest when a large amount of unlabeled data is available compared to a smaller labeled dataset. Considering the two corpora at hand, we approach the problem from a semi-supervised representation learning perspective to train recurrent sequence to sequence autoencoders [23] on unlabeled RL-Data (Sec. 2.3) and then use lower dimensional representations of SL-Data to train classifiers. Sequence-to-sequence learning is about translating sequences from one domain to another. Unsupervised Sequence-to-sequence representation learning was popularized in the use of machine translation [24]. It has also been employed for audio classification with success [25]. It offers the chance of resolving the overfitting problem experienced when training an end to end deep learning model.

First, mel-spectrogram of 126 bands are extracted with a window size of 320ms with 50% overlap. The raw audio files are clipped to 30 seconds in length. To reduce background noise, the spectrogram is thresholded below -30,-45,-60 and -75 dB. This results in four different spectrograms. The model is trained on all four of these separately, which results in four different feature sets. Both the encoder and decoder Recurrent Neural Network had 2 hidden layers with 256 Gated Recurrent Units each. The final hidden states of all the GRUs are concatenated into a 1024 dimensional feature vector. Fig. 3 portrays the reconstructed outputs for mel-spectrograms clipped below -30 dB and -45 dB. Four different feature vectors for the four different spectrograms are also concatenated to form fused features. Feature representations of SL-Data were used to train classifiers. The model is deployed and trained using the AUDEEP toolkit [26].

5. Supervised Learning with Segment-level Features

5.1. ComParE Acoustic Feature Set

In this sub-system, we utilize the acoustic feature set described in [27]. This feature set contains 6373 static features resulting from the computation of various functionals over LLD parameters [15]. The LLD parameters and functionals utilized are described in [27]. The features are extracted using the openSMILE toolkit [28].

5.2. Classifiers

We have implemented several machine learning algorithms for heart sound classification from the ComParE Acoustic feature set. The evaluated classifiers include: Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and Multi-Layer Perceptron (MLP). SVM classifier with complexity C= 10^{-4} and tolerance L= 0.3 outperformed the other classifiers.

6. Experimental Evaluation and Results

The evaluation metric for the INTERSPEECH ComParE Challenge is Unweighted Average Recall (UAR) since the datasets are class unbalanced. We also monitor classwise recall and accuracy for evaluation of model performance. Performance metrics on both the development and test set are listed on Table 1 with the training datasets mentioned. The Comp-SVM model, evaluated on the ComParE test set, acquired 45.9% UAR and 51.5% overall accuracy. Our transfer learning based model

Table 1: *Performance evaluation of proposed methods compared to the official baseline systems.*

Baseline Systems						
Model Name	Dataset	Features	Classifiers	UAR (%) dev	Acc. (%) dev	UAR (%) test
OPENSMILE[15]	INTERSPEECH ComParE HSS	ComParE Feature set	SVM	50.3	52.2	46.4
AUDEEP[15]	INTERSPEECH ComParE HSS	Fused Autoencoder Features	SVM	38.6	-	47.9
END2YOU[15]	INTERSPEECH ComParE HSS	CNN	LSTM	41.2	-	37.7
Fusion of best 2 systems [15]				-	-	56.2
Proposed Systems						
Model Name	Dataset	Features	Classifiers	UAR (%) dev	Acc. (%) dev	UAR (%) test
ComP-SVM	INTERSPEECH ComParE HSS	ComParE Feature set	SVM	52.1	53.9	45.9
RL-SVM	RL-Data SL-Data	-60 dB Autoencoder Features	SVM	42.9	48.9	-
RL-LDA	RL-Data SL-Data	-60 & -75dB Autoencoder Features	LDA	51.4	54.4	34.4
LP-tConv	TL-Data	tConv CNN	MLP	44.6	56.1	39.5
System Ensembles						
Ensemble System Name				UAR (%) dev	Acc. (%) dev	UAR (%) test
Fusion of Comp-SVM, RL-SVM and LP-tConv models				57.92	63.9	39.3
Hierarchical with Fusion				57.93	64.2	42.1

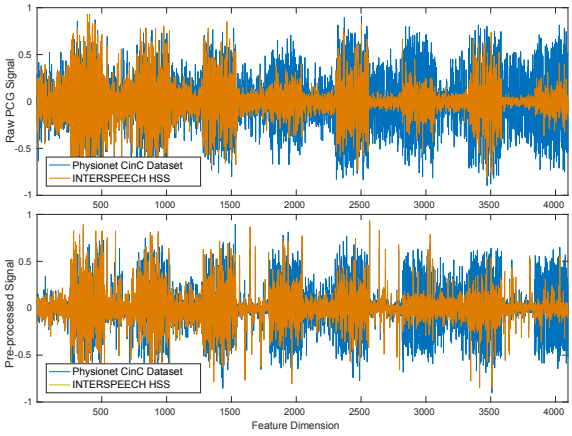

Figure 4: *Mean values of the* 4096 *features learned from the 4 mel-spectrograms by the RNN-Autoencoders.*

with a variant of our proposed tConv layer acquired improved performance compared to the end to end deep learning baseline (END2YOU). Training on a larger corpus has provided an improved performance on the development set using representation learning with significantly reduced performance on the test set. Comp-SVM, RL-SVM and LP-tConv models are ensembled using a majority voting algorithm. It yields UAR of 57.92% on the development set, and UAR of 39.2% on the test set. To improve the *Normal* hit rate a hierarchical decision system is implemented where an LP-tConv network trained on Physionet/Cinc Database is first used for binary classification between *Normal* and *Abnormal* recordings. Following that, an ensemble of Comp-SVM, RL-SVM and LP-tConv is used to classify between *Mild* and *Severe* classes. The hierarchical model has acquired a dev set UAR of 57.93% and test set UAR of 42.1%.

7. Discussion

Our proposed end to end LP-tConv model superseded the test set metric for the standalone baseline end to end model (END2YOU). Other proposed systems failed to beat the baseline systems test set UAR while it outperformed the development set UAR. This could indicate overfitting on the development set. On the other hand, for the baseline systems a tendency

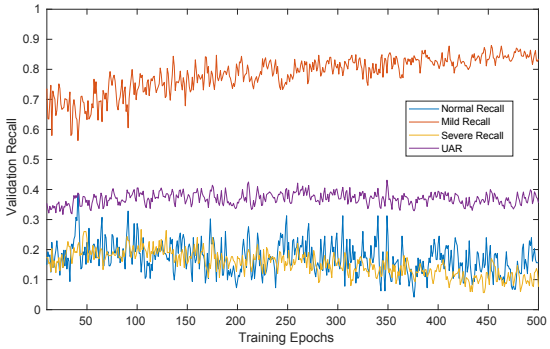

Figure 5: *Recall scores obtained on the validation data after each training epoch. A steady increase in the mild recall is visible while recall for the other classes are steadily decreasing.*

of overfitting on the test set was visible. This is because the individual approach/hyperparameters performing best on the test set has been chosen as baselines [15]. A generalized feature-classifier system should yield similar UAR on both development and test dataset if the development and test data distributions are consistent. This was noticeable only for the openSMILE features with an SVM classifier.

More interesting insights were revealed during the training of the recurrent autoencoders. The lower dimensional representations learned were different for the Physionet CinC Challenge database and the INTERSPEECH ComParE HSS database. The RL model was trained on both RL-data and the INTERSPEECH HSS database. Fig. 4 shows the mean of the concatenated (fused) representations learned from the 4 mel-spectrograms. A distinct difference can be visualized from feature dimension 1700. The last 2048 dimensions are representations learned from the -60 dB and the -75 dB mel-spectrograms, these are the dimensions where the feature means deviate the most. Quite interestingly, the -60 dB and -75 dB spectrogram features yield better results compared to the others. After training the model with preprocessed signals (resampled to 1000 Hz and band-pass filtered between 20-400 Hz), the representation differences in the mean reduced for certain dimensions. This could mean that the corresponding dimensions represent information from the higher end of the frequency spectrum. Another observation experienced during experimentation was the Normal Recall vs Mild/Severe Recall trade-off. While training an end to end LP-tConv model, we have seen a divergent behavior between the normal and mild/severe recall metrics (Fig. 5) which persisted even when the percentage of *Normal* recordings were more than *Mild* recordings.

8. Conclusions

In this work, we have presented an ensemble of classifiers for automatically detecting abnormal heart sounds of different severity levels for the INTERSPEECH 2018 ComParE Heart Beats Sub-Challenge. The primary framework was based on transfer learning of parameters from a 1D-CNN model pre-trained on the Physionet HS Classification dataset. We have also deployed unsupervised feature representation learning from mel-spectrograms using a deep autoencoder based architecture. Finally, we have also implemented a segment-level feature based system using the ComParE feature set and an SVM classifier. The final hierarchical ensemble of the systems provided with a UAR of 57.9% on the development dataset and 42.1% on the test dataset.

9. References

[1] U. Alam, O. Asghar, S. Q. Khan, S. Hayat, and R. A. Malik, "Cardiac auscultation: an essential clinical skill in decline," *Br. J. Cardiology*, vol. 17, no. 1, p. 8, 2010.

[2] H. Uğuz, "A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases," *J. Med. Syst.*, vol. 36, no. 1, pp. 61–72, 2012.

[3] A. Gharehbaghi, P. Ask, M. Lindén, and A. Babic, "A novel model for screening aortic stenosis using phonocardiogram," in *Proc. NBCBME*. Springer, 2015, pp. 48–51.

[4] R. SaraçOğLu, "Hidden markov model-based classification of heart valve disease with pca for dimension reduction," *J. Eng. Appl. Artif. Intell.*, vol. 25, no. 7, pp. 1523–1528, 2012.

[5] M. N. Homsi and P. Warrick, "Ensemble methods with outliers for phonocardiogram classification," *Physiol. Meas.*, vol. 38, no. 8, p. 1631, 2017.

[6] I. J. D. Bobillo, "A tensor approach to heart sound classification," in *Proc. IEEE CinC*, 2016, pp. 629–632.

[7] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-HSMM-based heart sound segmentation," *IEEE Trans. on Biomed. Eng.*, vol. 63, no. 4, pp. 822–832, 2016.

[8] V. Maknickas and A. Maknickas, "Recognition of normal–abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients," *Physiol. Meas.*, vol. 38, no. 8, p. 1671, 2017.

[9] F. Plesinger, I. Viscor, J. Halamek, J. Jurco, and P. Jurak, "Heart sounds analysis using probability assessment," *Physiol. Meas.*, vol. 38, no. 8, p. 1685, 2017.

[10] B. M. Whitaker, P. B. Suresha, C. Liu, G. D. Clifford, and D. V. Anderson, "Combining sparse coding and time-domain features for heart sound classification," *Physiol. Meas.*, vol. 38, no. 8, p. 1701, 2017.

[11] E. Kay and A. Agarwal, "Dropconnected neural networks trained on time-frequency and inter-beat features for classifying heart sounds," *Physiol. Meas.*, vol. 38, no. 8, p. 1645, 2017.

[12] M. Zabihi, A. B. Rad, S. Kiranyaz, M. Gabbouj, and A. K. Katsaggelos, "Heart sound anomaly and quality detection using ensemble of neural networks without segmentation," in *Proc. IEEE CinC*, 2016, pp. 613–616.

[13] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy, "Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds," in *Proc. IEEE CinC*, 2016, pp. 621–624.

[14] T.-c. I. Yang and H. Hsieh, "Classification of acoustic physiological signals based on deep learning neural networks with augmented features," in *Proc. IEEE CinC*, 2016, pp. 569–572.

[15] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *Proc. ISCA Interspeech*, 2018.

[16] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. Johnson *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiol. Meas.*, vol. 37, no. 12, p. 2181, 2016.

[17] C. D. Papadaniil and L. J. Hadjileontiadis, "Efficient heart sound segmentation and extraction using ensemble empirical mode decomposition and kurtosis features," *IEEE J. Biomed. Health. Inform.*, vol. 18, no. 4, pp. 1138–1152, 2014.

[18] S. 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 measurement*, vol. 31, no. 4, p. 513, 2010.

[19] R. Matei and G. Liviu, "A class of circularly-symmetric CNN spatial linear filters," vol. 19, pp. 299–316, 01 2006.

[20] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Proc. ISCA Interspeech*, 2015.

[21] A. I. Humayun, S. Ghaffarzadegan, Z. Feng, and T. Hasan, "Learning front-end filter-bank parameters using convolutional neural networks for abnormal heart sound detection," in *Proc. IEEE EMBC*. IEEE, 2018.

[22] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Adv. Neural. Inf. Process. Syst.*, 2014, pp. 3320–3328.

[23] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Adv. Neural. Inf. Process. Syst.*, 2014, pp. 3104–3112.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[25] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. of the DCASE 2017 Workshop*, 2017.

[26] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *arXiv preprint arXiv:1712.04382*, 2017.

[27] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. ACMMM*. ACM, 2010, pp. 1459–1462.