Learning Image-based Representations for Heart Sound Classification

Conference Paper · April 2018

DOI: 10.1145/3194658.3194671

Zhao Ren
Universität Augsburg
12 PUBLICATIONS **38** CITATIONS

Nicholas Cummins
Universität Augsburg
66 PUBLICATIONS **515** CITATIONS

Vedhas Pandit
Universität Augsburg
15 PUBLICATIONS **79** CITATIONS

Jing Han
Universität Augsburg
19 PUBLICATIONS **92** CITATIONS

Learning Image-based Representations for Heart Sound Classification

Zhao Ren
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
zhao.ren@informatik.uni-augsburg.de

Nicholas Cummins
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
nicholas.cummins@ieee.org

Vedhas Pandit
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
vedhas.pandit@informatik.uni-augsburg.de

Jing Han
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
jing.han@informatik.uni-augsburg.de

Kun Qian
Machine Intelligence and Signal Processing Group, Technische Universität München, Germany
andykun.qian@tum.de

Björn Schuller
GLAM – Group on Language, Audio & Music, Imperial College London, UK
schuller@ieee.org

ABSTRACT

Machine learning based heart sound classification represents an efficient technology that can help reduce the burden of manual auscultation through the automatic detection of abnormal heart sounds. In this regard, we investigate the efficacy of using the pre-trained Convolutional Neural Networks (CNNs) from large-scale image data for the classification of Phonocardiogram (PCG) signals by learning deep PCG representations. First, the PCG files are segmented into chunks of equal length. Then, we extract a scalogram image from each chunk using a wavelet transformation. Next, the scalogram images are fed into either a pre-trained CNN, or the same network fine-tuned on heart sound data. Deep representations are then extracted from a fully connected layer of each network and classification is achieved by a static classifier. Alternatively, the scalogram images are fed into an end-to-end CNN formed by adapting a pre-trained network via transfer learning. Key results indicate that our deep PCG representations extracted from a fine-tuned CNN perform the strongest, 56.2 % mean accuracy, on our heart sound classification task. When compared to a baseline accuracy of 46.9 %, gained using conventional audio processing features and a support vector machine, this is a significant relative improvement of 19.8 % ($p < .001$ by one-tailed z-test).

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**; • **Applied computing → Health care information systems**; **Health informatics**;

KEYWORDS

Heart Sound Classification, Phonocardiogram, Convolutional Neural Networks, Scalogram, Transfer Learning.

ACM Reference Format:
Zhao Ren, Nicholas Cummins, Vedhas Pandit, Jing Han, Kun Qian, and Björn Schuller. 2018. Learning Image-based Representations for Heart Sound Classification. In *Proceedings of Digital Health (DH'18)*. ACM, April 23-26, 2018, Lyon, France, 5 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Heart disease continues to be a leading worldwide health burden [16]. Phonocardiograph is a method of recording the sounds and murmurs made by heart beats, as well as the associated turbulent blood flow with a stethoscope, over various locations in the chest cavity [11]. *Phonocardiogram* (PCG), as the product of phonocardiograph, is widely employed in the diagnosis of heart disease. Enhancing conventional heart diseases diagnostic methods using the state-of-the-art automated classification techniques based on PCG recordings, is a rapidly growing field of machine learning research [13]. In this regard, the recent PhysioNet/ *Computing in Cardiology* (CinC) Challenge in 2016 [3], has encouraged the development of heart sound classification algorithms, by collecting multiple PCG datasets from different groups to construct a large, more than 20 hours of recordings, heart sound database. The two-class classification of normal/ abnormal heart sound was the core task of the PhysioNet/ CinC Challenge 2016.

In recent years, *Convolutional Neural Networks* (CNNs) have proven to be effective for a range of different signals and image classification tasks [7, 9]. In particular, large-scale CNNs have revolutionised visual recognition tasks as evidenced by their performances in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) [23]. On the back of the challenge, a large number of pre-trained CNNs have been made publicly available, such as AlexNet [10] and VGG [30]. Similarly, CNNs have also been successfully used for the detection of abnormal heart sounds [14].

Herein, we utilise the *Image Classification CNN* (ImageNet) to process *scalogram images* of PCG recordings for abnormal heart sound detection. Scalogram images are constructed using wavelet

Björn Schuller is also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany.

transformations [22]. Wavelets are arguably the predominate feature representation used for heart sound classification [8], and have successfully been applied in other acoustic classification tasks [19–21]. Moreover, instead of training CNNs from scratch, which can be a time-consuming task due in part to the large hyperparameter space associated with CNNs, we explore the benefits of using the aforementioned pre-trained ImageNet to construct robust heart sound classification models. Such an approach has been employed in other acoustic classification paradigms [1, 4], but to the best of the authors' knowledge it has not been verified for PCG based heart sound classification. Further, we also explore if transfer-learning based adaptation and updating of the ImageNet parameters can further improve the accuracy of classification.

The remainder of this paper is structured as follows: first, we describe our proposed approach in Section 2; the database description, experimental set up, evaluation method and results are then presented in Section 3; finally, our conclusions and future work plans are given in Section 4.

2 METHODOLOGY

In this section, we describe the classification paradigms we test for abnormal heartbeat detection. This consists of: (i) a conventional audio-based baseline system; (ii) two deep PCG representation systems combined with a *Support Vector Machine* (SVM) classifier; (iii) two end-to-end CNN based systems.

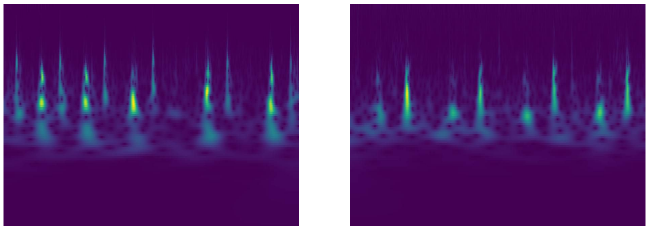
2.1 Baseline Classification System

As a baseline, we use a system based on the INTERSPEECH COMPUTATIONAL PARALINGUITIC CHALLENGE (ComParE) audio feature set [29], and SVM classification. The combination of ComParE features and SVM have been used in a range of similar acoustic classification tasks such as snore sound classification [28]. The ComParE feature set is a 6373 dimensional representation of an audio instance and is extracted using the openSMILE toolkit [5]; full details of the audio features presented in ComParE can be found in [5].

2.2 Scalogram Representation

In this study, to transform the PCG samples into images which can be processed by an ImageNet, the scalogram images are generated using the *morse* wavelet transformation [17] with 2 kHz sampling frequency. We have previously successfully used these scalogram images for acoustic scene classification [21]. When creating the images, we represent frequency, in *kHz*, on the vertical axis, and time, in *s*, on the horizontal axis. We use the *viridis* colour map, which varies from blue (low range) to green (mid range) to red (upper range), to colour the wavelet coefficient values. Further, the axes and margins marking are removed to ensure only the necessary information is fed into the ImageNet. Finally, the scalogram images are scaled to 224 × 224 for compatibility with the VGG16 ImageNet [30], which will be introduced in Section 2.3.

The scalogram images of a normal and an abnormal heartbeat are given in Figure 1. It can even be observed by human eyes that, there are some clear distinctions between the two classes in these (exemplar) images.



(a) Normal (*a0007.wav*) (b) Abnormal (*a0001.wav*)

Figure 1: The scalogram images are extracted from the first 4 s segments of normal/ abnormal heart sounds using the *viridis* colour map. The samples from which these scalogram images have been extracted are described in parentheses.

2.3 Convolutional Neural Networks

We use ImageNet to process the scalogram images for heart sound classification. The VGG16 ImageNet is chosen due to its successful application in the ILSVRC Challenge[1]. VGG16 is constructed from 13 ([2, 2, 3, 3, 3]) convolutional layers, five maxpooling layers, three fully connected layers {*fc6*, *fc7*, *fc*} and a soft-max layer for 1000 labels according to the image classification task in the ImageNet Challenge. The receptive field size of 3 × 3 is used in all of the convolutional layers. The full details of VGG16 are described in [30]. The structure and parameters of VGG16 are obtained from Pytorch[2]. Further, we use VGG16 for either *feature extraction* or *classification* by transfer learning, both of which are described in the following sub-sections.

2.4 Deep PCG Feature Representations

ImagNet has gathered considerable research interest as a feature extractor for a task of interest, e. g., [1]. In this regard, this subsection presents two methodologies for unsupervised PCG feature extraction using VGG16.

2.4.1 PCG Feature Extraction from ImageNet. The activations of the first fully connected layer *fc6* of VGG16 are employed as our feature representations. These features have previously proven to be effective in the task of acoustic scene classification [21]. Essentially, we feed the scalogram images into VGG16 and then the deep PCG feature representations of 4096 attributes are extracted as the activations of all neurons in the first fully connected layer *fc6*.

2.4.2 PCG Feature Extraction from adapted ImageNet. As VGG16 is normally employed for image classification tasks on a very different data from that required for heart sound classification, the feature extraction method described in the previous sub-section may yields a sub-optimal feature representation. We therefore also employ a transfer learning methodology (see Section 2.5.2) to adapt the parameters of VGG16 to better suit the task of abnormal heart sound detection. After the adaptation according to Section 2.5.2, the scalogram images are fed into the updated CNN model and a new

[1] http://www.image-net.org/challenges/LSVRC/
[2] http://pytorch.org/

set of deep representations (also with 4096 attributes) are extracted from the first fully connected layer *fc6*.

2.4.3 Classification Methodology. We perform classification of the heart sound samples into one of two classes: normal and abnormal. The process is achieved for the deep PCG feature representations via a linear SVM; the robustness of SVM for such a classification task is well-known in the literature [6]. Herein, our two deep feature representations are denoted as *pre-trained VGG+SVM* for the set-up described in Section 2.4.1 and *learnt VGG+SVM* for the set-up described in Section 2.4.2.

2.5 End-to-end ImageNet based Classification

With the aim of constructing a robust end-to-end heart sound CNN classifier, we adapt the parameters of VGG16 on the heart sound data by transfer learning. To achieve this, we use two different approaches, both of which are described below.

2.5.1 Learning Classifier of ImageNet. Noting that there are three fully connected layers in VGG16, we create our ImageNet classifier, herein denoted as *learning Classifier of VGG16*, by freezing the parameters of the convolutional layers and *fc6*, and updating (using scalogram images of heart sound data) the parameters of the final two fully connected layers and the soft-max layer for classification.

2.5.2 Learning ImageNet. In this method, herein denoted as *learning VGG*, we replace the last fully connected layer with a new one which has 2 neurons and a soft-max layer in order to achieve the 2-class classification task. We then update the *entire* network (again, using scalogram images of heart sound data) so that *all* VGG16 parameters are adapted to the heart sound data. This method represents a faster way to achieve a full CNN based classification than training an entire CNN from scratch with random initialisation of parameters.

2.6 Late-fusion Strategy

As the PCG recordings in the PhysioNet/ CinC Challenge are of varying lengths (cf. Section 3.1), we segment the recordings into non-overlapping chunks of 4 seconds. We therefore employ a late-fusion strategy to produce a single label (normal/ abnormal) per recoding. Our strategy is based on the probabilities of predictions, $p_i, i = 1, \ldots n$ of each i-th segment of a PCG sample, as outputted by the SVM or the soft-max layer; we choose the label of a PCG sample according to the highest probability $\max\{p_i\}$ gained from each chunk.

3 EXPERIMENTS

3.1 Database

Our proposed approaches are evaluated on the database of PhysioNet/ CinC Challenge 2016 [12]. This dataset is focused on classification of normal and abnormal heart sound recordings. As the test set labels for this data are not publicly available, we use the training set of the database and split it into a new training/ development/ test set. There are totally 3240 heart sound recordings collected from 947 pathological patients and healthy individuals. The dataset consists of six sub-databases from different research groups:

(1) **MIT heart sounds database:** *The Massachusetts Institute of Technology heart sounds database* (MIT) [31, 32] comprises 409 PCG and ECG recordings sampled at 44.1 kHz with 16 bit quantisation from 121 subjects, in which there are 117 recordings from 38 healthy adults and 134 recordings from 37 patients. The recording duration varies from 9 s to 37 s with a 32 s average length.

(2) **AAD heart sounds database:** *Aalborg University heart sounds database* (AAD) [25–27] is recorded at a 4 kHz sample rate and 16 bit quantisation. It contains 544 recordings from 121 healthy adults and 151 recordings from 30 patients. The recording length varies from 5 s to 8 s with an 8 s average length.

(3) **AUTH heart sounds database:** *The Aristotle University of Thessaloniki heart sounds database* (AUTH) [18] includes 45 recordings in total from 11 healthy adults and 34 patients. Each healthy adult/ patient gives one recording and the recording length varies from 10 s to 122 s with a 49 s average length. The sampling rate is 4 kHz with 16 bit quantisation.

(4) **UHA heart sounds database:** *The University of Haute Alsace heart sounds database* (UHA) [15] is sampled at 8 kHz with a 16 bit quantisation. It contains 39 recordings from 25 healthy adults and 40 recordings from 30 patients. The recording length varies from 6 s to 49 s with a 15 s average length.

(5) **DLUT heart sounds database:** *The Dalian University of Technology heart sounds database* (DLUT) [33] includes 338 recordings from 174 healthy adults and 335 recordings from 335 patients. The recording length varies from 8 s to 101 s with a 23 s average length. The sampling rate is 8 kHz with a 16 bit quantisation.

(6) **SUA heart sounds database:** *The Shiraz University adult heart sounds database* (SUA) [24] is constructed from 81 recordings from 79 healthy adults and 33 recordings from 33 patients. Except for three recordings sampled at 44.1 kHz and one at 384 kHz, the sampling rate is 8 kHz with 16 bit quantisation. The recording duration varies from 30 s to 60 s with a 33 s average length.

A detailed overview of database is described in Table 1. In this work, we split the dataset into a training set (including MIT, AUTH, UHA, and DLUT) and a test set (including AAD and SUA). Further, we carry out a three-fold cross validation by excluding the database MIT, AUTH, or UHA for fold 1, fold 2, or fold 3 (c. f., Table 2) correspondingly for validation, noting that due to its large size, DLUT is always for system training.

3.2 Setup

We generate scalogram images using the Matlab-2017 wavelet toolbox[3]. During training/ adaptation of VGG16, both for last two layers of VGG16 (cf. Section 2.5.1), and the entire network (cf. Section 2.5.2), the *learning rate* is 0.001, the *batch size* is 64, and the *epoch* is set as 50. The *cross entropy* is applied as the loss function and *stochastic gradient descent* is used as the optimiser. The deep representations (cf. Section 2.4), with a dimensionality of 4096, are extracted from *fc6* of VGG16.

[3] https://de.mathworks.com/products/wavelet.html

Table 1: An overview of the training and test partitions used in this work. The training set is structured by four sub-sets from four different databases of the PhysioNet/ CinC dataset, and the test set is by two. The PCG recordings in this dataset are annotated by the two-class labels (normal/ abnormal).

Dataset	Database	Recordings	Normal	Abnormal	Durations (s)			
					Total	Min	Max	Average
Training	MIT	409	117	292	13328.08	9.27	36.50	32.59
	AUTH	31	7	24	1532.49	9.65	122.00	49.44
	UHA	55	27	28	833.14	6.61	48.54	15.15
	DLUT	2141	1958	183	49397.15	8.06	101.67	23.07
Total		2636	2109	527	65090.86			
Test	AAD	490	386	104	3910.20	5.31	8.00	7.98
	SUA	114	80	34	3775.45	29.38	59.62	33.12
Total		604	466	138	7685.65			

Table 2: Performances comparison of the proposed approaches with baseline. The methods are evaluated on the 3-fold development set and the test set. The experimental results are evaluated by *Sensitivity* (*Se*), Specificity (*Sp*), and the *Mean Accuracy* (*MAcc*).

	Development set												Test set		
	fold 1			fold 2			fold 3			mean					
performance [%]	*Se*	*Sp*	*MAcc*	*Se*	*Sp*	*MAcc*	*Se*	*Sp*	*MAcc*	*Se*	*Sp*	*MAcc*	*Se*	*Sp*	*MAcc*
ComParE+SVM (baseline)	23.6	93.2	58.4	58.3	100.0	79.2	00.0	100.0	50.0	27.3	97.7	**62.5**	76.8	17.0	46.9
pre-trained VGG+SVM	57.2	70.9	64.1	41.7	85.7	63.7	17.9	81.5	49.7	38.9	79.4	59.1	24.6	87.1	55.9
learnt VGG+SVM	58.6	57.3	57.9	83.3	57.1	70.2	32.1	70.4	51.3	58.0	61.6	59.8	24.6	87.8	**56.2**
learning Classifier of VGG	68.2	51.3	59.7	79.2	14.3	46.7	35.7	40.7	38.2	61.0	35.4	48.2	33.3	63.7	48.5
learning VGG	83.6	40.2	61.9	95.8	28.6	62.2	53.6	44.4	49.0	77.7	37.7	57.7	12.3	95.7	54.0

When classifying by SVM, we use the LIBSVM library [2] with a linear kernel and optimise the SVM complexity parameter C $\in [10^{-5}; 10^{+1}]$ on the development partition. We present the best results from this optimisation as the final result.

3.3 Evaluation Method

According to the official scoring mechanism of the PhysioNet/ CinC Challenge 2016 [12], our predictions are evaluated by both *Sensitivity* (*Se*) and *Specificity* (*Sp*). For two-class classification, *Se* and *Sp* are defined as:

$$Se = \frac{TP}{TP + FN}, \tag{1}$$

$$Sp = \frac{TN}{TN + FP}, \tag{2}$$

where TP denotes the number of true positive abnormal samples, FN denotes the number of false negative abnormal samples, TN denotes the number of true negative normal samples, and FP denotes the number of false positive normal samples.

Finally, the *Mean Accuracy* (*MAcc*) is given as the overall score of the predictions, which is defined as:

$$MAcc = (Se + Sp)/2. \tag{3}$$

3.4 Results

The experimental results of the baseline and proposed methods are shown in Table 2. All CNN-based approaches achieve improvements in *MAcc* over the baseline on test set. Although this consistency is not seen on the development set, the *MAcc*s achieved on the test set indicate that the deep representation features extracted from scalogram images perform stronger and more robust than conventional audio features when performing heart sound classification.

When comparing the methods 'learning Classifier of VGG' and 'learning VGG', it is clear from the results that adapting the entire CNNs is definitely more effective than only updating the last two fully connected layers. Moreover, an in-general trend of the SVM classification of features extracted from either the pre-trained or the learnt VGG topologies performing stronger than the CNN classifiers can be observed. This could be due in part to the SVM classifiers being better to suit to the relatively smaller amounts of training data available in the PhysioNet/ CinC dataset than the soft-max classifiers.

Finally, the strongest performance, 56.2 % *MAcc*, is obtained on the test set using the method 'learnt VGG+SVM'. This *MAcc* offers a significant relative improvement of 19.8 % on our baseline classifier ($p < .001$ by one-tailed z-test), ComParE features and a SVM. Therefore, our learnt CNN model is shown to extract more salient deep representation features for abnormal heart sound detection when compared with features gained from the pre-trained VGG16 model.

4 CONCLUSIONS

We proposed to apply and adapt pre-trained *Image Classification Convolutional Neural Networks* (ImageNet) on scalogram images of *Phonocardiogram* (PCG) for the task of normal/ abnormal heart sound classification. Deep PCG representations extracted from a task-adapted version of the popular ImageNet VGG16 were shown to be more robust for this task than the widely used ComParE audio feature set. The combination of learnt VGG features and a SVM significantly ($p < .001$ by one-tailed z-test) outperformed the ComParE based baseline system. We speculate this success is due to the autonomous nature of the feature extraction associated with the 'learnt VGG' topology; the representations are adapted to the dataset and therefore are more robust than a 'fixed' conventional feature set.

In future work, data augmentation will be investigated for heart sound classification to compensate for the unbalanced nature of the dataset. Further, a new ImageNet topology based on the scalogram images will be developed and validated on a variety of heart sound datasets, e. g., AudioSet[4], to build a robust ImageNet for heart sound classification.

ACKNOWLEDGMENTS

This work was partially supported by the German national BMBF IKT2020-Grant under grant agreement No. 16SV7213 (EmotAsS), the EU Horizon 2020 Innovation Action Automatic Sentiment Estimation in the Wild under grant agreement No. 645094 (SEWA), and the EU H2020 / EFPIA Innovative Medicines Initiative under grant agreement No. 115902 (RADAR-CNS).

REFERENCES

[1] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 2017. Snore sound classification using image-based deep spectrum features. In *Proc. INTERSPEECH.* Stockholm, Sweden, 3512–3516.

[2] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (Apr. 2011), 1–27.

[3] Gari D. Clifford, Chengyu Liu, Benjamin Moody, David Springer, Ikaro Silva, Qiao Li, and Roger G. Mark. 2016. Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016. In *Proc. Computing in Cardiology Conference (CinC).* Vancouver, Canada, 609–612.

[4] Jun Deng, Nicholas Cummins, Jing Han, Xinzhou Xu, Zhao Ren, Vedhas Pandit, Zixing Zhang, and Björn Schuller. 2016. The University of Passau open emotion recognition system for the multimodal emotion challenge. In *Proc. CCPR.* Chengdu, China, 652–666.

[5] Florian Eyben, Felix Weninger, Florian Groß, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proc. ACM Multimedia.* Barcelona, Spain, 835–838.

[6] Steve R. Gunn. 1998. Support vector machines for classification and regression. *ISIS technical report* 14, 1 (May 1998), 5–16.

[7] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proc. CVPR.* Columbus, OH, 1725–1732.

[8] Edmund Kay and Anurag Agarwal. 2016. DropConnected neural network trained with diverse features for classifying heart sounds. In *Proc. Computing in Cardiology Conference (CinC).* Vancouver, Canada, 617–620.

[9] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS.* Lake Tahoe, NV, 1097–1105.

[11] Aubrey Leatham. 1952. Phonocardiography. *British Medical Bulletin* 8, 4 (1952), 333–342.

[12] Chengyu Liu et al. 2016. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement* 37, 12 (Nov. 2016), 2181–2213.

[13] Ilias Maglogiannis, Euripidis Loukis, Elias Zafiropoulos, and Antonis Stasis. 2009. Support vectors machine-based identification of heart valve diseases using heart sounds. *Computer Methods and Programs in Biomedicine* 95, 1 (July 2009), 47–61.

[14] Vykintas Maknickas and Algirdas Maknickas. 2017. Recognition of normal–abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients. *Physiological Measurement* 38, 8 (July 2017), 1671–1679.

[15] Ali Moukadem, Alain Dieterlen, Nicolas Hueber, and Christian Brandt. 2013. A robust heart sounds segmentation module based on S-transform. *Biomedical Signal Processing and Control* 8, 3 (May 2013), 273–281.

[16] Dariush Mozaffarian et al. 2016. Heart disease and stroke statistics–2016 update: A report from the American Heart Association. *Circulation* 133, 4 (Jan. 2016), e38–e360.

[17] Sofia C. Olhede and Andrew T. Walden. 2002. Generalized morse wavelets. *IEEE Transactions on Signal Processing* 50, 11 (Nov. 2002), 2661–2670.

[18] Chrysa D. Papadaniil and Leontios J. Hadjileontiadis. 2014. Efficient heart sound segmentation and extraction using ensemble empirical mode decomposition and kurtosis features. *IEEE Journal of Biomedical and Health Informatics* 18, 4 (July 2014), 1138–1152.

[19] Kun Qian, Christoph Janott, Vedhas Pandit, Zixing Zhang, Clemens Heiser, Winfried Hohenhorst, Michael Herzog, Werner Hemmert, and Björn Schuller. 2017. Classification of the excitation location of snore sounds in the upper airway by acoustic multifeature analysis. *IEEE Transactions on Biomedical Engineering* 64, 8 (Aug. 2017), 1731–1741.

[20] Kun Qian, Christoph Janott, Zixing Zhang, Clemens Heiser, and Björn Schuller. 2016. Wavelet features for classification of vote snore sounds. In *Proc. ICASSP.* Shanghai, China, 221–225.

[21] Zhao Ren, Vedhas Pandit, Kun Qian, Zijiang Yang, Zixing Zhang, and Björn Schuller. 2017. Deep sequential image features on acoustic scene classification. In *Proc. of DCASE Workshop.* Munich, Germany, 113–117.

[22] Olivier Rioul and Martin Vetterli. 1991. Wavelets and signal processing. *IEEE Signal Processing Magazine* 8, 4 (Oct. 1991), 14–38.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (Dec. 2015), 211–252.

[24] Maryam Samieinasab and Reza Sameni. 2015. Fetal phonocardiogram extraction using single channel blind source separation. In *Proc. ICEE.* Tehran, Iran, 78–83.

[25] Samuel E. Schmidt, Claus Holst-Hansen, Claus Graff, Egon Toft, and Johannes J. Struijk. 2010. Segmentation of heart sound recordings by a duration-dependent hidden Markov model. *Physiological Measurement* 31, 4 (Mar. 2010), 513–529.

[26] Samuel E. Schmidt, Claus Holst-Hansen, John Hansen, Egon Toft, and Johannes J. Struijk. 2015. Acoustic features for the identification of coronary artery disease. *IEEE Transactions on Biomedical Engineering* 62, 11 (Nov. 2015), 2611–2619.

[27] Samuel E. Schmidt, Egon Toft, Claus Holst-Hansen, and Johannes J. Struijk. 2010. Noise and the detection of coronary artery disease with an electronic stethoscope. In *Proc. CIBEC.* Cairo, Egypt, 53–56.

[28] Björn Schuller, Stefan Steidl, Anton Batliner, Elika Bergelson, Jarek Krajewski, Christoph Janott, Andrei Amatuni, Marisa Casillas, Amdanda Seidl, Melanie Soderstrom, Anne Ss Warlaumont, Guillermo Hidalgo, Sebastian Schnieder, Clemens Heiser, Winfried Hohenhorst, Michael Herzog, Maximilian Schmitt, Kun Qian, Yue Zhang, George Trigeorgis, Panagiotis Tzirakis, and Stefanos Zafeiriou. 2017. The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Proc. INTERSPEECH.* Stockholm, Sweden, 3442–3446.

[29] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. INTERSPEECH.* Lyon, France, 148–152.

[30] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR.* San Diego, CA, no pagination.

[31] Zeeshan Syed, Daniel Leeds, Dorothy Curtis, Francesca Nesta, Robert A. Levine, and John Guttag. 2007. A framework for the analysis of acoustical cardiac signals. *IEEE Transactions on Biomedical Engineering* 54, 4 (Apr. 2007), 651–662.

[32] Zeeshan Hassan Syed. 2003. *MIT automated auscultation system.* Ph.D. Dissertation. Massachusetts Institute of Technology.

[33] Hong Tang, Ting Li, Tianshuang Qiu, and Yongwan Park. 2012. Segmentation of heart sounds based on dynamic clustering. *Biomedical Signal Processing and Control* 7, 5 (Sep. 2012), 509–516.

[4] https://research.google.com/audioset/dataset/heart_sounds_heartbeat.html