Machine Learning with Applications 9 (2022) 100322

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Deep attention-based neural networks for explainable heart sound classification

Zhao Ren [a,b,*], Kun Qian [c,**], Fengquan Dong [d,e], Zhenyu Dai [f], Wolfgang Nejdl [b], Yoshiharu Yamamoto [g], Björn W. Schuller [a,h]

[a] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159, Germany
[b] L3S Research Center, Leibniz University Hannover, Hannover 30159, Germany
[c] School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China
[d] University of Hong Kong – Shenzhen Hospital, Shenzhen 518009, China
[e] Department of Cardiology, Shenzhen University General Hospital, Shenzhen 518055, China
[f] Department of Cardiovascular, Wenzhou Medical University First Affiliated Hospital, Wenzhou 325035, China
[g] Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo 113-0033, Japan
[h] GLAM – Group on Language, Audio & Music, Imperial College London, London SW7 2AZ, UK

ARTICLE INFO

Keywords:
Computer audition
Heart sound classification
Sensor signal processing
Digital health

ABSTRACT

Cardiovascular diseases are the leading cause of death and severely threaten human health in daily life. There have been dramatically increasing demands from both the clinical practice and the smart home application for monitoring the heart status of individuals suffering from chronic cardiovascular diseases. However, experienced physicians who can perform efficient auscultation are still lacking in terms of number. Automatic heart sound classification leveraging the power of advanced signal processing and deep learning technologies has shown encouraging results. Nevertheless, a lack of explanation for deep neural networks is a limitation for the applications of automatic heart sound classification. To this end, we propose explaining deep neural networks for heart sound classification with an attention mechanism. We evaluate the proposed approach on the heart sounds shenzhen corpus. Our approach achieves an unweighted average recall of 51.2% for classifying three categories of heart sounds, i. e., normal, mild, and moderate/severe. The experimental results also demonstrate that the global attention pooling layer improves the performance of the learnt representations by estimating the contribution of each unit in high-level features. We further analyse the deep neural networks by visualising the attention tensors.

1. Introduction

As reported by the World Health Organisation (WHO), Cardiovascular diseases (CVDs) are the first leading cause of death globally, which made 17.9 million people dead in 2016 (representing 31% of all global deaths) (Wolrd Health Organisation (WHO), 2017). More seriously, this number is predicted to be around 23 million per year by 2030 (Benjamin et al., 2019). Early-stage diagnosis and proper management of CVDs can be very beneficial to mitigate the high costs and social burdens by coping with serious CVDs (Hu et al., 2016; Schwamm et al., 2017). Auscultation of the heart sounds, as a cheap, convenient, and non-invasive method, has been successfully used by physicians for over a century (Dwivedi et al., 2018). However, this clinical skill needs tremendous training and is still difficult for more than 20% of the less experienced medical interns to efficiently use (Mangione, 2001). Therefore, developing an automatic auscultation framework can facilitate the early cost-effective screening of CVDs, and at the same time, manage the progression of its condition (Dwivedi et al., 2018).

Computer audition (CA) and its applications in healthcare (Qian, Li et al., 2020) have yielded encouraging results in the past decades. Due to its non-invasive and ubiquitous characteristic, CA-based methods can facilitate automatic heart sound analysis studies, which have already attracted a plethora of efforts (Dwivedi et al., 2018). Additionally, benefited from the fast development of machine learning (ML), particularly, its subsets, i. e., deep learning (DL), and the prevalent smart sensors, wearables, devices, etc., intelligent healthcare can be implemented feasibly in this era of AIoT (artificial intelligence-enabled internet of things). A systematical and comprehensive review of the existing literature on heart sound analysis via ML was provided in the study (Dwivedi et al., 2018). In the early works, designing efficient

* Corresponding author at: L3S Research Center, Leibniz University Hannover, Hannover 30159, Germany.
** Corresponding author.
E-mail addresses: zren@l3s.de (Z. Ren), qian@bit.edu.cn (K. Qian), fengquan.dong@foxmail.com (F. Dong), zhenyudai@foxmail.com (Z. Dai), nejdl@l3s.de (W. Nejdl), yamamoto@p.u-tokyo.ac.jp (Y. Yamamoto), schuller@informatik.uni-augsburg.de, bjoern.schuller@imperial.ac.uk (B.W. Schuller).

https://doi.org/10.1016/j.mlwa.2022.100322
Received 24 January 2022; Received in revised form 30 April 2022; Accepted 3 May 2022
Available online 14 May 2022

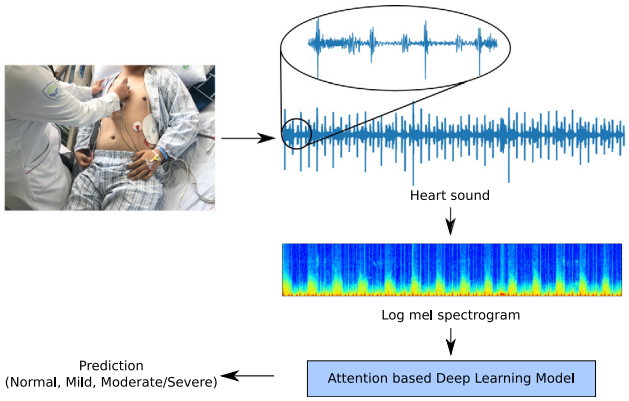

Fig. 1. The overview scheme of our heart sound classification procedure.

features ranging from classic Fourier transformation to multi-resolution analysis (e. g., wavelet transformation) dominated the well-documented literature in this field. In recent years, using DL models for analysing and extracting high-level representations from heart sounds has increasingly been studied (Clifford et al., 2017). Furthermore, as indicated in the study (Dong et al., 2020), the current trend is to classify the heart sounds from the whole audio recording without any segmentation step. On the one hand, the state-of-the-art DL methods aim to build a deep end-to-end architecture that can learn high-level representations from the heart sound itself without any human hand-crafted features. On the other hand, the DL models are restrained by the generalisation of the learnt representations from a limited data set. However, with the DL-based systems of heart sounds analysis, black-box DL models cannot produce transparent and understandable decisions for physicians to provide the next physical examination and appropriate treatment. Making explainable decisions via DL-based systems is a trend to enhance the trust of physicians in the systems and promote their application in the medical area (Holzinger et al., 2017). In the recent study (Xu et al., 2017), a promising attention mechanism was proposed to explain the DL models via visualising the internal layers.

To this end, we propose a novel attention-based deep representation learning method for heart sound classification in this study (Fig. 1). The proposed approach is validated on an open database, i. e., the Heart Sounds Shenzhen (HSS) database (Dong et al., 2020), hence rendering our studies reproducible and sustainable. The main contributions of our work are: First, by leveraging the power of a global attention pooling layer, the DL models can learn more robust and generalised high-level representations from the heart sound. Second, we make a comprehensive investigation and comparison of the topologies of DL models, i, e., convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Third, we compare the proposed method with other state-of-the-art approaches using the same database and standard processing. Fourth, we explore the visualisation of the learnt high-level representations of our proposed DL models using the attention mechanism, which can contribute to an explainable AI (XAI) (Adadi & Berrada, 2018). Last but not least, we indicate the current limitations and give our perspectives in this domain, which can be good guidance for future work.

The remainder of this paper will be structured as follows. First, a brief description of related work will be given in Section 2. Then, we introduce the database and methods used in Section 3. The experimental results and discussion are illustrated in Sections 4 and 5, respectively. Finally, we summarise our work in Section 6.

2. Related work

Classic machine learning. In the classic ML paradigm, human hand-crafted feature extraction is a prerequisite, which aims to design a series of efficient and robust features from the signals for specific tasks, e. g., heart sound classification. Among the features, wavelet transformation (WT) based representations showed efficient and excellent performance. For instance, wavelet features fed into the least square support vector machine (LSSVM) can enable to recognise the cases of normal, aortic insufficiency, aortic stenosis, atrial septal defect, mitral regurgitation, and mitral stenosis (Ari et al., 2010). Moreover, Uğuz (2012) designed entropy features of sub-bands by using discrete wavelet transformation (DWT) for classifying heart sounds. Similarly, tunable-Q wavelet transformation (TQWT) based features that characterise the various types of murmurs in cardiac sound signals were introduced in the study (Patidar et al., 2015). Wavelet packet transformation (WPT) based features were used in another study (Zheng et al., 2015), by which a full decomposition tree can be generated in a one-level decomposition process. Besides using the directly extracted low-level descriptors (LLDs) of the wavelet features, some high-level representations can also be derived. For example, auto-correlation features can be extracted from the sub-band envelopes that are calculated from the sub-band coefficients of the heart sound by DWT (Deng & Han, 2016). A combination of WT and WPT energy-based features combined with a deep RNN model was proposed in the study (Qian et al., 2019). Compared with the conventional short-time Fourier transformation (STFT) based features used for heart sound classification (Wang et al., 2007), wavelet features can provide a multi-resolution analysis of the non-stationary signals (heart sounds). This capacity helps to optimise the Heisenberg-alike time–frequency trade-off in time–frequency transformations (De Bruijn, 1967). Nevertheless, wavelet transformation still has its own drawbacks. In particular, designing a suitable *wavelet function* is not an easy job, which demands tremendous empirical experiments for specific tasks.

Deep Learning. Benefiting from the fast development of DL, the heart sound feature extraction can be realised without domain knowledge. Supervised learning and unsupervised learning approaches for analysing heart sounds are introduced as follows.

(i) Supervised Learning. The higher representations of the heart sounds can be automatically extracted from (pre-trained) CNNs and be fed into a classifier, e. g., SVM (Ren, Cummins et al., 2018). In recent work, Fernando et al. introduced the attention-based deep learning model for the heart sound segmentation task, and indicated that their model outperformed the state-of-the-art baseline methods (Fernando et al., 2020).

(ii) Unsupervised Learning. (Amiriparian et al., 2018) introduced an unsupervised representation learning method using an auto-encoder-based recurrent neural network in the paradigm of sequence-to-sequence (Seq2Seq) learning.

Attention Mechanism. With the generated high-level representations, most end-to-end deep representation learning methods, particularly CNNs and RNNs, use a global pooling layer to summarise the high-dimensional representations into one-dimensional vectors for later classification (Akhtar & Ragavendran, 2020; Ren, Kong et al., 2018). For example, global max-pooling selects the maximum value from each two-dimensional feature map in CNNs (Ren, Kong et al., 2018). Yet, our previous study has shown that global max-pooling loses the contribution of the other smaller values (Ren et al., 2019). Global attention pooling was proposed in the study (Ren, Kong et al., 2018) to improve the performance of CNNs through estimating the contribution of each unit in the feature maps to the classification task. An attention mechanism was also employed to explain the decisions via visualising the internal layers of DL models in the studies (Ren et al., 2019; Xu et al., 2017). With the inspiration of global attention pooling (Xu et al., 2017), we will show the effectiveness of CNNs with attention at the time–frequency level, and RNNs with attention at the time level, respectively. Notably, the input of the deep learning models is the log Mel spectrograms of heart sound signals.

Table 1
The data partitions, i. e., train, dev(elopment), and test sets, of the HSS corpus at the three classes, i. e., normal, mild, and mod(erate)/sev(ere), and subject numbers.

#	Subject	Normal	Mild	Mod./Sev.	Σ
Train	100	84	276	142	502
Dev	35	32	98	50	180
Test	35	28	91	44	163
Σ	170	144	465	236	845

3. Materials and methods

In this section, the HSS corpus, which were collected for heart sound classification, will be firstly introduced. Afterwards, two DL topologies, including a CNN and an RNN, are presented, and the attention mechanisms applied to each of them are described in detail. Finally, the evaluation metrics for the task of heart sound classification will be given.

3.1. HSS corpus

The HSS corpus was established by Shenzhen University General Hospital, Shenzhen, China (Dong et al., 2020). Please note that the study (Dong et al., 2020) was approved by the ethics committee of the Shenzhen University General Hospital. During the data collection, 170 participants (Female: 55, Male: 115, Age: 65.4 ± 13.2 years) were involved. Specifically, the heart sound signals were recorded from four positions on the body of each subject, including auscultatory mitral, aortic valve auscultation, pulmonary valve auscultation, and auscultatory areas of the tricuspid valve, through an electronic stethoscope (Eko CORE, USA) using Bluetooth 4.0 and 4 kHz sampling rate. Then, experienced cardiologists annotated the data into three categories: *normal*, *mild*, and *moderate/severe* by using Echocardiography as the golden standard. Finally, 845 audio recordings, each of which has around 30 s, were obtained, i. e., approximately 7 h. Considering subject-independency, and balanced age and gender distribution, the HSS corpus was split into three data sets: train, dev(elopment), and test sets (cf. Table 1). For more details on the HSS collection and further information, interested readers are suggested to refer to the study (Dong et al., 2020).

3.2. Deep learning models

In essence, DL is a series of non-linear transformations of the inputs, resulting in the highly abstract representations which have shown effectiveness in audio classification tasks (Amiriparian et al., 2017; Ren, Cummins et al., 2018). For this study, two typical DL topologies, i. e., a CNN (with a strong feature extraction capacity) and an RNN (which can capture the context information from time-series data), will be investigated.

3.2.1. Convolutional neural network

With a strong capability of feature extraction, CNN models have been applied to heart sound classification in previous research (Ryu et al., 2016; Tschannen et al., 2016). As shown in Fig. 2, a CNN model generally contains a stack of convolutional layers and local max-pooling layers to extract high-level representations. Convolutional layers capture abstract features using a set of convolutional kernels, which achieve convolution operations on the input or the feature maps from the intermediate layers. At the mth layer, $m = 1, \ldots, M$, where M is the total number of layers, an $I \times P \times Q$ feature map $\boldsymbol{h}_m$ is produced, where I is the number of channels, and $P \times Q$ stands for the size of $\boldsymbol{h}_m$ at each channel. While the $(m+1)$th layer is a convolutional layer, the jth channel of $\boldsymbol{h}_{m+1}$ is calculated by

$$\boldsymbol{h}^j_{m+1} = \sum_{i=1}^{I} \boldsymbol{w}^{ij}_{m+1} * \boldsymbol{h}^i_m + b^j_{m+1}, \tag{1}$$

where $\boldsymbol{h}^i_m$ is the ith channel of $\boldsymbol{h}_m$, $\boldsymbol{w}^{ij}_{m+1}$ denotes the (i,j)th convolutional kernel, $*$ is the convolutional operation, and b^j_{m+1} is the bias. Each two-dimensional convolutional kernel works on the feature maps at each channel, therefore the convolutional layers can learn the representations at the time–frequency level. Notably, batch normalisation and an activation function of rectified linear unit (ReLU) are utilised to deal with the output of each convolutional layer, as batch normalisation usually improves the stability of CNNs, and both of them can accelerate the convergence speed (Ide & Kurita, 2017).

Convolutional layers with batch normalisation and a ReLU activation function are mostly followed by local pooling layers, which reduce the computational cost via downsampling the feature maps (Kobayashi, 2019). Through local pooling layers, the robustness of CNNs is also improved against the input variation (Kobayashi, 2019). Since local max-pooling has been successfully employed in our previous study (Ren, Kong et al., 2018), we use local max-pooling layers following each convolutional layer.

3.2.2. Recurrent neural network

RNNs can extract sequential representations from time-series data using a set of recurrent layers (cf. Fig. 3). Each recurrent layer contains a sequence of recurrent units, each of which is used to process the corresponding time step of the input data. The hidden states, output from each recurrent layer, are fed into the next recurrent layer. Finally, the hidden states of the final recurrent layer are used to predict the classes of the samples.

We define the number of the total time steps by T. At the tth time step, $t = 1, \ldots, T$, a traditional recurrent unit computes its output via a weighted sum of the input x_t and the hidden state h_{t-1}. Due to the vanishing gradient problem caused by the traditional recurrent unit (Hochreiter, 1998), in particular, two recurrent units were proposed in the literature: Long Short-Term Memory (LSTM) cells (Hochreiter & Schmidhuber, 1997), and Gated Recurrent Units (GRUs) (Chung et al., 2014).

At the tth time step, an LSTM unit consists of an input gate i_t, an output gate o_t, a forget gate f_t, and a cell state c_t. The procedure of an LSTM unit is defined by

$$i_t = \sigma(\boldsymbol{w}_i x_t + \boldsymbol{u}_i h_{t-1} + b_i), \tag{2}$$

$$f_t = \sigma(\boldsymbol{w}_f x_t + \boldsymbol{u}_f h_{t-1} + b_f), \tag{3}$$

$$o_t = \sigma(\boldsymbol{w}_o x_t + \boldsymbol{u}_o h_{t-1} + b_o), \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(\boldsymbol{w}_c x_t + \boldsymbol{u}_c h_{t-1} + b_c), \tag{5}$$

$$h_t = o_t \odot \tanh(c_t), \tag{6}$$

where $\boldsymbol{w}$ and $\boldsymbol{u}$ are the weight matrices, b denotes the bias, σ stands for a logistic sigmoid function, and $\odot$ means the element-wise multiplication. Compared to the traditional recurrent unit, an LSTM cell can control what information to remember using an input gate, and what to forget using a forget gate.

Different from an LSTM cell, a GRU contains a reset gate r_t and an update gate z_t at the t time step. The procedure of a GRU is defined by

$$r_t = \sigma(\boldsymbol{w}_r x_t + \boldsymbol{u}_r h_{t-1} + b_r), \tag{7}$$

$$z_t = \sigma(\boldsymbol{w}_z x_t + \boldsymbol{u}_z h_{t-1} + b_z), \tag{8}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tanh(\boldsymbol{w}_h x_t + \boldsymbol{u}_h(r_t \odot h_{t-1}) + b_h). \tag{9}$$

With two gates inside one unit, a GRU has fewer parameters than an LSTM cell. As both LSTM–RNN and GRU–RNN have been employed in audio classification tasks (Dong et al., 2020; Ren, Qian et al., 2018), the effectiveness of them are explored in this study.

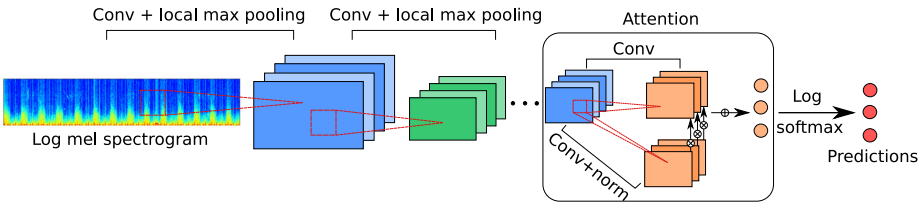

Fig. 2. The structure of the chosen CNN model with attention. The input are log Mel spectrograms. The CNN model consists of several convolutional layers, local max-pooling layers, an attention layer, and a log softmax layer for classification.

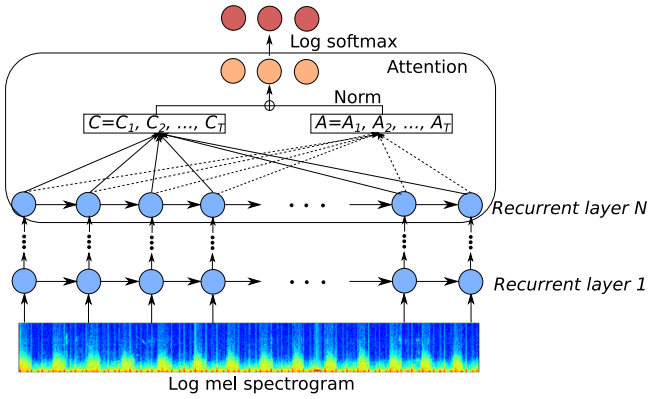

Fig. 3. The structure of the chosen RNN model. The RNN model learns sequential representations from the logmel spectrograms, and the features from the final recurrent layer are then processed by an attention layer and a log softmax layer for classification. In the attention layer, $A = A_1, A2, \ldots, A_T$ is the attention vector, and $C = C_1, C_2, \ldots, C_T$ is the classification vector, while T denotes the frame number of each log Mel spectrogram.

In a similar way to CNNs, a layer normalisation (Ba et al., 2016) and an activation function of scaled exponential linear unit (SELU) (Phankokkruad & Wacharawichanant, 2019) are set to follow each recurrent layer, since layer normalisation can stabilise the hidden state dynamics in RNNs (Ba et al., 2016), and the SELU activation function has been successfully applied in the previous study (Phankokkruad & Wacharawichanant, 2019).

3.3. Attention mechanism

It is essential to interpret the key parts of the input inside a deep learning model, especially in the applications of medical diagnosis. As aforementioned in Section 2, global attention pooling can evaluate the contribution of each unit in a representation. We will now introduce the attention mechanisms in CNNs and RNNs, respectively.

3.3.1. Attention in CNN

While a log Mel spectrogram is fed into a CNN model, the feature map $\boldsymbol{h}_M$ output by the final layer before the attention mechanism has three dimensions $I' \times P' \times Q'$, where I' is the number of channels, and $P' \times Q'$ denotes the feature map size at the time–frequency level. To achieve the heart sound classification, the dimensions of $\boldsymbol{h}_M$ are reduced from three into one. During this procedure of dimension reduction, the global attention pooling evaluates how much each time–frequency bin in $\boldsymbol{h}_M$ devotes to the final predictions by estimating a weight value for each bin. As shown in Fig. 2, the global attention pooling consists of two components: the top one has a convolution layer, and the bottom one is comprised of a convolutional layer and a normalisation operation. In the top component, the convolutional layer is set up with 1×1 kernels and an output channel of the class number. In the bottom component, the convolutional layer has the same hyperparameters as that in the top one. Afterwards, to calculate the weight tensor of $\boldsymbol{h}_M$, an activation function is employed to rectify the values of the feature map from the convolutional layer in the bottom component. Both softmax and sigmoid functions can rectify the values into the interval of $[0, 1]$. Further, normalisation is applied to the rectified feature map F using

$$F^* = \frac{F}{\sum_{p=1}^{P'} \sum_{q=1}^{Q'} F_{pq}}, \tag{10}$$

where F^* is the output of the bottom component. Next, the feature map from the top component is multiplied with F^*, leading to an element-wise product, which is then summed to a vector with the length equalling the number of classes. Finally, log softmax is employed to fit the chosen negative log-likelihood (NLL) loss function.

3.3.2. Attention in RNN

The representation from the recurrent layers has two dimensions $T \times Q''$, where Q'' denotes the length of the feature at each time frame. While summarising the representation to a vector for classification, it would be worthwhile to explain the essential time frames using global attention pooling in RNNs. As the length of the time frames is equal to that of the original log Mel spectrogram, an attention mechanism can show more details at the frame level in RNNs than in CNNs.

As shown in Fig. 3, the global attention pooling in RNNs also includes two components as the attention mechanism in CNNs. In a similar way to the attention mechanism in CNNs, the left component (corresponding to the top one in Fig. 2) herein contains a one-dimensional convolutional layer, in which the kernel size is 1 and the output channel number is equal to the class number, leading to a classification tensor C of size $T \times class\ number$. The right component (corresponding to the bottom one in Fig. 2) consists of a convolutional layer with the same setting as that in the left component and a normalisation procedure. In the right component, the convolutional layer is also followed by an activation function (softmax or sigmoid) to rectify the values of the representation. Then, normalisation is applied to the rectified representation A using

$$A^* = \frac{A}{\sum_{t=1}^{T} A_t}, \tag{11}$$

where A^* is the normalised feature in the right component. The element-wise product of A^* and C is then followed by a log softmax layer for the heart sound classification.

3.4. Evaluation metrics

To evaluate the performance of the proposed models, the unweighted average recall (UAR) is employed as the main evaluation metric by taking the imbalanced characteristic of the HSS database and the inherent phenomena into account. Compared to another popular evaluation metric, weighted average recall (WAR), *aka* accuracy, UAR shows more reasonable in measuring the performance of a model trained by imbalanced data (Schuller et al., 2009). The value of UAR is defined as:

$$\text{UAR} = \frac{\sum_{i=1}^{N_c} recall_i}{N_c}, \tag{12}$$

where $recall_i$ is the *recall* achieved for the ith class, and N_c denotes the number of classes ($N_c = 3$ in this study).

When comparing two methods' performances, we use a one-tailed z-test (Dietterich, 1998) by checking if a finding is significant ($p < 0.05$) or not. Additionally, the area under the receiver operating characteristic curve (AUC-ROC) is calculated to evaluate the performance.

4. Experimental results

We give a brief description of our experimental setup at first. Then, we present and discuss the results achieved in this study.

4.1. Setup

First, a series of 936 × 64 log Mel spectrograms are extracted from the audio signals in the HSS corpus using a Hamming window of 256 samples width with 50% overlap and 64 Mel frequency bins. During training, all models are learnt with an Adam optimiser and a batch size of 32. The initial learning rate is experimentally set to 0.0001, and is reduced into 90% at each 100-th iteration with the aim of stabilising the training process. The setting of the initial learning rate is based on the experiments of training models in our prior studies (Ren, Kong et al., 2018; Ren, Qian et al., 2018). Finally, the learnt models at the 3 000-th iteration are used to predict the audio samples in the development/test set.

The CNN models employ a flattening layer or a global pooling layer before the final log softmax layer for classification. The flattening layer is to flatten the multi-dimensional feature maps into a vector with retaining all units in each feature map. In contrast, the global pooling layer in CNN summarise the multi-dimensional feature maps into a vector with the length of the channel number by calculating the maximum or average values along the axes other than the channel axis, i. e., global max-pooling and global average-pooling, and the global pooling layer in RNN summarise the feature maps into a vector with the length of feature dimensions at each time step. In our prior study (Ren, Kong et al., 2018), global pooling outperformed flattening as it reduces redundant information from the feature maps. In this regard, we apply flattening and global pooling for comparison in this work. In contrast, the RNN models generally select the feature at the last time step for further procedure. To be consistent with the experiments on CNN models, the selection of the last time step is also compared to global pooling. The structures before the flattening or global pooling layer in the deep neural networks are empirically set as follows (cf. Table 2).

- The CNN models consist of four convolutional layers with output channels 64, 128, 256, and 256. Each convolutional layer is followed by a local max-pooling layer with 2 × 2 kernels.
- Both LSTM–RNN and GRU–RNN models contain three recurrent layers with output dimensions 256, 1 024, and 256.

Finally, the representations after the flattening layer/global pooling layer/last-time step selection are fed into a linear layer for the final prediction.

To investigate the effect of the balanced training set on the DL models, we compare the results on the original imbalanced HSS data and balanced HSS training data produced by a random upsampling strategy aiming at class balance (Zhang & Schuller, 2012). The random upsampling strategy randomly selects data samples from the class with less data to increase the sample number, therefore all classes will have the same sample number.

4.2. Results

The experimental results (UARs in [%]) of all three DL topologies (CNN, LSTM–RNN, and GRU–RNN) are shown in Table 3. The best result (a UAR of 51.2%) is achieved by the CNN model with an attention mechanism (using a sigmoid function). The best results for LSTM–RNN and GRU–RNN are 42.6% UAR and 46.8% UAR, respectively. To produce convincing results, subject-independent experiments are used in our work; therefore the performance is not very high. All of the best results are significantly better than the chance level (in a one-tailed z-test, $p < 0.001$ for CNN, $p < 0.05$ for LSTM–RNN, and $p < 0.01$ for GRU-RNN). We can see that, an attention-based mechanism can significantly improve the corresponding DL models in recognising heart sound. For instance, CNN with sigmoid-attention (a UAR of 51.2%) performs better than a CNN with flattening (a UAR of 37.6%), and a CNN with max-pooling (a UAR of 38.4%) (in a one-tailed z-test, $p < 0.01$), and a GRU–RNN with softmax-attention (a UAR of 46.8%) outperforms GRU–RNNs without attention (UARs of 36.1% and 35.2%) (in a one-tailed z-test, $p < 0.05$). The upsampling strategy can slightly improve the performances of the best RNN models. Compared to other state-of-the-art studies, our proposed method can perform better than most performances achieved by single models (cf. Table 4). Notably, the best performance 56.2% on the test set is achieved by the ConPaRE baseline with fusion methods, indicating fusion of multiple models is helpful to improve the performance. As our aim is to verify the effectiveness of attention and explain the attention models, the single models are used for comparison rather than embedding multiple models.

When looking at the confusion matrices (cf. Fig. 4), we find that the best CNN and GRU–RNN models outperform the best LSTM–RNN model in recognising the 'Mild' type of heart sounds. For all the three models, both the 'Normal' and 'Mod./Sev.' types of heart sounds are incorrectly recognised as the 'Mild' type of heart sounds. Figs. 5 and 6 present the macro-averaged receiver operating characteristic (ROC) curves and the visualisation of the best three proposed attention-based DL models on each class, respectively.

Table 2
The details of the employed CNN and RNN models in this work. "conv" is the convolutional layer, "ch" means the output channel number, "k" is the kernel size, and "out" denotes the output dimension of each RNN layer.

CNN	RNN
conv (ch:64, k: 5 × 5), local max-pooling (k: 2 × 2)	LSTM/GRU layer(out: 256)
conv (ch:128, k: 5 × 5), local max-pooling (k: 2 × 2)	LSTM/GRU layer(out: 1 024)
conv (ch:256, k: 5 × 5), local max-pooling (k: 2 × 2)	LSTM/GRU layer(out: 256)
conv (ch:256, k: 5 × 5), local max-pooling (k: 2 × 2)	
Flattening/global pooling/attention	Last-time step/global pooling/attention
Linear layer	

Table 3
The results comparison of different deep learning topologies on the HSS corpus.

	w/o upsampling		w/upsampling	
UAR [%]	Dev	Test	Dev	Test
CNN				
Flattening	35.6	37.6	35.6	39.9
Global max-pooling	41.7	38.4	39.3	38.5
Attention-softmax	31.5	43.1	38.3	47.3
Attention-sigmoid	40.1	**51.2**	39.6	50.5
LSTM–RNN				
Last-time step	39.3	36.1	40.7	35.7
Global max-pooling	32.9	38.9	34.6	38.1
Attention-softmax	40.0	39.6	39.0	39.4
Attention-sigmoid	39.6	38.9	42.0	**42.6**
GRU–RNN				
Last-time step	39.0	36.5	37.4	36.1
Max-pooling	38.7	35.8	40.7	35.2
Attention-softmax	30.8	44.7	35.3	**46.8**
Attention-sigmoid	34.9	44.2	34.5	45.7

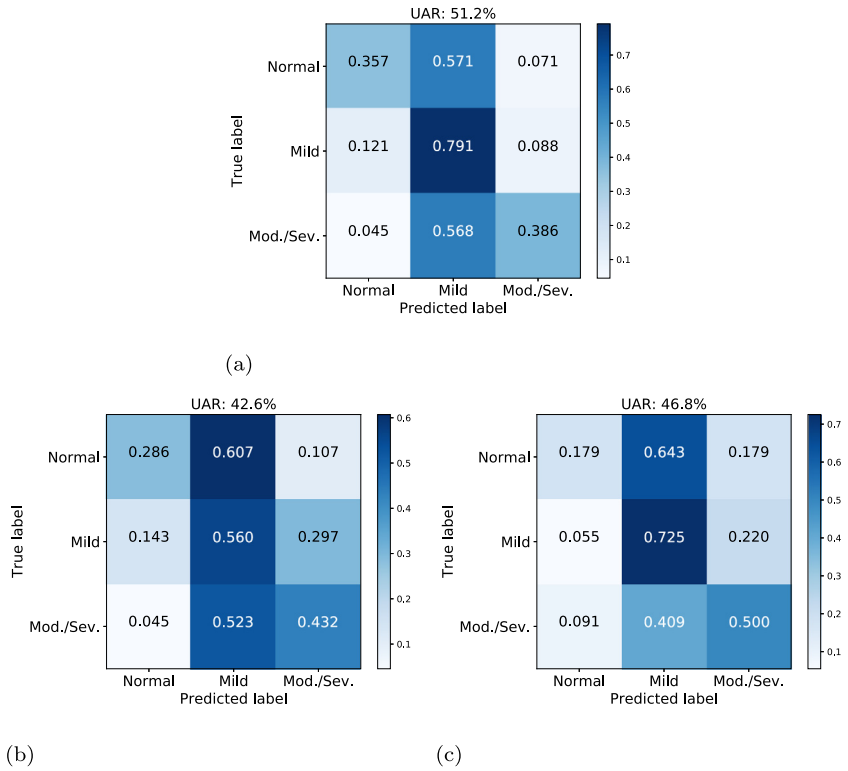

Fig. 4. Confusion matrices (normalised) achieved by the best models on the test set of the HSS corpus. The best three models are (a) CNN, (b) LSTM–RNN, and (c) GRU–RNN, respectively.

Table 4
The results comparison among the state-of-the-art methods and our proposed model on the HSS corpus.

UAR [%]	Dev	Test
ComParE baseline (End2You) (Schuller et al., 2018)	41.2	37.7
ComParE baseline (openSMILE) (Schuller et al., 2018)	50.3	46.4
ComParE baseline (openXBOW) (Schuller et al., 2018)	42.6	52.3
ComParE baseline (fusion) (Schuller et al., 2018)	–	**56.2**
Ensemble of transfer learning (Humayun et al., 2018)	57.9	42.1
Utterance-level feature and SVMs (Gosztolya et al., 2018)	53.2	49.3
Seq2Seq autoencoders and SVMs (Amiriparian et al., 2018)	35.2	47.9
Wavelets and RNNs (Qian et al., 2019)	–	43.0
Log Mel features and SVMs (Dong et al., 2020)	46.5	49.7
Our proposed approach	40.1	51.2

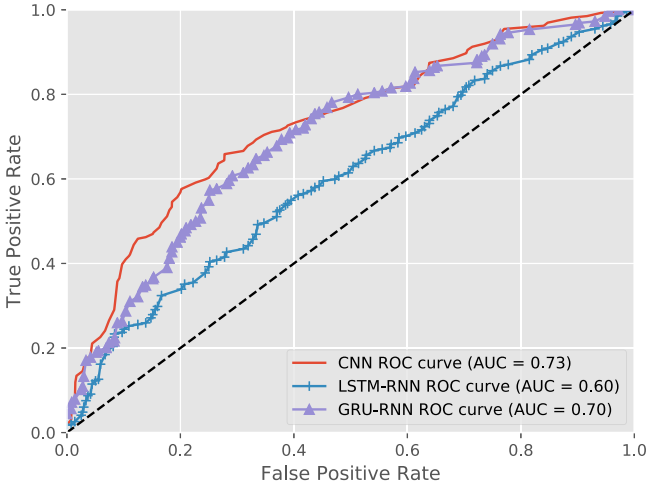

Fig. 5. Comparison of the macro-average receiver operating characteristic (ROC) curves of the best three models on the test set of the HSS corpus. The corresponding area-under-curve (AUC) is also computed for each model.

5. Discussion

In this section, we summarise the findings from this study. Afterwards, we indicate the limitations and future work by providing our perspectives.

5.1. Findings of this study

In most cases, the results on the test set are better than those on the development set. The reason might be that the models trained on both training and development sets are verified on the test set; the models trained on only the training set are verified on the development set. In Fig. 4, some samples with the classes of 'Normal' and 'Mod./Sev.' are classified into 'Mild', probably due to the data imbalance. A CNN model is found to be superior to an RNN in recognising heart sounds in this study. As shown in Fig. 5, the ROC curve of the considered CNN and GRU–RNN can yield a higher true positive rate at a given false positive rate compared to the LSTM–RNN, and the true positive rate of the CNN is superior to or comparable to that of our GRU–RNN. Finally, the area under the ROC curve (AUC) of the CNN is the highest in those of the three models.

As depicted in Fig. 6, compared to 'Normal' or 'Mild' types of heart sounds, the 'Mod./Sev.' type shows more irregular waveforms and spectrograms. In addition, by checking the learnt high-level representations of the CNN models, the 'Mod./Sev.' types of heart sounds can have a higher number of higher energy components than the other two types at similar frequency bands. Such irregular changes in frequency bands via the time axis of the heart sound might be caused by pathological changes in the heart. When looking at the learnt representations of the RNN models (cf. Fig. 6), we can see the periodic signal's characteristics in the 'Normal' types of the heart sound. It is worth exploring the

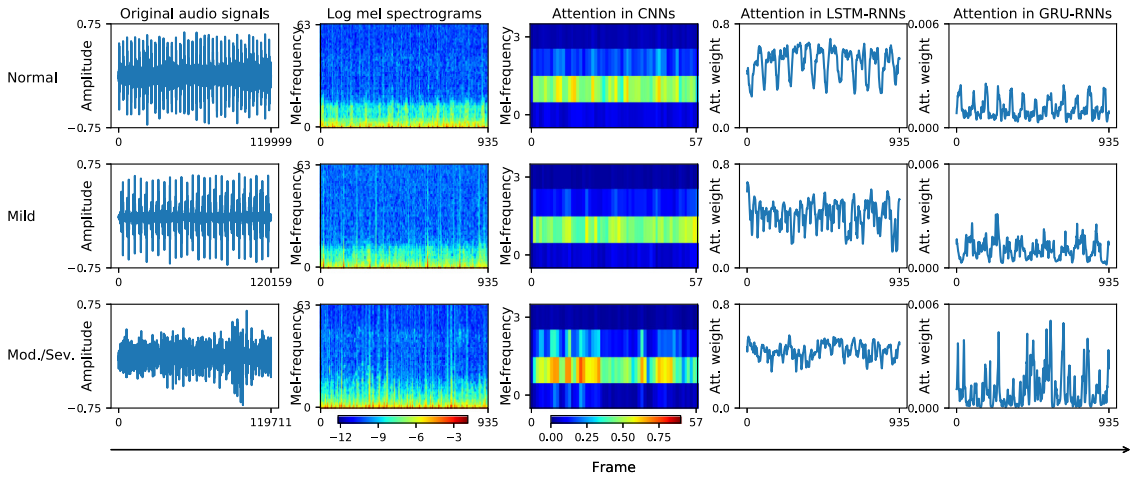

Fig. 6. Visualisation of three examples in the HSS corpus with the classes Normal, Mild, and Moderate/Severe, respectively. Each example consists of an original audio signal, its corresponding log Mel spectrogram, the attention matrix in the CNNs, the attention vector in the LSTM–CNNs, and the attention vector in the GRU–RNNs.

fundamental mechanism of CVDs and their corresponding properties in heart sound changes.

5.2. Ablation study

In this study, we explore the effectiveness of random upsampling for improving the performance. We find that simple data augmentation (upsampling) cannot yield significantly better results than using the original data set (cf. Table 3). We may think that the random upsampling technique cannot generate sufficiently informative instances for improving the models' performances.

5.3. Comparison with the state-of-the-art

Compared to the state-of-the-art approaches on the HSS corpus (cf. Table 4), our proposed approach performs better than most single-model methods and is comparable to the ComParE baseline with features extracted by openXBOW (Schuller et al., 2018). Particularly, our approach outperforms the log Mel features and SVMs in Dong et al. (2020), indicating that deep neural networks can better learn a non-linear mapping between the inputs and the labels than the traditional machine learning method.

5.4. Limitations and perspectives

Data size limitation is the biggest challenge in the current study. Moreover, similar to other clinical data studies, e. g., snore sound (Qian, Janott et al., 2020), asking experienced medical experts to annotate massive data is expensive, time-consuming, and even unavailable in practice. Even though the data augmentation did not show excellent performance in this study, it is a necessary step in improving the DL models' generalisation and robustness. More recently, some advanced data augmentation technologies, e. g., the *generative adversarial networks* (GANs) (Goodfellow et al., 2014) can be considered. In future work, we should explore using more sophisticated data augmentation technologies for heart sound classification. Moreover, *(labelled) data scarcity* is a challenging issue for almost all of the biomedical areas including heart sound. One should consider using unsupervised learning, semi-supervised learning, active learning, and cooperative learning paradigms in future studies.

The best model's result is encouraging but modest. In a future effort, one should consider using hybrid network architectures (Yu et al., 2017) or model fusion strategies (Qian et al., 2017). Even though we can find promising results achieved by the deep attention-based models, the inherited mechanism is still unclear. We tried to visualise the learnt representations of the hidden layers, but it failed to make any consolidated conclusion. Another direction is to explore the learnt representations by DL models, which aims to present the interpretations between the model architectures and the pathological meaning of the heart sound. An explainable AI is essential for intelligent medical applications.

6. Conclusion

In this work, we proposed a novel attention-based deep representation learning method for heart sound classification. We also investigated and compared different topologies of the DL models and found the considered CNN model as the best option in this study. The efficacy of the proposed method was successfully validated by the publicly accessible HSS corpus. We also compared the results with other state-of-the-art works and pointed out the current limitations and future directions. For a three-category classification task, the proposed approach achieved an unweighted average recall of 51.2%, which outperformed the other models trained by traditional human handcrafted features or other deep learning approaches. In future work, we will improve our model's generalisation and explainability for the heart sound classification task.

CRediT authorship contribution statement

Zhao Ren: Conceptualization, Methodology, Software, Writing – original draft. **Kun Qian:** Conceptualization, Methodology, Writing – original draft. **Fengquan Dong:** Data curation, Resources. **Zhenyu Dai:** Data curation, Resources. **Wolfgang Nejdl:** Writing – review & editing. **Yoshiharu Yamamoto:** Writing – review & editing. **Björn W. Schuller:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network (MSCA-ITN-ETN) project under grant agreement No. 766287 (TAPAS), Germany, the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS), Germany, the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor with grant No. 01DD20003, the BIT Teli Young Fellow Program from the Beijing Institute of Technology, China, the Zhejiang Lab's International Talent Fund for Young Professionals (Project HANAMI), China, the JSPS Postdoctoral Fellowship for Research in Japan (ID No. P19081) from the Japan Society for the Promotion of Science (JSPS), Japan, the Grants-in-Aid for Scientific Research (No. 19F19081 and No. 17H00878) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, the Ministry of Science and Technology of the People's Republic China (2021ZD0201900), the Medical Scientific Research Foundation of Guangdong Province of China (No. ,A2021333), Shenzhen Science and Technology Innovation Commission Project (No. JCYJ20190808120613189), and the Shenzhen Pea-cock Program-Project Development Fund (No. 20190904141C). The authors would like to thank the colleagues who collected the HSS corpus.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. http://dx.doi.org/10.1109/ACCESS.2018.2870052.

Akhtar, N., & Ragavendran, U. (2020). Interpretation of intelligence in CNN-pooling processes: A methodological survey. *Neural Computing and Applications*, *32*, 879–898. http://dx.doi.org/10.1007/s00521-019-04296-5.

Amiriparian, S., Freitag, M., Cummins, N., & Schuller, B. (2017). Sequence to sequence autoencoders for unsupervised representation learning from audio. In *Proc. DCASE Workshop* (pp. 17–21). Munich, Germany.

Amiriparian, S., Schmitt, M., Cummins, N., Qian, K., Dong, F., & Schuller, B. (2018). Deep unsupervised representation learning for abnormal heart sound classification. In *Proc. EMBC* (pp. 4776–4779). Honolulu, HI.

Ari, S., Hembram, K., & Saha, G. (2010). Detection of cardiac abnormality from PCG signal using LMS based least square SVM classifier. *Expert Systems with Applications*, *37*(12), 8019–8026. http://dx.doi.org/10.1016/j.eswa.2010.05.088.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. In *NIPS Deep Learning Symposium* (p. 14). Barcelona, Spain.

Benjamin, E. J., Muntner, P., & Bittencourt, M. S. (2019). Heart disease and stroke statistics-2019 update: a report from the American heart association. *Circulation*, *139*(10), e56–e528. http://dx.doi.org/10.1161/CIR.0000000000000659.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proc. NIPS deep learning and representation learning workshop* (pp. 1–9). Montreal, Canada.

Clifford, G. D., Liu, C., Moody, B., Millet, J., Schmidt, S., Li, Q., Silva, I., & Mark, R. G. (2017). Recent advances in heart sound analysis. *Physiological Measurement*, *38*(8), E10–E25. http://dx.doi.org/10.1088/1361-6579/aa7ec8.

De Bruijn, N. (1967). Uncertainty principles in Fourier analysis. In *Inequalities (Proc. Sympos. Wright-Patterson Air Force Base, Ohio, 1965)* (pp. 57–71).

Deng, S.-W., & Han, J.-Q. (2016). Towards heart sound classification without segmentation via autocorrelation feature and diffusion maps. *Future Generation Computer Systems*, *60*, 13–21. http://dx.doi.org/10.1016/j.future.2016.01.010.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923. http://dx.doi.org/10.1162/089976698300017197.

Dong, F., Qian, K., Zhao, R., Baird, A., Li, X., Dai, Z., Dong, B., Metze, F., Yamamoto, Y., & Schuller, B. W. (2020). Machine listening for heart status monitoring: Introducing and benchmarking HSS–the heart sounds shenzhen corpus. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 2082–2092. http://dx.doi.org/10.1109/JBHI.2019.2955281.

Dwivedi, A. K., Imtiaz, S. A., & Rodriguez-Villegas, E. (2018). Algorithms for automatic analysis and classification of heart sounds–a systematic review. *IEEE Access*, *7*, 8316–8345. http://dx.doi.org/10.1109/ACCESS.2018.2889437.

Fernando, T., Ghaemmaghami, H., Denman, S., Sridharan, S., Hussain, N., & Fookes, C. (2020). Heart sound segmentation using bidirectional LSTMs with attention. *IEEE Journal of Biomedical and Health Informatics*, *24*(6), 1601–1609. http://dx.doi.org/10.1109/JBHI.2019.2949516.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Proc. NIPS* (pp. 2672–2680). Montreal, Canada.

Gosztolya, G., Grósz, T., & Tóth, L. (2018). General utterance-level feature extraction for classifying crying sounds, atypical & self-assessed affect and heart beats. In *Proc INTERSPEECH* (pp. 531–535). Hyderabad, India.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116. http://dx.doi.org/10.1142/S0218488598000094.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable ai systems for the medical domain? (p. 28). ArXiv, https://arxiv.org/abs/1712.09923v1.

Hu, J., Cui, X., Gong, Y., Xu, X., Gao, B., Wen, T., Lu, T. J., & Xu, F. (2016). Portable microfluidic and smartphone-based devices for monitoring of cardiovascular diseases at the point of care. *Biotechnology Advances*, *34*(3), 305–320. http://dx.doi.org/10.1016/j.biotechadv.2016.02.008.

Humayun, A., Khan, M., Ghaffarzadegan, S., Feng, Z., & Hasan, T. (2018). An ensemble of transfer, semi-supervised and supervised learning methods for pathological heart sound classification. In *Proc INTERSPEECH* (pp. 127–131). Hyderabad, India.

Ide, H., & Kurita, T. (2017). Improvement of learning for CNN with ReLU activation by sparse regularization. In *Proc. IJCNN* (pp. 2684–2691). Anchorage, AK.

Kobayashi, T. (2019). Global feature guided local pooling. In *Proc. ICCV* (pp. 3365–3374). Seoul, Korea.

Mangione, S. (2001). Cardiac auscultatory skills of physicians-in-training: A comparison of three english-speaking countries. *The American Journal of Medicine*, *110*(3), 210–216. http://dx.doi.org/10.1016/S0002-9343(00)00673-2.

Patidar, S., Pachori, R. B., & Garg, N. (2015). Automatic diagnosis of septal defects based on tunable-Q wavelet transform of cardiac sound signals. *Expert Systems with Applications*, *42*(7), 3315–3326. http://dx.doi.org/10.1016/j.eswa.2014.11.046.

Phankokkruad, M., & Wacharawichanant, S. (2019). A comparison of efficiency improvement for long short-term memory model using convolutional operations and convolutional neural network. In *Proc. ICOIACT* (pp. 608–613). Yogyakarta, Indonesia.

Qian, K., Janott, C., Schmitt, M., Zhang, Z., Heiser, C., Hemmert, W., Yamamoto, Y., & Schuller, B. W. (2020). Can machine learning assist locating the excitation of snore sound? A review. *IEEE Journal of Biomedical and Health Informatics*, *25*, 1233–1246. http://dx.doi.org/10.1109/JBHI.2020.3012666.

Qian, K., Li, X., Li, H., Li, S., Li, W., Ning, Z., Yu, S., Hou, L., Tang, G., Lu, J., Li, F., Duan, S., Du, C., Cheng, Y., Wang, Y., Gan, L., Yamamoto, Y., & Schuller, B. W. (2020). Computer audition for healthcare: Opportunities and challenges. *Frontiers in Digital Health*, *2*, 1–4. http://dx.doi.org/10.3389/fdgth.2020.00005.

Qian, K., Ren, Z., Dong, F., Lai, W.-H., Schuller, B. W., & Yamamoto, Y. (2019). Deep wavelets for heart sound classification. In *Proc. ISPACS* (pp. 1–2). Taipei, Taiwan, China.

Qian, K., Ren, Z., Pandit, V., Yang, Z., Zhang, Z., & Schuller, B. (2017). Wavelets revisited for the classification of acoustic scenes. In *Proc. DCASE Workshop* (pp. 108–112). Munich, Germany.

Ren, Z., Cummins, N., Pandit, V., Han, J., Qian, K., & Schuller, B. W. (2018). Learning image-based representations for heart sound classification. In *Proc. DH* (pp. 143–147). Lyon, France.

Ren, Z., Kong, Q., Han, J., Plumbley, M., & Schuller, B. (2019). Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes. In *Proc. ICASSP* (pp. 56–60). Brighton, UK.

Ren, Z., Kong, Q., Qian, K., Plumbley, M., & Schuller, B. (2018). Attention-based convolutional neural networks for acoustic scene classification. In *Proc. DCASE* (pp. 39–43). Surrey, UK.

Ren, Z., Qian, K., Zhang, Z., Pandit, V., Baird, A., & Schuller, B. (2018). Deep scalogram representations for acoustic scene classification. *IEEE/CAA Journal of Automatica Sinica*, *5*(3), 662–669.

Ryu, H., Park, J., & Shin, H. (2016). Classification of heart sound recordings using convolution neural network. In *Proc. CinC* (pp. 1153–1156). Vancouver, Canada.

Schuller, B. W., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 emotion challenge. In *Proc. INTERSPEECH* (pp. 312–315). Brighton, UK.

Schuller, B., Steidl, S., Batliner, A., Marschik, P. B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F., Rathner, E.-M., Bartl-Pokorny, K. D., Einspieler, C., Zhang, D., Baird, A., Amiriparian, S., Qian, K., Ren, Z., Schmitt, M., Tzirakis, P., & Zafeiriou, S. (2018). The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. In *Proc INTERSPEECH* (pp. 122–126). Hyderabad, India.

Schwamm, L. H., Chumbler, N., Brown, E., Fonarow, G. C., Berube, D., Nystrom, K., Suter, R., Zavala, M., Polsky, D., Radhakrishnan, K., Lacktman, N., Horton, K., Malcarney, M.-B., Halamka, J., & Tiner, A. C. (2017). Recommendations for the implementation of telehealth in cardiovascular and stroke care: A policy statement from the American heart association. *Circulation*, *135*(7), e24–e44. http://dx.doi.org/10.1161/CIR.0000000000000475.

Tschannen, M., Kramer, T., Marti, G., Heinzmann, M., & Wiatowski, T. (2016). Heart sound classification using deep structured features. In *Proc. CinC* (pp. 565–568). Vancouver, Canada.

Uğuz, H. (2012). Adaptive neuro-fuzzy inference system for diagnosis of the heart valve diseases using wavelet transform with entropy. *Neural Computing and Applications*, *21*(7), 1617–1628. http://dx.doi.org/10.1007/s00521-011-0610-x.

Wang, P., Lim, C. S., Chauhan, S., Foo, J. Y. A., & Anantharaman, V. (2007). Phonocardiographic signal analysis method using a modified hidden Markov model. *Annals of Biomedical Engineering*, *35*(3), 367–374. http://dx.doi.org/10.1007/s10439-006-9232-3.

Wolrd Health Organisation (WHO) (2017). Cardiovascular diseases (CVDs) key facts. URL: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

Xu, Y., Kong, Q., Huang, Q., Wang, W., & Plumbley, M. D. (2017). Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging. In *Proc. INTERSPEECH* (pp. 3083–3087). Stockholm, Sweden.

Yu, S., Cheng, Y., Xie, L., Luo, Z., Huang, M., & Li, S. (2017). A novel recurrent hybrid network for feature fusion in action recognition. *Journal of Visual Communication and Image Representation*, *49*, 192–203. http://dx.doi.org/10.1016/j.jvcir.2017.09.007.

Zhang, Z., & Schuller, B. (2012). Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition. In *Proc. INTERSPEECH* (pp. 362–365). Portland, OR.

Zheng, Y., Guo, X., & Ding, X. (2015). A novel hybrid energy fraction and entropy-based approach for systolic heart murmurs identification. *Expert Systems with Applications*, *42*(5), 2710–2721. http://dx.doi.org/10.1016/j.eswa.2014.10.051.