Machine Listening for Heart Status Monitoring: Introducing and Benchmarking HSS—The Heart Sounds Shenzhen Corpus

Fengquan Dong, Kun Qian, *Member, IEEE*, Zhao Ren, *Student Member, IEEE*, Alice Baird, Xinjian Li, Zhenyu Dai, Bo Dong, Florian Metze, *Senior Member, IEEE*, Yoshiharu Yamamoto, *Member, IEEE*, and Björn W. Schuller, *Fellow, IEEE*

***Abstract*—Auscultation of the heart is a widely studied technique, which requires precise hearing from practitioners as a means of distinguishing subtle differences in heart-beat rhythm. This technique is popular due to its non-invasive nature, and can be an early diagnosis aid for a range of cardiac conditions. Machine listening approaches can support this process, monitoring continuously and allowing for a representation of both mild and chronic heart conditions. Despite this potential, relevant databases and benchmark studies are scarce. In this paper, we introduce our publicly accessible database, the Heart Sounds Shenzhen Corpus (HSS), which was first released during the recent INTERSPEECH 2018 ComParE Heart Sound sub-challenge. Additionally, we provide a survey of machine learning work in the area of heart sound recognition, as well as a benchmark for HSS utilising standard acoustic features and machine learning models. At best our support vector machine with Log Mel features achieves 49.7% unweighted average recall on a three category task (normal, mild, moderate/severe).**

***Index Terms*—Heart sound, cardiovascular disease, machine listening, artificial intelligence, healthcare.**

I. Introduction

THE heart sound, recorded via the Phonocardiogram (PCG), has been widely used in medical practice for its simplistic, efficient, and cost-effective screening of a number of cardiovascular diseases (CVD), which annually result in 45% of all deaths in Europe [1]. However, extensive training and experience for auscultation are needed for physicians [2]. Furthermore, it was reported that, on average just 20% of less experienced medical interns can make efficient use of the auscultation method to measure the heart conditions [3]. In the past two decades, building an intelligent machine to monitor the status of the heart via the information extracted from the PCG, has become increasingly popular within audio signal processing, and machine learning [4]. In the era of artificial intelligence (AI), and internet of things (IoT), developing an intelligent machine listening based system, can be beneficial for cardiology physicians and ultimately patients suffering from CVD to better understand their current health status.

It is encouraging to see that, the variety of approaches published in the literature demonstrate the feasibility to automatic diagnosis of CVD via machine learning and signal processing techniques. However, the limitations highlighted in the existing studies are: First, publicly accessible heart sound databases are extremely limited [5], which dramatically limits further and reproducible research. Table I shows the known publicly accessible heart sound databases. Currently, the PhysioNet CinC Challenge database [7] is the largest, containing different PCG signals collected from eight different medical centres. However, the data acquisition system, environment, and the annotation procedures are not consistent. This may results in obstacles and uncertainties for building an intelligent model. Secondly, most of the previous studies ignore subject-independency, which might

This work was supported in part by the Natural Science Foundation of Shenzhen University General Hospital under Grant SUGH2018QD013, China, in part by the Zhejiang Lab's International Talent Fund for Young Professionals (Project HANAMI), China, in part by the JSPS Postdoctoral Fellowship for Research in Japan (ID No. P19081) from the Japan Society for the Promotion of Science (JSPS), Japan, in part by the Grants-in-Aid for Scientific Research (No. 19F19081 and No. 17H00878) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and in part by Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network (MSCA-ITN-ETN) Project under grant agreement No. 766287 (TAPAS), as well as the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B). *Fengquan Dong and Kun Qian contributed equally to this work. (Corresponding authors: Kun Qian; Zhao Ren.)*

F. Dong and B. Dong are with the Department of Cardiology, Shenzhen University General Hospital, Shenzhen 518055, China (e-mail: fengquan.dong@foxmail.com; yinghan.dong@163.com).

K. Qian and Y. Yamamoto are with the Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo 113-0033 Japan (e-mail: qian@p.u-tokyo.ac.jp; yamamoto@p.u-tokyo.ac.jp).

Z. Ren and A. Baird are with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159 Germany (e-mail: zhao.ren@informatik.uni-augsburg.de; alice.baird@informatik.uni-augsburg.de).

X. Li and F. Metze are with the Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: xinjianl@cs.cmu.edu; fmetze@cs.cmu.edu).

Z. Dai is with the Department of Cardiovascular Medicine of the First Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, China (e-mail: zhenyudai@foxmail.com).

B. W. Schuller is with the GLAM—Group on Language, Audio and Music, Imperial College London, SW7 2AZ London, U.K., and also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany (e-mail: schuller@ieee.org).

TABLE I
A COMPARISON WITH OTHER PUBLICLY ACCESSIBLE HEART SOUND DATABASES

	Year	Number of Instances	Heart Sound Categories
PASCAL Heart Sound Challenge Dataset [6]	2011	Dataset A: 176 Dataset B: 656	Dataset A: Normal, Murmurs, Extra Heart Sound, and Artifact Dataset B: Normal, Murmurs, and Extra Systoles
PhysioNet CinC Challenge Dataset [7]	2016	3 240	Normal, Abnormal, and Uncertain
HSS (Proposed)	2018	845	Normal, Mild, and Moderate/Severe

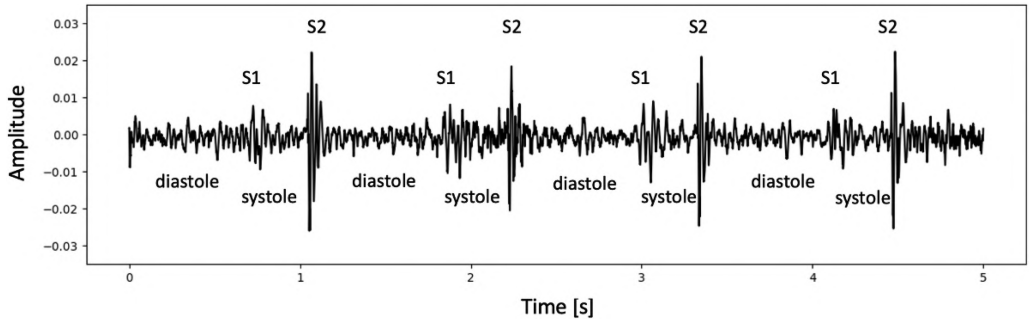


Fig. 1. The PCG sample of one heart sound audio file in HSS corpus.

make the final results overoptimistic. Thirdly, the experimental paradigm, e.g., the selected feature sets, the selected classification algorithms, the validation method, and the evaluation metrics were not standardised, which makes the related work difficult to reproduce or compare.

In order to overcome the aforementioned challenges, we introduce a conventional and reproducible benchmark for our publicly accessible heart sound database. The main contributions of this work are: 1) We release a standard heart sound database with consistent data collection equipment, rigid ground-truth examination, and a reasonable data partitioning principle. We use the name of **Heart Sounds Shenzhen (HSS)** corpus, as it first appeared at the INTERSPEECH 2018 COMPARE challenge Heart Sound sub-challenge [8]. To our best knowledge, HSS is the most recent and largest expert-annotated heart sound database collected from one single medical centre. 2) We use standard open source toolkits, e.g., OPENSMILE, LIBSVM, and TensorFlow, to benchmark the baseline of the introduced database, which will make the study reproducible and sustainable. 3) We provide a comprehensive investigation and comparison on typical acoustic features and classifiers, which will give a guidance to colleagues in the same community to consider future research directions.

The remainder of this paper will be organised as follows: At first, the background and related work will be described in Section II. Subsequently, the database, and the methods used for the benchmark work will be introduced in Section III. We will give experimental results and a discussion in Section IV and Section V, respectively. Finally, an overall conclusion is given in Section VI.

II. BACKGROUND & RELATED WORK

A plethora of research has been invested in the field to observe efficient and robust machine listening based systems for analysing heart sounds. Among these studies, there are two main research directions: segmentation, and classification. The segmentation of heart sounds is to separate the PCG signals into their fundamental components, i.e., the first (S1), and the second (S2) heart sounds (see Fig. 1). The S1 is caused by the closure of the mitral and the tricuspid valves, while S2 is caused by the closure of the aortic and the pulmonary valves. S1 and S2 are normal sounds. Nevertheless, the third, and the fourth heart sounds, i.e., S3 and S4, murmurs, and ejection clicks, usually refer to some disease, or anomaly [4]. As indicated by Renna *et al.* [9], segmentation can contribute to the feature extraction for an individual component, the detection of extra sound components, and the extraction of information from the analysis of waveforms associated to the S1 and S2 sounds. Therefore, in many studies on machine listening for heart sound monitoring, segmentation is the first step of a classification system. Nigam and Priemer proposed a complexity-based algorithm for segmentation of the PCG signals [10]. They indicated that, their method is invariant to amplitude and frequency variations of the heart sound, which can achieve improved time gates for heart sounds as compared to energy-based segmentation. When analysing the PCG, some other studies made a combination of the PCG and the electrocardiogram (ECG) signals. Syed *et al.* proposed a framework to perform analysis of acoustical cardiac signals via the PCG and the simultaneously recorded ECG signals [11]. In their work, the ECG signals were used to locate the R wave, and the T wave for each heart beat. In addition, a K-means algorithm was applied to select the interesting candidate intervals for further analysis, e.g., those which may contain pathological heart sounds. In the study of [12], three representations, i.e., the normalised Shannon energy, the envelope information of the Hilbert transformation, and the cardiac sound characteristic waveform (CSCW), were compared. From the results, the CSCW representation showed superior performance to the other two approaches. Furthermore, the sequential characteristic of the PCG was investigated. Schmidt *et al.* proposed a duration-dependent hidden Markov model (DHMM) for segmentation of heart sounds [13]. In their study, the model was based on the duration of the events,

the amplitude of the signal envelope, and a predefined model structure. A dynamic clustering algorithm for segmentation of heart sounds was proposed in [14], in which the heart sounds were segmented based on the hybrid analysis of clustering and medical knowledge. The S-transformation was applied to segment the heart sounds into S1, S2, and diastole in [15], and evaluated against white additive Gaussian noise. Varghees and Ramachandran proposed a method for automated robust heart sound activity detection, which could be applied to the real-time wireless cardiac health monitoring, and electronic stethoscope devices [16]. The ensemble empirical mode decomposition (EEMD) combined with kurtosis features were used in [17], to separate the heart cycle into four components, i.e., diastole, S1, systole, and S2. A hidden semi-Markov model (HSMM), extended with the use of logistic regression for emission probability estimation, was introduced in [18]. In addition, a modified Viterbi algorithm was involved for decoding the most likely sequence of states. Chen *et al.* introduced a deep neural network (DNN) method for recognising S1 and S2 without any duration and interval information [19]. Ozbek and Shamsi introduced a new entropy bound with low computational complexity for differential Shannon entropy estimation with a kernel density approach in [20], by which a bound for the Kullback-Leibler divergence between two Gaussian mixture models was defined. A deep recurrent neural network (DRNN) was used in [21], fed with spectral and envelope features, to detect the state sequence. A multilayer perceptron (MLP) neural network using Cochleagram features for S1-S2 identification was proposed in [22], which was demonstrated to be superior than other acoustic features reported earlier. A recent study [9] showed that, a deep convolutional neural network (DCNN) can achieve a promising performance when combined with HMM and HSMM.

For representation, wavelet transformation has been found to be efficient for the recognition of heart sounds when utilising conventional classifiers, e.g., MLP [23], or *k*-nearest neighbour (*k*-NN) [24]. Ahlstrom *et al.* made an investigation of the features which could be extracted as a means for distinguishing a pathological murmur from a physiological murmur [25]. Pudil's sequential floating forward selection (SFFS) method was used for selecting the subset of multi-domain features including Shannon energy, wavelets, fractal dimensions, and recurrence quantification analysis. Wang *et al.* indicated that, a Mel-frequency cepstral coefficients (MFCCs) based hidden Markov model (HMM) system may have promising performance on heart sound classification, when comparing with other features extracted from the time domain, and short-time Fourier transformation (STFT) [26]. Maglogiannis *et al.* found that, a support vector machine (SVM) based model can perform well in recognition of aortic stenosis (AS), aortic regurgitation (AR), mitral stenosis (MS), and mitral regurgitation (MR) heart sounds [27]. An effective and simple method for PCG classification was proposed in [28], by which the time-frequency representation and feature reduction were investigated. Ari *et al.* studied the capacity of wavelet features and least square SVM (LSSVM), which can be effective to classify the cases of the normal, the aortic insufficiency, the aortic stenosis, the atrial septal defect, the mitral regurgitation, and the mitral stenosis [29]. The discrete Fourier transformation (DFT) and Burg autoregressive (AR) were involved in feature extraction for classifying the normal, the pulmonary and the mitral stenosis heart valve diseases via a NN classifier [30]. Furthermore, the same author in the aforementioned work, also found the entropy features of sub-bands by discrete wavelet transformation (DWT). This can be used to classify the heart sounds via adaptive neuro-fuzzy inference system (ANFIS) classifiers [31]. Schmidt *et al.* studied multiple features for heart sound classification via a quadratic discriminant function [32]. Patidar *et al.* introduced tunable-Q wavelet transformation (TQWT) based features for heart sound classification in [33], by which the sub-band information extracted can characterise the various types of murmurs in cardiac sound signals. The wavelet packet transformation (WPT) was involved in [34], for detecting the abnormality of heart sounds and discriminating heart murmurs by a SVM classifier. Gharehbaghi *et al.* indicated that, the growing time SVM (GTSVM) can be superior to the conventional SVM in recognition of innocent and pathological murmurs [35]. Guillermo *et al.* proposed a radial wavelet neural-network (RWNN) classifier for heart murmurs detection [36], by which the performances achieved by MLP and extreme learning machine (ELM) were also investigated and compared. Deng and Han introudced a novel framework for heart sound classification without a segmentation step [37], by which the autocorrelation features were extracted from the sub-band envelopes calculated from the sub-band coefficients of the heart sound signal by DWT. The DCNN fed with the mel-frequency spectral coefficients was applied to normal/abnormal heart sound classification in [38]. In their study, a publicly accessible database, i.e., the PhysioNet/Computing in Cardiology (CinC) Challenge database [7] was used. Other techniques like wavelet analysis [39], probability assessment [40], and sparse coding [41], were investigated on the same task as previously mentioned.

Among other studies in recent years, innovative directions were given in two points: First, finding novel features, e.g., empirical wavelet transformation [42], tensor decomposition from the scaled spectrogram [43], curve fitting and fractal features [44], transfer learning based representations [45], and deep unsupervised learned representations [46]. Second, improving the decision making process, e.g., modified neighbor annealing (MNA) [47], and classification tree [48]. One common trend from these recent studies is to eliminate the reference of ECG signals for segmentation of the heart sounds, which is of great importance for a holistic automated system for diagnosis of heart diseases via the PCG signals.

In this work, we focus on introducing our large scale publicly accessible heart sound database. We use standard acoustic features and machine learning models to benchmark the baseline of the database, which are reproducible and comparable. Additionally, all the toolkits used in this work can be freely downloaded and used for research purposes.

III. Materials and Methods

In this section, we will firstly introduce the proposed publicly accessible heart sound database, i.e., HSS. Then, the acoustic features and machine learning models used for the benchmark

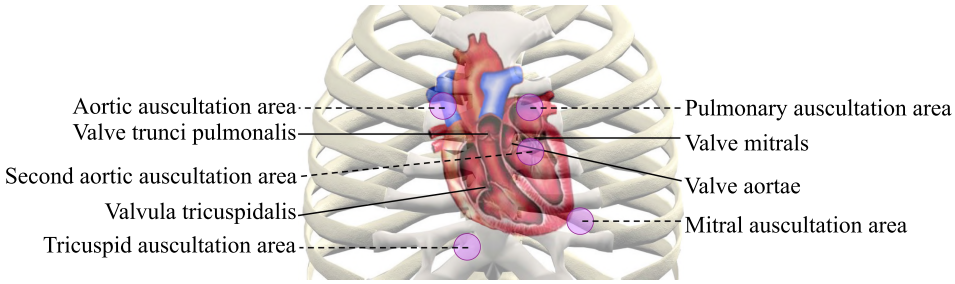
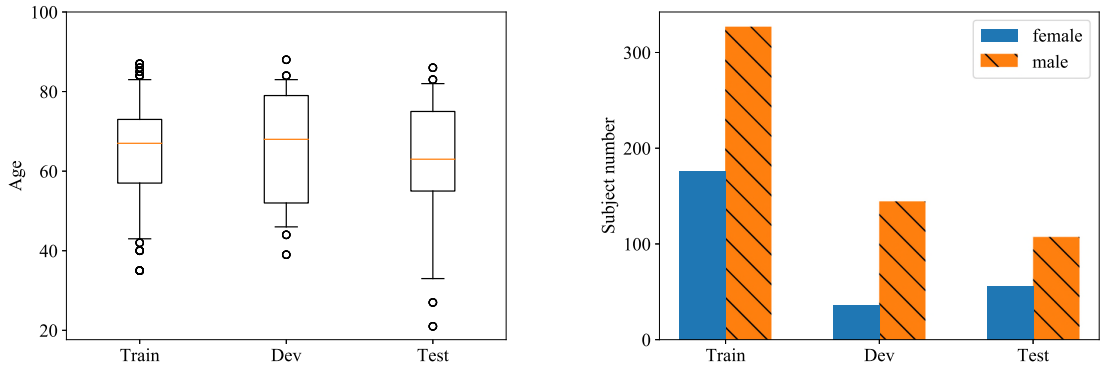


Fig. 2. The anatomical positions of the heart. We collected the heart sounds from four locations of one subject's body: auscultatory mitral, aortic valve auscultation, pulmonary valve auscultation, and auscultatory areas of the tricuspid valve.



(a) Age Distribution (b) Gender Distribution

Fig. 3. The demographics of the data partition. There are no significant differences between each set in terms of age or gender in HSS corpus.

work will be given. Finally, we give the evaluation metrics adopted in this study.

A. HSS Database

This study was approved by the ethic committee of the Shenzhen University General Hospital. There were 170 subjects involved (female: 55, male:115) with various health conditions including coronary heart disease, heart failure, arrhythmia, hypertension, hyperthyroid, valvular heart disease and congenital heart disease amongst others. The ages of participants ranged from 21 to 88 years (65.4 ± 13.2 years). The heart sound audio recording was collected with an electronic stethoscope (Eko CORE, USA) set up via Bluetooth 4.0, at a 4 kHz sampling rate. The data was acquired from four locations on the body (cf. Fig. 2), i.e., auscultatory mitral, aortic valve auscultation, pulmonary valve auscultation, and auscultatory areas of the tricuspid valve. For each area as mentioned previously, a duration of 30 seconds on average (ranging from 29.808 s to 30.152 s) in a sitting or supine position of the subjects was recorded, which resulted in 845 recordings within 422.82 min length from the 170 subjects.

Experienced cardiologists annotated the data through use of the golden standard, i.e., Echocardiography. In Echocardiography, the mitral and tricuspid valves use area ratios to predict the reflux: mild (less than 30%), moderate (30%–50%), severe (greater than 50%). Correspondingly, the HSS has three category heart sounds to be classified: normal, mild, and moderate/severe. Due to the diverse nature of human conditions such as these, we consider a subject independent approach, and split the data into train, development (dev), and test sets. The distribution of gender, class, and age are taken into account when partitioning the data (see Fig. 3), ending up with 502/180/163 instances for the train/development/test sets collected from 100/35/35 subjects. The detailed information of data split information can be found in Table II.

TABLE II
THE NUMBER [#] OF INSTANCES IN EACH DATA SET OF THE HSS CORPUS. FEMALE SUBJECTS: 55, MALE SUBJECTS:115. AGE: 65.4 ± 13.2 YEARS

#	Train	Dev	Test	Σ
Normal	84	32	28	144
Mild	276	98	91	465
Moderate/Severe	142	50	44	236
Σ	502	180	163	845

B. Acoustic Features

Extracting efficient representations from the analysed signal is a vital step in the paradigm of machine learning. In this study, we use standard acoustic features extracted by the open source toolkit OPENSMILE [49], [50], which was also used to train the baseline system in the INTERSPEECH 2018 COMPARE challenge Heart Sound sub-challenge [8].

TABLE III
THE LLDS FOR COMPARE FEATURE SET. RASTA: RELATIVE SPECTRAL TRANSFORM; HNR: HARMONICS TO NOISE RATIO; RMSE: ROOT MEAN SQUARE ENERGY

55 spectral LLDs	**Group**
MFCCs 1–14	Cepstral
Psychoacoustic sharpness, harmonicity	Spectral
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	Spectral
Spectral energy 250–650 Hz, 1 k–4 kHz	Spectral
Spectral flux, centroid, entropy, slope	Spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	Spectral
Spectral variance, skewness, kurtosis	Spectral
6 voicing related LLDs	**Group**
F_0 (SHS and Viterbi smoothing)	Prosodic
Prob. of voicing	Voice Quality
log HNR, jitter (local and δ), shimmer (local)	Voice Quality
4 energy related LLDs	**Group**
RMSE, zero-crossing rate	Prosodic
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic

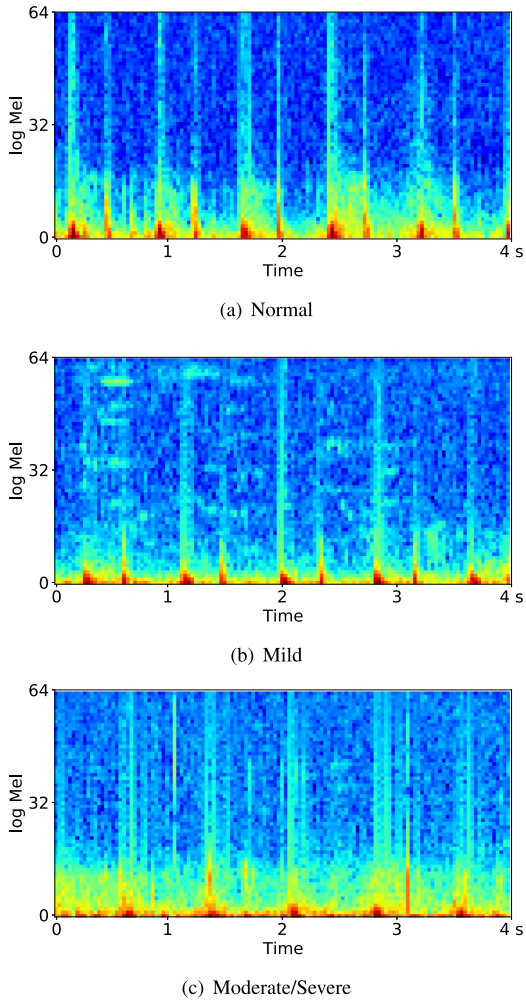
1) Low-Level Descriptors: The low-level descriptors (LLDs) are extracted from the frame-level analysed signals, which can be used to represent the physiological characteristics of the heart sound. In this study, we use the sophisticated COMPARE feature set, which includes 65 basic acoustic features (see Table III). The configuration file known as ComParE_2016, included in the 2.3 version of OPENSMILE is used. The detailed information about these LLDs can be referred to [51]. In addition, motivated by the success achieved in our previous studies [45], we also extract a Log Mel feature set. Log Mel spectrograms of typical heart sounds can be seen in Fig. 4.

2) Functionals: We can assume that, the change of the LLDs over a given period of time can carry important information for building a model to learn the inherited characteristics of the analysed audio signals [51]. For heart sound, we apply *supera-segmental* features, i.e., we apply *functionals* [51] to the aforementioned LLDs extracted from the time signal. The mechanism of functionals is to map the time series based LLDs to a scalar value per applied functional (e.g., minimum, maximum, mean), which results in a single, fixed dimension vector independent of the length of the input heart sound clip. Details on the used functionals from the COMPARE feature set can be found in [51].

C. Machine Learning Models

In this study, we investigate and compare two typical kinds of machine learning models, i.e., Support Vector Machine (SVM) and Recurrent Neural Network (RNN), which represents the static and the dynamic models, respectively.

1) Support Vector Machine: The mechanism of an SVM [52] is to find a set of hyperplanes in a multi-dimensional space that instances of different classes can be separated by (see Fig. 5). For a classification task, a subset of data points from the training set, called as *support vectors*, which have the widest possible gap, will be selected as pivots to support the hyperplane on both sides of the *margin*. Then, the instances from the test set will be firstly



(a) Normal

(b) Mild

(c) Moderate/Severe

Fig. 4. The Log Mel spectrogram images of the three heart beat classes. The three images are extracted from the first 4 s audio clip of the audio files.

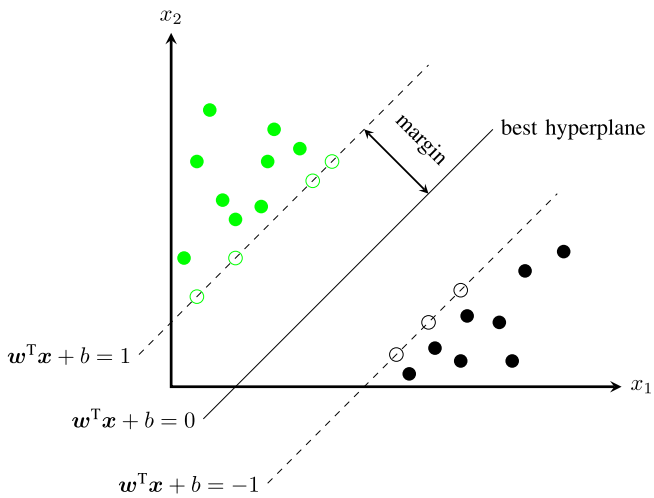


Fig. 5. The mechanism of training a SVM classifier in a binary classification scenario. The best hyperplane can be achieved by maximising the margin during the process of training a SVM classifier. The support vectors (indicated by circles) are the data points with the widest possible gap; $\boldsymbol{w}$ is a normal vector; b is a bias.

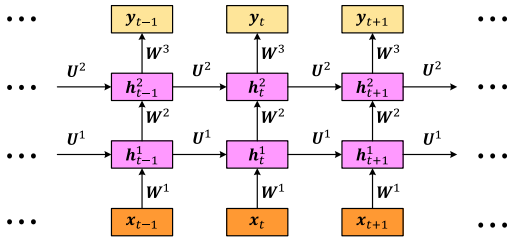


Fig. 6. The structure of a RNN model. $\boldsymbol{W}$, and $\boldsymbol{U}$ represent the neuron weight matrix, and the recurrent weight matrix, respectively. $\boldsymbol{h}$ represents a hidden layer.

mapped to this multi-dimensional space, and the predictions are given based on which side of the gap they will fall onto.

In a binary classification scenario, an SVM aims to find the optimal margin separating hyperplane by solving the following optimisation problem [52]:

$$\begin{aligned} &\text{minimise}: \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) - \sum_{i=1}^{n}\alpha_i, \\ &\text{subject to}: \sum_{i=1}^{n}\alpha_i y_i = 0,\ 0 \le \alpha_i \le C,\ i = 1,\ldots,n, \end{aligned} \tag{1}$$

where α_i is called the *Lagrange* multiplier of a training sample $(\boldsymbol{x}_i, y_i)$, and C is a hyper-parameter to control the generalisation of the trained SVM model. To make the SVM model available to analyse linear or nonlinear separable problems, a *kernel function*, i.e., $\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ is used. There are some commonly used kernel functions including *linear*, *polynomial*, and *radial basis function* (RBF). The optimisation problem mentioned above can be solved by the *sequential minimal optimisation* (SMO) algorithm [53].

It should be noted that, SVMs were originally designed for binary classification problems, however there are some methods developed to make it feasible for SVMs to solve multi-class problems. Among these methods, there are two popular ones, i.e., *one-versus-all*, and *one-versus-one* [54]. In this study, the *one-versus-one* is used, by which the final prediction of a test sample will be given by the most frequently voted classifier among the ones trained in pairs by a binary SVM classifier at a first step.

2) Recurrent Neural Network: In the past few years, *deep learning* [55] has dominated the machine learning community, achieving high performance in many fields including speech recognition [56], image recognition [57], or object detection [58], with many advantages shown, particularly for speech in the health domain [59]. In this study, we investigate the RNN [60], for its performance in recognising heart sounds. Unlike the normal feed-forward neural network (FNN), an RNN can learn the contextual information from the sequential input (see Fig. 6), taking into account the inherent time dependencies of the heart beat signal.

Nevertheless, when training a RNN model, inevitably problems occur, such as *vanishing gradient* [61] in the *backpropagation through time* (BPTT) [62], which can restrain the RNN in learning long-term contextual information. To overcome the issue of vanishing gradient, more complicated topologies, such as having a *memory cell* to preserve long-term information are needed. Two popular structures can achieve this aim, i.e., *long short-term memory* (LSTM) cells [63] and *gated recurrent unit* (GRU) cells [64].

In an LSTM cell (see Fig. 7(a)), the inputs include the output from the layer below (here for simplicity, we use $\boldsymbol{x}_t$), and the output from the current layer at a previous time step ($\boldsymbol{h}_{t-1}$). A candidate value $\tilde{\boldsymbol{c}}_t$ is generated by the aforementioned inputs. There are three gates in an LSTM cell, i.e., *input gate* ($\boldsymbol{i}$), *output gate* ($\boldsymbol{o}$), and *forget gate* ($\boldsymbol{f}$). The mechanism of an LSTM cell can be described as:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i\boldsymbol{x}_t + \boldsymbol{U}_i\boldsymbol{h}_{t-1} + \boldsymbol{b}_i), \tag{2a}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o\boldsymbol{x}_t + \boldsymbol{U}_o\boldsymbol{h}_{t-1} + \boldsymbol{b}_o), \tag{2b}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f\boldsymbol{x}_t + \boldsymbol{U}_f\boldsymbol{h}_{t-1} + \boldsymbol{b}_f), \tag{2c}$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c\boldsymbol{x}_t + \boldsymbol{U}_c\boldsymbol{h}_{t-1} + \boldsymbol{b}_c), \tag{2d}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t, \tag{2e}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \sigma_h(\boldsymbol{c}_t), \tag{2f}$$

where σ is the logistic sigmoid function, and $\odot$ denotes the element-wise multiplication. In addition, $\boldsymbol{c}$, and $\boldsymbol{b}$ represent the *cell state*, and the bias matrix, respectively. We can see that, if the *forget gate* is open (gate activation values are close to one) and the *input gate* is closed (gate activation values are close to zero), the activation of the cell cannot be overwritten by the new inputs. At this point, the information from the previous time steps can then be accessed by opening the *output gate*.

Compared with LSTMs, GRUs have a simpler structure (see Fig. 7(c)): an *update gate* $\boldsymbol{z}$, a *reset gate* $\boldsymbol{r}$, an activation $\boldsymbol{h}$, and a candidate activation $\tilde{\boldsymbol{h}}$. The mechanism of a GRU cell can be defined as:

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r\boldsymbol{x}_t + \boldsymbol{U}_r\boldsymbol{h}_{t-1} + \boldsymbol{b}_r), \tag{3a}$$

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z\boldsymbol{x}_t + \boldsymbol{U}_z\boldsymbol{h}_{t-1} + \boldsymbol{b}_z), \tag{3b}$$

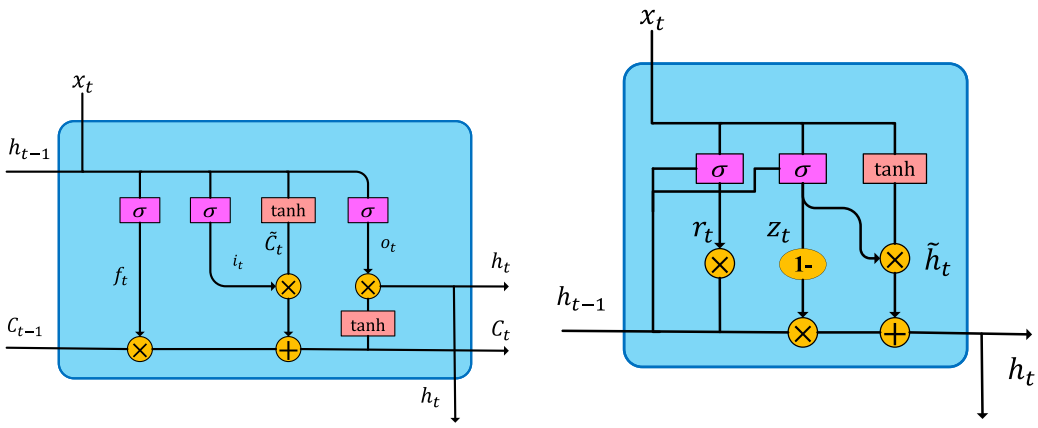
$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_h\boldsymbol{x}_t + \boldsymbol{U}_h(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{b}_h), \tag{3c}$$

$$\boldsymbol{h}_t = \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t + (\boldsymbol{1} - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1}. \tag{3d}$$

The unit will be overwritten if the *update gate* is closed, which helps the model to remember the existing contextual information from inputs for a long series of time steps. In this way, the error can be back-propagated without too much attenuation by passing through the *update gate* when it is open [64]. For simplicity, we use LSTM, and GRU to represent the implemented LSTM RNN, and GRU RNN, respectively in the proceedings sections.

D. Fusion Strategy

There are two main fusion strategies in this work, i.e., early fusion, and late fusion. In the scenario of early fusion, features will be concatenated directly before being fed into the machine learning model. On the other hand, late fusion is implemented by independently training the models with different feature sets, and the final prediction will be made by using a voting method. In this study, we investigate two late fusion strategies (refer to [65]), i.e., the *majority voting* (MV), and the *margin sampling voting* (MSV). For MV, the final prediction will be given to the one



(a) An LSTM Cell (b) A GRU Cell

Fig. 7. The structures of the LSTM and the GRU cells.

TABLE IV
THE UARS ([%]) ACHIEVED BY THE PROPOSED MODELS ON THE DEVELOPMENT AND THE TEST SETS. THE RESULTS SHOWN BY THE DEVELOPMENT SET ARE ACHIEVED BY THE OPTIMISED PARAMETERS. CHANCE LEVEL: 33.3%.

			SVM (UAR [%])		LSTM (UAR [%])		GRU (UAR [%])	
Feature Type	Delta	#Features	Dev	Test	Dev	Test	Dev	Test
MFCCs	coeff + delta	1400	39.4	31.9	36.6	40.8	36.6	37.8
	coeff	756	41.5	29.9	36.6	40.0	36.7	36.9
	delta	644	41.4	46.2	36.0	38.1	34.6	33.8
Log Mel	coeff + delta	800	45.6	48.3	36.6	41.6	39.0	**42.3**
	coeff	432	46.5	**49.7**	38.1	34.1	38.9	38.5
	delta	368	44.2	43.3	40.4	**43.3**	42.6	37.8
eGeMAPS	coef	88	44.3	46.0	38.0	37.6	37.1	34.4
F1–F3	coeff + delta	600	33.7	37.1	37.2	33.4	38.7	33.4
	coeff	324	38.0	40.3	36.4	34.9	38.1	34.2
	delta	276	33.7	35.7	35.6	34.1	37.4	35.8
ALL w/o F1–F3	coeff + delta	6373	50.3	46.4	41.1	39.3	39.0	34.6
	coeff	3425	45.5	46.0	39.4	40.5	40.7	37.6
	delta	2948	42.7	49.1	38.3	43.2	37.8	41.9
ALL	coeff + delta	6973	48.0	44.3	38.8	36.9	44.1	35.3
	coeff	3749	45.8	47.4	39.6	36.3	39.1	39.8
	delta	3224	45.1	46.8	40.0	41.2	40.0	38.3

which has been voted for most commonly by the individually trained classifiers. In the strategy of MSV, the final prediction will be set to the choice made by the 'most confident' individual classifier, which has the biggest *margin sampling* value [66], i.e., the difference between the first and the second highest posterior probability value estimated by the model.

E. Evaluation Metrics

As is common with health-related data, the HSS is an imbalanced database (see Table II), making it suitable to be evaluated by the unweighted average recall (UAR) [67] rather than the traditionally used weighted average recall (WAR), i.e., the accuracy. The UAR is defined as:

$$\text{UAR} = \frac{\sum_{i=1}^{N_c} \text{Recall}_i}{N_c}, \tag{4}$$

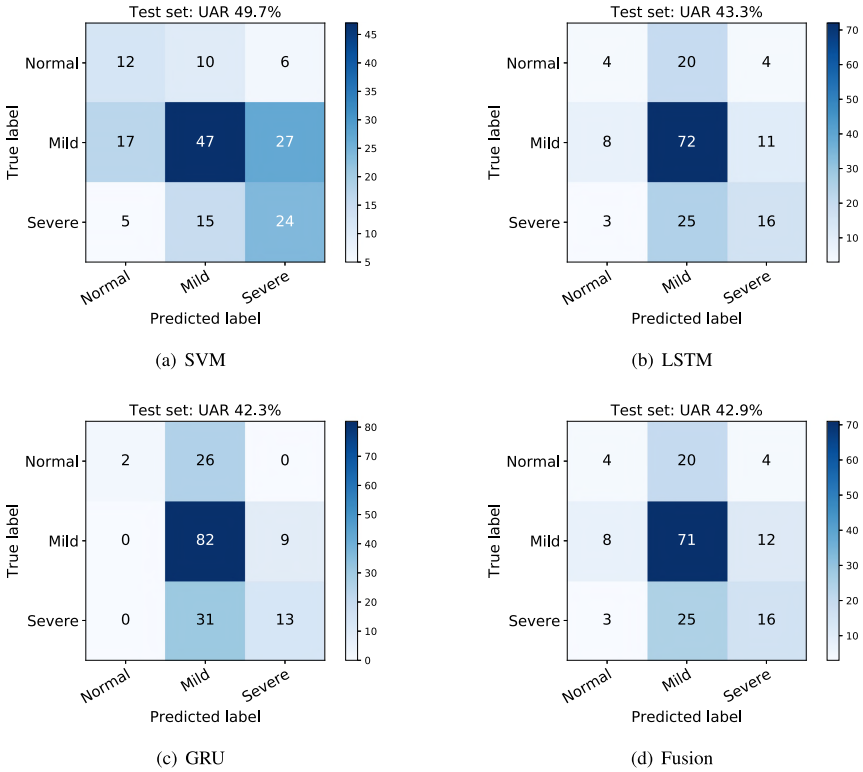
where N_c is the number of classes (here in this study, $N_c = 3$).

IV. Experimental Results

In this section, the experimental results for benchmarking the HSS database will be given. To make the relevant studies comparable and reproducible as outline, we use standard and opensource toolkits.

A. Experimental Setup

The acoustic features (both of LLDs and functionals) are extracted by our toolkit, openSMILE [50]. The SVM model is implemented by LIBSVM [68]. The RNN model is implemented based on Keras. All the hyper-parameters of the classifiers are tuned and optimised on the development set, and applied to the test set after concatenation of the train and development sets. We selected the *linear* kernel for the SVM model as it achieved excellent performance in the previous experiments [8]. The C-value of the SVM is optimised within a grid searching strategy



(a) SVM (b) LSTM

(c) GRU (d) Fusion

Fig. 8. Confusion matrices of the best models achieved on the test set.

from $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1.0\}$. For training the RNN models, we use a three-layer structure (480–120–60) for both of the LSTM and GRU cells. In addition, a highway layer (60), a fully connected layer (30), and a final linear layer with softmax activation are added to the LSTM, or GRU layers. The optimiser for RNN models is 'rmsprop'. The batch size is set as 128, and the learning rate is 0.001. During training RNNs, we train the model for 90 epochs on the training set, and compare the performance on development set at each 10 epochs. Then, the best result is chosen from the nine models, and the corresponding epoch number is applied to the model used for test set.

We need to note that, to avoid each audio instance containing a large number of sequences, we first segment the audio instances into clips with a window size of 4 000 ms and a step size of 3 200 ms (the first segment level). Then, we segment the audio clips with a window size of 800 ms and a step size of 400 ms (the second segment level) for training the sequential model, i.e., the RNN models. The final prediction will be obtained via the predictions of the first segment level using the MSV strategy.

B. Results

Table IV shows the UARs achieved by the proposed models on the development, and the test set, respectively. Generally, in this study, SVM (49.7% of UAR) shows a better performance than RNN, for both the LSTM RNN (43.3% of UAR) and the GRU RNN (42.3% of UAR). The confusion matrices of the best models achieved on the test set are illustrated in Fig. 8. We find that, the RNN models (LSTM, or GRU) outperform the SVM model on the recall of 'mild' type of heart sounds. On the other hand, the SVM model can achieve better recalls for both of the 'normal,' and the 'severe' types of heart sounds. We tried to fuse the best models (see Table V, and Fig. 8(d)) via the strategy of majority voting (MV), and margin sampling voting (MSV). However, the best result (42.9% of UAR) achieved on the test set still yields to the single SVM model (49.7% of UAR).

TABLE V
THE UARS ([%]) ACHIEVED BY LATE FUSION OF THE BEST THREE MODELS OF SVM, LSTM, AND GRU. CHANCE LEVEL: 33.3%.

	MV		MSV	
UAR[%]	Dev	Test	Dev	Test
SVM+LSTM	40.5	**45.2**	43.5	42.9
SVM+GRU	39.0	42.7	39.0	42.3
LSTM+GRU	42.1	45.1	39.0	42.3
SVM+LSTM+GRU	40.8	45.1	39.0	42.3

V. DISCUSSION

In this section, we will discuss the current benchmark, the comparison with other published work on HSS, and limitations, as a means of suggesting future directions.

A. Benchmarks of This Work

We can see that, the Log Mel feature set performs the best in this study (refer to Table IV). The SVM model trained by Log Mel benchmarks the HSS in this study within an UAR of 49.7%. The eGeMAPS feature set has the minimum dimension, which can reach a comparable performance when fed into a

TABLE VI
CURRENT STATE-OF-THE-ART WORK ON THE HSS DATABASE IN PUBLISHED LITERATURE. LDA: LINEAR DISCRIMINANT ANALYSIS; MLP: MULTILAYER PERCEPTRON

	UAR [%]	Main Methods
INTERSPEECH 2018 COMPARE Challenge Baseline [8]	56.2	Late fusion of COMPARE Features + openXBOW, and auDeep
Humayun *et al.* [69]	42.1	1D-CNN Feature Transfer Learning, COMPARE Features SVM, LDA, MLP
Gosztolya *et al.* [70]	49.3	COMPARE Features, BoAW, Frame-Level DNN Posteriors Binned Features, SVM, Late Fusion
Amiriparian *et al.* [46]	47.9	Sequence to Sequence Autoencoders, SVM, Early Fusion

SVM model (UAR at 46.0%). On the other hand, when feeding all of the features (6 373 dimensions) into the models, the performance cannot be improved. In fact, training the models within redundant features may lead to a decrease in performance. The performance of formant-based features (F1-F3) yields to the MFCCs. We can think that, formants carry important information about the structure of the sound generator. However, the complicated anatomical structure of the heart can make formants themself insufficient for revealing the representations of the pathological sounds.

When comparing the classifiers trained by the Log Mel feature set, SVM outperforms RNN (LSTM, or GRU). It is reasonable to think that, due to the limited number of instances, the capacity to learning sequential information from the heart sound is restrained. In future work, we will continuously collect the heart sound data. The fused models cannot outperform the single models in this work (see Table V). Therefore, another future direction is to find a more efficient fusion strategy to exploit the recognition capacity of each individual classifier.

B. Comparison With Other Work

We show the current state-of-the-art work on the HSS corpus in published literature in Table VI. We should note that, untill now, the INTERSPEECH 2018 COMPARE challenge baseline keeps the highest UAR for the HSS. This record was achieved by a late fusion (MV) by two optimised models, i.e., a SVM trained by COMPARE LLDs processed by the bag-of-audio-words (BoAW) approach [71], and an SVM trained by representations learnt by deep sequence to sequence autoencoders [72]. The models can be reproduced by our open source toolkits, OPENXBOW [73] and auDeep [74]. There were no winners in the INTERSPEECH 2018 COMPARE challenge heart sound sub-challenge. In addition to the official baseline paper [8], there are three published works which used our HSS database [46], [69], [70]. In the study of [69], the authors proposed a 1D-CNN time-convolution (tConv) layers based model pre-trained by the PhysioNet CinC Challenge database to learn higher representations from the heart sounds. In addition, they also investigated representation learning (RL) by sequence to sequence autoencoders. Finally, in their study, an ensemble model by hierarchically fusing the SVM models trained by a COMPARE feature set, and the model trained by tConv layers achieved the an UAR of 42.1% on test set. Gosztolya *et al.* proposed an interesting idea to extract the BoAW representations from the frame-level DNN posteriors [70]. However, as they indicated, this idea failed in the heart sound recognition task due to the insufficient information provided from the sliding windows, which could only contain one heart beat. Finally, the highest UAR achieved in their study (49.3%) was a model by late fusion strategy based on the instance-wise posterior estimates of SVM models trained on the COMPARE feature set, the BoAW representations from MFCCs, and the binned energy feature set. An unsupervised feature learning method based on sequence to sequence autoencoders contributed to the INTERSPEECH 2018 COMPARE challenge baseline, and is described in deatils in [46]. An early fusion of different models trained by clipping the amplitude below certain thresholds (−30 dB, −45 dB, −60 dB, and −75 dB) utilising the AUDEEP open source toolkit [75] reached a UAR of 47.9% in the work.

We can see from the aforementioned studies, the COMPARE feature set, and the SVM model dominated the results. These expert designed hand-crafted features, and the conventional popular classifier, are demonstrated to be robust and efficient for heart sound classification. The state-of-the-art deep learning based techniques can be promising in extracting some higher representations without any domain knowledge. However, the performance between the development set and the test set has a substantial gap [8], [46], [69]. One reasonable explanation for this phenomenon is that, due to the extremely limited data size, overfitting occurred in the development set. Positively, we are happy to see that, there are an increasing number of works using our HSS database. We continue to find novel representations, or robust models for the heart sound classification from the community which shares the common interests.

C. Current Limitations and Outlook

As previously discussed, the baseline [8] was 56.2% in INTERSPEECH 2018 COMPARE Challenge heart sound sub-challenge, and it relied on a combination of all used optimised models by using sophisticated late fusion strategies. For single models, the current benchmark is 49.7% in this work. First of all, the limited data size constrains the development of state-of-the-art deep learning techniques. On the other hand, future directions can be given by using a more advanced data augmentation method, like generative adversarial networks (GANs) [76], which recently were successfully applied to the study of snore sound recognition [77]. Secondly, unlike the typical audio signals, e.g., speech or music, the heart sound is a kind of

physiological signals. Therefore, some more advanced signal processing methods should be considered. For instance, the wavelet transformation had been found very efficient in our previous studies on snore sounds [78]–[82]. Thirdly, annotating the heart sound data is an expensive, time-consuming task, which needs professionally trained experts in cardiology. In order to reduce the future human expert annotation work, active learning [83], [84], and cooperative learning [85] can be introduced, and applied to this dataset. Last but not least, the fundamental knowledge, e.g., the relationship between the acoustical representations and the anatomical changes in the heart, should be investigated deeply in future work.

VI. Conclusion

In this study, we firstly introduced a publicly accessible database, i.e., HSS. Then, the state-of-the-art techniques in physiological audio classification were described. A benchmark experiment was given based on the methodologies proposed by this work. We discussed the results and the limitations, and pointed out some future directions. A SVM model trained within the Log Mel features achieved the best UAR (49.7%) in this work.

Acknowledgment

The authors would like to thank the colleagues who collected and annotated the heart sound corpus. The HSS corpus (train and dev sets) will be provided by request currently (only for research purpose), and will be released partially to public for research purpose in future challenges or workshops.

References

[1] E. Wilkins *et al.*, *European Cardiovascular Disease Statistics*. Brussels, Belgium: European Heart Network, 2017.

[2] D. Roy, J. Sargeant, J. Gray, B. Hoyt, M. Allen, and M. Fleming, "Helping family physicians improve their cardiac auscultation skills with an interactive CD-ROM," *J. Continuing Educ. Health Professions*, vol. 22, no. 3, pp. 152–159, 2002.

[3] S. Mangione, "Cardiac auscultatory skills of physicians-in-training: A comparison of three english-speaking countries," *Amer. J. Med.*, vol. 110, no. 3, pp. 210–216, 2001.

[4] S. Ismail, I. Siddiqi, and U. Akram, "Localization and classification of heart beats in phonocardiography signalsa comprehensive review," *EURASIP J. Adv. Signal Process.*, vol. 2018, no. 1, 2018, Art. no. 26.

[5] G. D. Clifford *et al.*, "Recent advances in heart sound analysis," *Physiological Meas.*, vol. 38, no. 8, pp. E10–E25, 2017.

[6] P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor, *The PASCAL Classifying Heart Sounds Challenge*. 2011. [Online]. Available: http://www.peterjbentley.com/heartchallenge/index.html

[7] C. Liu *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiological Meas.*, vol. 37, no. 12, pp. 2181–2213, 2016.

[8] B. Schuller *et al.*, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 122–126.

[9] F. Renna, J. H. Oliveira, and M. T. Coimbra, "Deep convolutional neural networks for heart sound segmentation," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2435–2445, Nov. 2019.

[10] V. Nigam and R. Priemer, "Accessing heart dynamics to estimate durations of heart sounds," *Physiological Meas.*, vol. 26, no. 6, 2005, Art. no. 1005.

[11] Z. Syed, D. Leeds, D. Curtis, F. Nesta, R. A. Levine, and J. Guttag, "A framework for the analysis of acoustical cardiac signals," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 4, pp. 651–662, Apr. 2007.

[12] S. Choi and Z. Jiang, "Comparison of envelope extraction algorithms for cardiac sound signal segmentation," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 1056–1069, 2008.

[13] S. E. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent hidden markov model," *Physiological Meas.*, vol. 31, no. 4, pp. 513–529, 2010.

[14] H. Tang, T. Li, T. Qiu, and Y. Park, "Segmentation of heart sounds based on dynamic clustering," *Biomed. Signal Process. Control*, vol. 7, no. 5, pp. 509–516, 2012.

[15] A. Moukadem, A. Dieterlen, N. Hueber, and C. Brandt, "A robust heart sounds segmentation module based on s-transform," *Biomed. Signal Process. Control*, vol. 8, no. 3, pp. 273–281, 2013.

[16] V. N. Varghees and K. Ramachandran, "A novel heart sound activity detection framework for automated heart sound analysis," *Biomed. Signal Process. Control*, vol. 13, pp. 174–188, 2014.

[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 Informat.*, vol. 18, no. 4, pp. 1138–1152, Jul. 2014.

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

[19] T.-E. Chen *et al.*, "S1 and S2 heart sound recognition using deep neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 372–380, Feb. 2017.

[20] I. Y. Ozbek and H. Shamsi, "Heart sound localization in respiratory sound based on a new computationally efficient entropy bound," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 105–114, Jan. 2017.

[21] E. Messner, M. Zöhrer, and F. Pernkopf, "Heart sound segmentation—An event detection approach using deep recurrent neural networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1964–1974, Sep. 2018.

[22] S. Das, S. Pal, and M. Mitra, "Supervised model for cochleagram feature based fundamental heart sound identification," *Biomed. Signal Process. Control*, vol. 52, pp. 32–40, 2019.

[23] Y. M. Akay, M. Akay, W. Welkowitz, and J. Kostis, "Noninvasive detection of coronary artery disease," *IEEE Eng. Med. Biol. Mag.*, vol. 13, no. 5, pp. 761–764, Nov./Dec. 1994.

[24] P. Bentley, P. Grant, and J. McDonnell, "Time-frequency and time-scale techniques for the classification of native and bioprosthetic heart valve sounds," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 1, pp. 125–128, Jan. 1998.

[25] C. Ahlstrom *et al.*, "Feature extraction for systolic heart murmur classification," *Ann. Biomed. Eng.*, vol. 34, no. 11, pp. 1666–1677, 2006.

[26] P. Wang, C. S. Lim, S. Chauhan, J. Y. A. Foo, and V. Anantharaman, "Phonocardiographic signal analysis method using a modified hidden Markov model," *Ann. Biomed. Eng.*, vol. 35, no. 3, pp. 367–374, 2007.

[27] I. Maglogiannis, E. Loukis, E. Zafiropoulos, and A. Stasis, "Support vectors machine-based identification of heart valve diseases using heart sounds," *Comput. Methods Programs Biomed.*, vol. 95, no. 1, pp. 47–61, 2009.

[28] L. Avendano-Valencia, J. Godino-Llorente, M. Blanco-Velasco, and G. Castellanos-Dominguez, "Feature extraction from parametric time–frequency representations for heart murmur detection," *Ann. Biomed. Eng.*, vol. 38, no. 8, pp. 2716–2732, 2010.

[29] S. Ari, K. Hembram, and G. Saha, "Detection of cardiac abnormality from PCG signal using LMS based least square SVM classifier," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8019–8026, 2010.

[30] 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.

[31] H. Uğuz, "Adaptive neuro-fuzzy inference system for diagnosis of the heart valve diseases using wavelet transform with entropy," *Neural Comput. Appl.*, vol. 21, no. 7, pp. 1617–1628, 2012.

[32] S. E. Schmidt, C. Holst-Hansen, J. Hansen, E. Toft, and J. J. Struijk, "Acoustic features for the identification of coronary artery disease," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 11, pp. 2611–2619, Nov. 2015.

[33] S. Patidar, R. B. Pachori, and N. Garg, "Automatic diagnosis of septal defects based on tunable-q wavelet transform of cardiac sound signals," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3315–3326, 2015.

[34] Y. Zheng, X. Guo, and X. Ding, "A novel hybrid energy fraction and entropy-based approach for systolic heart murmurs identification," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2710–2721, 2015.

[35] A. Gharehbaghi, M. Borga, B. J. Sjöberg, and P. Ask, "A novel method for discrimination between innocent and pathological heart murmurs," *Med. Eng. Phys.*, vol. 37, no. 7, pp. 674–682, 2015.

[36] J. E. Guillermo, L. J. R. Castellanos, E. N. Sanchez, and A. Y. Alanis, "Detection of heart murmurs based on radial wavelet neural network with kalman learning," *Neurocomputing*, vol. 164, pp. 307–317, 2015.

[37] S.-W. Deng and J.-Q. Han, "Towards heart sound classification without segmentation via autocorrelation feature and diffusion maps," *Future Gener. Comput. Syst.*, vol. 60, pp. 13–21, 2016.

[38] V. Maknickas and A. Maknickas, "Recognition of normal–abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients," *Physiological Meas.*, vol. 38, no. 8, pp. 1671–1684, 2017.

[39] P. Langley and A. Murray, "Heart sound classification from unsegmented phonocardiograms," *Physiological Meas.*, vol. 38, no. 8, pp. 1658–1670, 2017.

[40] F. Plesinger, I. Viscor, J. Halamek, J. Jurco, and P. Jurak, "Heart sounds analysis using probability assessment," *Physiological Meas.*, vol. 38, no. 8, pp. 1685–1700, 2017.

[41] 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," *Physiological Meas.*, vol. 38, no. 8, pp. 1701–1713, 2017.

[42] V. N. Varghees and K. Ramachandran, "Effective heart sound segmentation and murmur classification using empirical wavelet transform and instantaneous phase for electronic stethoscope," *IEEE Sensors J.*, vol. 17, no. 12, pp. 3861–3872, Jun. 2017.

[43] W. Zhang, J. Han, and S. Deng, "Heart sound classification based on scaled spectrogram and partial least squares regression," *Biomed. Signal Process. Control*, vol. 32, pp. 20–28, 2017.

[44] M. Hamidi, H. Ghassemian, and M. Imani, "Classification of heart sound signal using curve fitting and fractal dimension," *Biomed. Signal Process. Control*, vol. 39, pp. 351–359, 2018.

[45] Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, and B. Schuller, "Learning image-based representations for heart sound classification," in *Proc. Int. Conf. Digit. Health*, 2018, pp. 143–147.

[46] S. Amiriparian, M. Schmitt, N. Cummins, K. Qian, F. Dong, and B. Schuller, "Deep unsupervised representation learning for abnormal heart sound classification," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 4776–4779.

[47] G. Eslamizadeh and R. Barati, "Heart murmur detection based on wavelet transformation and a synergy between artificial neural network and modified neighbor annealing methods," *Artif. Intell. Med.*, vol. 78, pp. 23–40, 2017.

[48] M. E. Karar, S. H. El-Khafif, and M. A. El-Brawany, "Automated diagnosis of heart sounds using rule-based classification tree," *J. Med. Syst.*, vol. 41, no. 4, 2017, Art. no. 60.

[49] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE–the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[50] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 835–838.

[51] F. Eyben, *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction*. Cham, Switzerland: Springer, 2015.

[52] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[53] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999, pp. 185–208.

[54] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[55] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[56] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[58] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2553–2561.

[59] N. Cummins, A. Baird, and B. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[60] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.

[61] S. Hochreiter *et al.*, "Gradient flow in recurrent nets: The difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*, J. F. Kolen and S. C. Kremer, Eds. Piscataway, NJ, USA: IEEE Press, 2001, pp. 237–244.

[62] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990.

[63] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[64] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Deep Learn. Representation Learn. Workshop*, 2014, pp. 1–9.

[65] K. Qian, *Automatic General Audio Signal Classification*. Munich, Germany: Technical Univ. Munich, 2018.

[66] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in *Proc. Int. Symp. Intell. Data Anal.*, 2001, pp. 309–318.

[67] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 312–315.

[68] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011. [Online]. Available: software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

[69] A. I. Humayun, M. Khan, S. Ghaffarzadegan, Z. Feng, and T. Hasan, "An ensemble of transfer, semi-supervised and supervised learning methods for pathological heart sound classification," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 127–131.

[70] G. Gosztolya, T. Grósz, and L. Tóth, "General utterance-level feature extraction for classifying crying sounds, atypical & self-assessed affect and heart beats," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 531–535.

[71] S. Rawat, P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze, "Robust audio-codebooks for large-scale event detection in consumer videos," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 2929–2933.

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

[73] M. Schmitt and B. Schuller, "openXBOW-introducing the Passau open-source crossmodal bag-of-words toolkit," *J. Mach. Learn. Res.*, vol. 18, no. 96, pp. 1–5, 2017.

[74] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6340–6344, 2017.

[75] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2017, pp. 17–21.

[76] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[77] Z. Zhang, J. Han, K. Qian, C. Janott, Y. Guo, and B. Schuller, "Snore-GANs: Improving automatic snore sound classification with synthesized data," *IEEE J. Biomed. Health Informat.*, to be published.

[78] K. Qian, C. Janott, Z. Zhang, C. Heiser, and B. Schuller, "Wavelet features for classification of vote snore sounds," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 221–225.

[79] K. Qian *et al.*, "Snore sound recognition: on wavelets and classifiers from deep nets to kernels," in *Proc. IEEE Eng. Med. Biol. Soc.*, 2017, pp. 3737–3740.

[80] K. Qian *et al.*, "Classification of the excitation location of snore sounds in the upper airway by acoustic multi-feature analysis," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1731–1741, Aug. 2017.

[81] K. Qian *et al.*, "Teaching machines on snoring: A benchmark on computer audition for snore sound excitation localisation," *Archives Acoust.*, vol. 43, no. 3, pp. 465–475, 2018.

[82] K. Qian *et al.*, "A bag of wavelet features for snore sound classification," *Ann. Biomed. Eng.*, vol. 47, no. 4, pp. 1000–1011, 2019.

[83] K. Qian, Z. Zhang, A. Baird, and B. Schuller, "Active learning for bird sounds classification," *Acta Acustica United Acustica*, vol. 103, no. 3, pp. 361–364, 2017.

[84] K. Qian, Z. Zhang, A. Baird, and B. Schuller, "Active learning for bird sound classification via a kernel-based extreme learning machine," *J. Acoustical Soc. Amer.*, vol. 142, no. 4, pp. 1796–1804, 2017.

[85] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 115–126, Jan. 2015.