A Smart Normal/Abnormal Heart Sound Classifier using NVIDIA Jetson™ Developer and Multiple AI Algorithms

Xingyao Wang[1,2], Yao Zhang[3], Yang Zhang[3], Hongxiang Gao[1], Chengyu Liu[1*]

Affiliations:

[1] School of Instrument Science and Engineering, Southeast University, Nanjing, China

[2] School of Information Science and Engineering, Southeast University, Nanjing, China

[3] Lenovo Group, Beijing, China

* Corresponding author: Chengyu Liu, E-mail: chengyu@seu.edu.cn

Abstract: In the past few decades, analysis of heart sound signals (i.e., the phonocardiogram or PCG), especially for automated heart sound segmentation and classification, has been widely studied and has been reported to have the potential value to detect pathology accurately in clinical applications. However, the current electronic auscultation device can only record and store heart sound data and without any AI-based analysis and clinic-condition functions.So we develop the smart heart sound device. This device is based on the NVIDIA Jetson TX2 and we build AI models with combination of CNN and RNN. For data, we use the data from Physionet/CinC Challenge 2016 to train the model. Eventually the smart heart sound device can record and appear people's heart sound signal, automatically analyze the signal to see whether is normal/abnormal and feed back to users. In addition, the device can be accessed to IoT system, so the users' data is able to be sent to family doctors immediately and doctors will find out the specific disease.

Key words: heart sound; phonocardiogram (PCG); heart sound classification; heart sound segmentation

1. Introduction

1.1 NVIDIA® Jetson™ Developer Challenge

NVIDIA Jetson Developer Challenge's purpose is to encourage all great developers, engineers, scientists, startups, and students to transform robotics, industrial IoT, healthcare, security, or any other industry with a powerful AI solution built on the **NVIDIA® Jetson™** platform.

Our team did heart sound signal processing for a long time. We start from the traditional algorithm, time domain, frequency-domain and non-linear analysisand then combine with the quickly developedAI technologies, we worked with cooperation with Yao and Yang from Lenovo. So when we saw the challenge from Challenge Rocket Website, we decided to develop a Intelligent medical device to do realtime analysis on heart sound signal.

1.2 Heart sound-based cardiovascular diseases detection

Cardiovascular diseases (CVDs) have the most patients (about 422.7 million) in the world and continue to be the leading cause of morbidity and mortality worldwide. An estimated 17.3 million people died from CVDs, representing 31% of all global deaths [1] (see Figure 1).

One of the first steps in evaluating the cardiovascular system in clinical practice is the auscultation of the heart sounds, which is an essential part of the physical examination and may reveal many pathologic cardiac conditions such as arrhythmias, valve disease, heart failure, and more. Heart sounds provide important initial clues in disease evaluation, serve as a guide for further diagnostic examination, and thus play an important role in the early detection for CVDs [2].

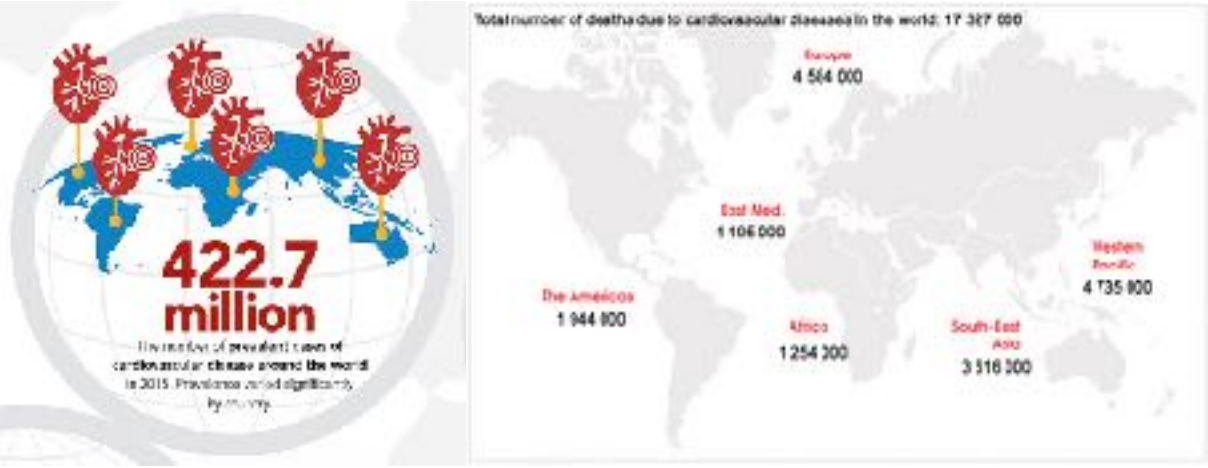

Figure 1. Prevalent of cardiovascular diseases (CVDs) in the world. Left: total number of the patients; right: total number of deaths due to CVDs in the world.

During the cardiac cycle, the heart first experiences electrical activation, which then leads to mechanical activity in the form of atrial and ventricular contractions. This in turn forces blood between the chambers of the heart and around the body, as a result of the opening and closure of the heart valves. This mechanical activity, and the sudden start or stop of the flow of blood within the heart, gives rise to vibrations of the entire cardiac structure [3]. These vibrations are audible on the chest wall, and listening for specific heart sounds can give an indication of the health of the heart. An audio recording (or graphical) time series representation of the resultant sounds, transduced at the chest surface is known as a heart sound recording or phonocardiogram (PCG).

Four locations are most often used to listen to and transduce the heart sounds, which are named according to the positions in which the valves can be best heard [4]:

- Aortic area - centred at the second right intercostal space.
- Pulmonic area - in the second intercostal space along the left sternal border.
- Tricuspid area - in the fourth intercostal space along the left sternal edge.
- Bicuspid area - at the cardiac apex, in the fifth intercostal space on the midclavicular line.

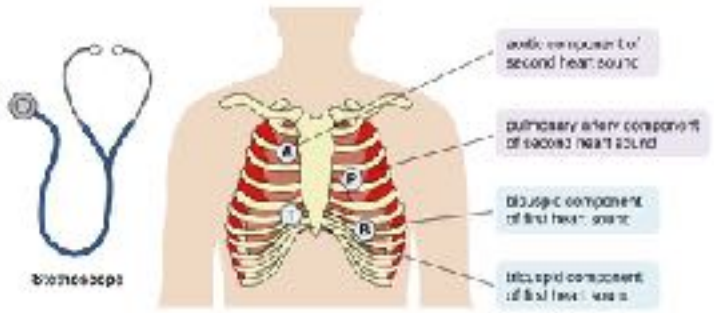

Figure 2. Four typical locations used to listen to heart sound signals.

Fundamental heart sounds (FHSs) usually include the first (S1) and second (S2) heart sounds [3]. S1 occurs at the beginning of isovolumetric ventricular contraction, when already closed mitral and tricuspid valves suddenly reach their elastic limit due to the rapid increase in pressure within the ventricles. S2 occurs at the beginning of diastole with the closure of the aortic and pulmonic valves (See Figure 3.) While the FHSs are the most recognizable sounds of the heart cycle, the mechanical activity of the heart may also cause other audible sounds, such as the third heart sound (S3), the fourth heart sound (S4), systolic ejection click (EC), mid-systolic click (MC), the diastolic sound or opening snap (OS), as well as heart murmurs caused by turbulent, high-velocity flow of blood.

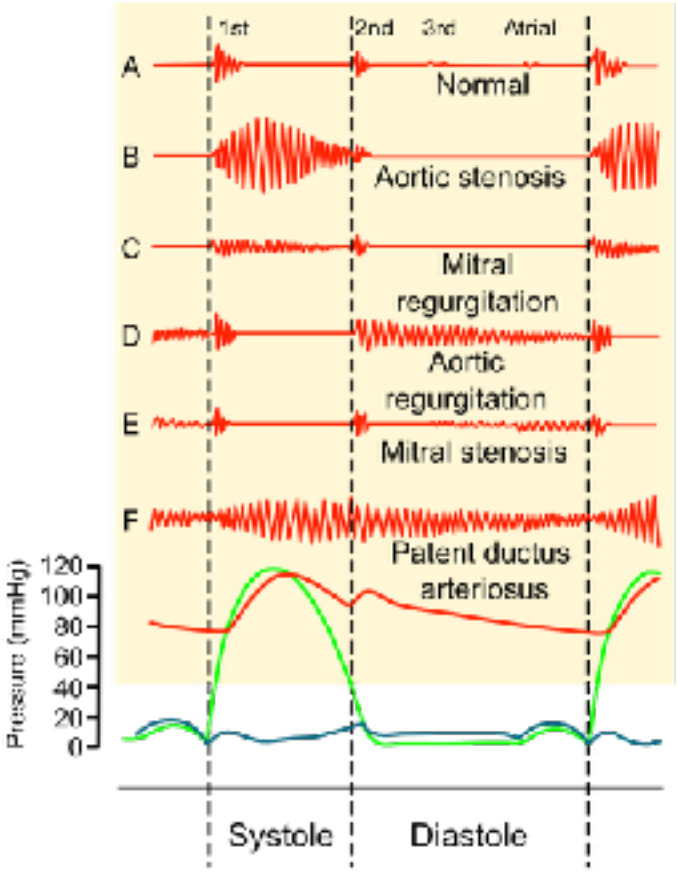

Figure 3. Phonocardiograms (above) from normal and abnormal heart sounds with pressure diagrams (below). Red indicates aortic pressure, green ventricular pressure and blue atrial pressure. Reproduced under the CC BY-SA 3.0 license and adapted from [5].

1.3 Aim of this challenge

Our purpose is to develop a smart heat sound device which has two specific features:

- Small and Portable

As an intelligent terminal product, we hope that the heart sound smart box will be an assistant to a doctor and a regular home health guardian.

- IoT system

The heart sound intelligent box can not only directly calculate the initial diagnosis results at the front end, but also connect with the server, store the data in the cloud, and send it to the doctor at any time and anywhere. In remote mountainous areas and rural areas where medical levels are backward, people can make a preliminary screening of diseases with the help of smart heart sound boxes. Meanwhile, with the help of cloud, patients can transmit data to doctors far away from the city, and doctors can make further diagnosis and instructions based on the collected heart sounds.

2. System architecture

2.1 Summary of system architecture

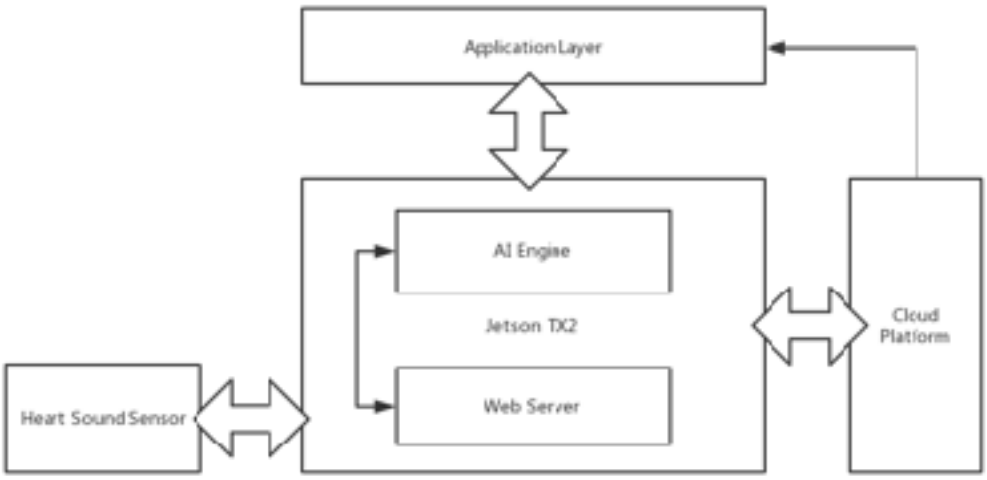

Figure 4. Architecture of the developed smart heart sound system.

3. AI-based classification algorithms

3.1 Signal pre-processing

3.2 Heart sound segmentation

Springer's heart sound segmentation algorithm is a hidden HSMM-based segmentation method. The HSMM model needs to be trained with the reference annotation of S1 and S2

sounds in order to obtain the model parameters. Once the model training is completed, the model can be used for segmenting heart sound recordings directly without any other input information (other than the derived features). In this study, the HSMM model was trained using the training-a database since it includes both heart sound and ECG signals and provides a more robust and independent location of the heart sounds. The training process was detailed described in [6]. We briefly summarize it in four steps [7]:

- Step 1: Obtain the locations of R-peak and end-T-wave in ECG signals as the reference positions for S1 and S2 sounds. R-peaks were detected and confirmed using the combination of four detectors: "gQRS" [8], "jQRS" [9], a parabolic fitting method [10] and wavelet-based method [11]. T-wave end points were also detected and confirmed using the combination of four detectors: "ecgpuwave" [8, 12], a sliding window area method [13], a wavelet-based method [11] and the Trapezium area method [14]. The agreement between the R-peak and end-T-wave detectors was assessed to derive the ECG signal quality index. First, the agreement between all four R-peak detectors was measured as an F_1 score using the "bxb" algorithm, available from Physionet [8]. Then, the R-peak detector with the lowest F_1 score was excluded. Over 4 s window, the ECG signal quality was labeled as the F_1 score of agreement between the remaining three R-peak detectors. In windows with 100% F_1 score, ECG episodes were determined as good signal quality if all three R-peak detectors were within the 100 ms tolerance. For detected end-T-wave positions, the annotation furthest from the median of the four annotations was excluded. Then ECG episodes were determined as good signal quality if the remaining three annotations were all within the 100 ms tolerance of each other. ECG episodes corresponding to poor signal quality were excluded for training the HSMM model.
- Step 2: The four heart sound states were labeled using the reference locations of R-peak and end-T-wave as shown in Figure 1. The period from each detected R-peak plus the mean S1 duration was labelled as an S1 sound. The maximum value of the Hilbert envelope of heart sound within a given window centred on the end-T-wave, was marked as centre of S2 sound. The period equal to mean S2 duration, centred on this maximum position, was labelled as an S2 sound. The period between S1 and S2 was labelled as systole, and the period between S2 and S1 in next beat cycle was labelled as diastole. The mean S1 duration was set to 122 ms, the mean S2 duration was set to 92 ms, the special window was set as the mean S2 duration plus the standard deviation of S2, i.e., 114 ms. All the parameter values were reported from the Schmidt's HSMM model [15].
- Step 3: Employ Springer's HSMM approach. This model was developed from the Schmidt's HSMM approach, which is a standard HMM (i.e., $\lambda = (A, B, \pi)$) plus a probability model of the time remaining in each heart sound state, i.e., $\lambda = (A, B, \pi, p)$, where A is the transmission matrix of the four heart sound states, B is the observation distribution matrix, π is the initial heart sound state distribution and p is the probability density function of the time expected to remain in each heart sound state. Since p was added to the iteration process of B, only the state transition matrix A is Markovian. Therefore the model is referred to as a hidden semi-Markov model. Springer's HSMM model improved the Schmidt's HSMM model in three ways: 1) It uses a logistic regression derived observation function for B matrix to replace the Gaussian distribution function; 2) it extended the Viterbi algorithm to predict the possible state durations beyond the beginning and end of the heart sound signal, to give the state durations at the boundary points; and 3) it uses a combination of four

envelope features of the heart sounds for the model inputs, including the homomorphic envelope, Hilbert envelope, wavelet envelope and power spectral density envelope [6].

- Step 4: Parameter setting and model training. In Springer's HSMM model, the transmission probabilities in matrix A are initialized to 0, except for the possible transitions between successive states (i.e., S1→systole, systole→S2, S2→diastole and diastole→S1), which were set to 1. The initial state distribution probabilities in matrix π were set to be equal to 0.25 for all four states. The matrix $B$and p were trained by running the modified Viterbi algorithm over the training data. The four envelopes mentioned above were calculated and were normalized on a per-recording basis by subtracting the mean and dividing by the standard deviation of each recording. After normalization, the four envelope feature vectors are down-sampled to 50 Hz using a poly-phase antialiasing filter to increase the speed of computation. Their envelope feature vectors, as well as the corresponding reference annotation of the four heart sound states, are inputted into Springer's HSMM model for training.

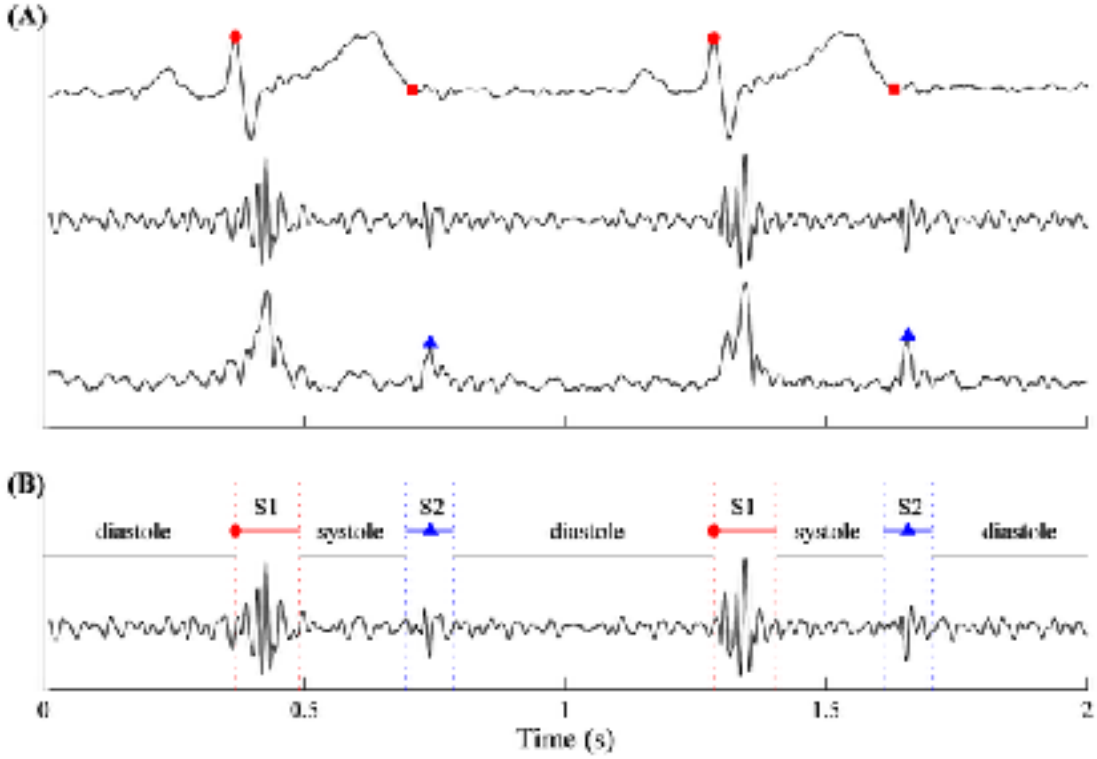

Figure 1. Demonstration of how to label the four heart sound states. (A) From top to bottom, ECG, heart sound and the Hilbert envelope of heart sound were shown. The detected R-peak (circle '●') and end-T-wave (square '▪') in ECG signal, as well as the locations of the maximum value from the Hilbert envelope of heart sound (triangle '▲'), were also given. (B) Heart sound states were labelled using the reference locations of R-peak (circle '●') and the maximum value from the Hilbert envelope of heart sound (triangle '▲') using the threshold parameters reported in the Schmidt's HSMM model [15]. Adapted from [7].

3.3 Feature calculation

Springer's segmentation code [16] was used to segment each selected heart sound recording to generate the time durations for the four states: S1, systole, S2 and diastole. Twenty features were extracted from the position information of the four states as follows [2, 7]:

1. m_RR: mean value of RR intervals
2. sd_RR: standard deviation (SD) of RR intervals
3. m_IntS1: mean value of S1 intervals
4. sd_IntS1: SD of S1 intervals
5. m_IntS2: mean value of S2 intervals
6. sd_IntS2: SD of S2 intervals
7. m_IntSys: mean of systolic intervals
8. sd_IntSys: SD of systolic intervals
9. m_IntDia: mean of diastolic intervals
10. sd_IntDia: SD of diastolic intervals
11. m_Ratio_SysRR: mean of the ratio of systolic interval to RR of each heart beat
12. sd_Ratio_SysRR: SD of the ratio of systolic interval to RR of each heart beat
13. m_Ratio_DiaRR: mean of ratio of diastolic interval to RR of each heart beat
14. sd_Ratio_DiaRR: SD of ratio of diastolic interval to RR of each heart beat
15. m_Ratio_SysDia: mean of the ratio of systolic to diastolic interval of each heart beat
16. sd_Ratio_SysDia: SD of the ratio of systolic to diastolic interval of each heart beat
17. m_Amp_SysS1: mean of the ratio of the mean absolute amplitude during systole to that during the S1 period in each heart beat
18. sd_Amp_SysS1: SD of the ratio of the mean absolute amplitude during systole to that during the S1 period in each heart beat
19. m_Amp_DiaS2: mean of the ratio of the mean absolute amplitude during diastole to that during the S2 period in each heart beat
20. sd_Amp_DiaS2: SD of the ratio of the mean absolute amplitude during diastole to that during the S2 period in each heart beat

3.4 AI-based classification method

Acoustic signal is complex and difficult to classify. Using handcrafted feature extraction and shallow classifiers must learn the distribution of different data in advance. Thus we aim to design a heart sound classifier utilizing deep features. Fig. 1 shows the flowchart of our method.

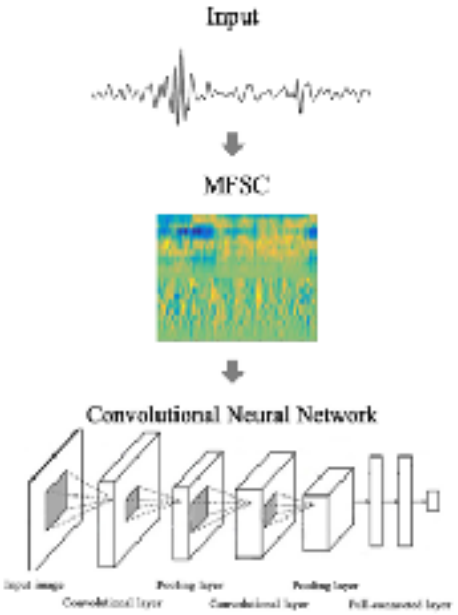

Fig. 1 The flow chart of our method

1. Spectrum

For time-series data like voice data or sensor data, the spectrogram visual representation is essential for extracting interpretable features. It is a very detailed, accurate image of audio, displayed in either 2D or 3D. Audio is shown on a graph according to time and frequency, with brightness or height indicating amplitude. Whereas a waveform shows how acoustic signals amplitude changes over time, the spectrogram shows this change for every frequency component in the signal. Moreover, selection of frequency-domain resolution in the spectrogram also provides an implicit way to remove noise in the data.

Specifically, the spectrogram is the magnitude squared of the short-time Fourier transform(STFT). It would divide time signal to short segments of equal length and then computing STFT on each segment. Suppose we have a signal x[n] and for every time m we multiply it by a window of length N and we take the FFT. This can be expressed as:

$$X_{STFT}(e^{j\omega}, m) = X(\omega, m) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n}$$

The spectrogram is the magnitude square of the elements in STFT{X}:

$$X_{spectrogram}(e^{j\omega}, m) = |X_{STFT}(e^{j\omega}, m)|^2$$

1. MFSC

Mel-frequency spectral coefficients (MFSC) can by derived from Mel-frequency cepstral coefficients (MFCC). MFCC simulates the auditory characteristics of humans, and its sensitivity to different frequency signals is different. It can not only extract the semantic information, but also can extract the speaker's characteristics. The acoustic signal x(n) is preprocessed as $x_i(m)$ where i represents the i-th frame signal. And then we compute fast fourier transformation (FFT) and power spectrum for each frame. MFSC is defined as:

$$S(i,m) = \sum_{k=0}^{N-1} E(i,k) H_m(k) \quad 0 \le m < M$$

where i represents i-th frame signal, k represents the k-th line in the frequency domain, M is the number of channels, and $H_m(k)$ refers to the magnitude frequency response of the filter m.

3、Convolutional Neural Network

Once the spectrogram and MFCC have been computed, they can be processed by the deep learning model. Deep learning is a new area of machine learning research. Its algorithms transform their inputs through more layers than shallow learning algorithms. A large set of layers can be built to extract a hierarchy of features from low level to high level. Deep learning models include Auto-encoder, Deep Neural Network (DNN), Deep Belief Networks (DBN), Convolutional Neural Network (CNN) and etc. CNN is the most efficient approach for image and speech recognition. The major difference between CNN and ordinary Neural Networks is that CNN architectures make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture. These then make the forward function more efficient to implement and vastly reduce the amount of parameters in the network. The convolutional neural networks architecture and training procedure are shown in Fig. 2.

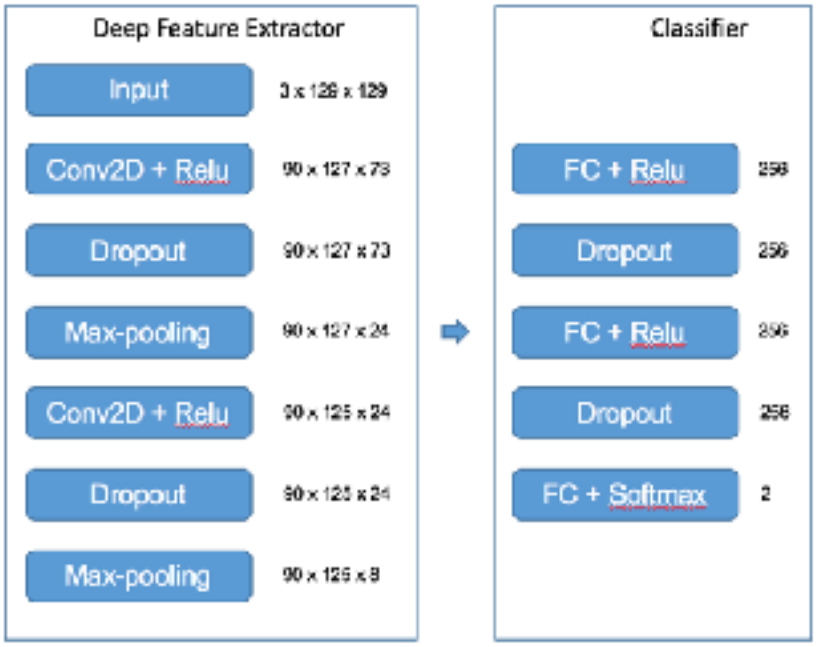

Fig. 2 The architecture of CNN

Convolutional layer is to capture the local dependencies of the data. Each convolutional layer is computed as:

$$y_j^l = \sigma(b_j^l + \sum_i k_{ij} * x_i^{l-1})$$

Another key characteristic of CNN is scale-invariant feature preservation, which is achieved by the pooling layer. The pooling layer is given by:

$$y_j^l = \sigma(\beta_j^l \cdot down(x_j^{l-1}) + b_j^l)$$

During the training process, our goal is to minimize the loss function in the back-ward propagation. The optimizers such as stochastic gradient descent (SGD), RMSprop and Adam are used to update the weights of hidden layers.

3.5 Evaluation approach

Let $\{x_1, x_2, \cdots, x_i, \cdots, x_N\}$ denote the manually annotated onset positions for one of the four heart sound states. A tolerance parameter δ is used for determining the true positive (TP), false positive (FP) and false negative (FN) segmentation for the evaluated heart sound

segmentation algorithms. For the ith manually annotated onset position x_i, we counted the numbers of the state onsets from the automatic segmentation results in two time regions: $x_i - \delta \leq \text{stateonsets} \leq x_i + \delta$ and $x_i + \delta < \text{stateonsets} < x_{i+1} - \delta$. Let N_1 and N_2 denote the counted numbers in these two time regions respectively.

For the current heart sound state, the automatically segmented onset was expected to appear in the time region $x_i - \delta \leq \text{stateonsets} \leq x_i + \delta$ and should notappear in the other time interval $x_i + \delta < \text{stateonsets} < x_{i+1} - \delta$. The TP, FP and FN for each manually annotated heart beat cycle were then defined as:

- TP: if $N_1 > 0$, $TP = TP + 1$, means that there is an expected state onset in the expected time region.
- FP: 1) if $N_1 > 1$, $FP = FP + N_1 - 1$, means that there are more than one segmented state onsets in the expected time region; 2) if $N_2 > 0$, $FP = FP + N_2$, means there is/are false segmented state onset/onsets in the unexpected time region.
- FN: if $N_1 = 0$, $FN = FN + 1$, means that there is a missing segmented state onset in the expected time region.

The metrics of sensitivity (Se, or *recall*), positive predictivity (P_+, or *precision*) and accuracy (Acc) are therefore defined as:

$$Se = \frac{TP}{TP + FN} \times 100\% \tag{1}$$

$$P_+ = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

$$Acc = \frac{TP}{TP + FP + FN} \times 100\% \tag{3}$$

However, the Acc metric does not provide an adequate representation of the results since no true negatives are included. We therefore also calculated F_1, the harmonic mean of Se and P_+ as in [6], defined as:

$$F_1 = \frac{2 \times Se \times P_+}{Se + P_+} \times 100\% \tag{4}$$

The Se, P_+ and Acc metrics are also calculated and reported to allow comparison to previous works.

The tolerance parameter δ (defining how close a detection and a annotation can be to cunt as a match) has a key effect on the evaluation metrics. For QRS detection evaluation in ECG signals, the ANSI/AAMI EC57 standard recommends a tolerance of 150 ms for identifying a

coincident automatic and reference ECG R-peak annotation [17]. In the PhysioNet/CinC Challenge 2013, focused on fetal ECG beat detection, a tolerance of 100 ms was used for determining the true/false matching between reference and detected annotations [18, 19]. For heart sound signals, four state onsets within a single heart cycle, which are shorter than 100 ms must be determined as true or false onset/offset detections. The tolerance therefore must be shorter. In Schmidt's HSMM approach, a tolerance of 60 ms was used and a heart sound was determined as a TP if the middle of the segmented sound state (S1 and S2) was closer than this tolerance to the middle of the manually annotated state [15]. In Springer's HSMM approach, the references were ECG features and thus a tolerance of 100 ms was used, i.e., a TP S1 onset was found to be within this tolerance of the R-peak, and a TP S2 onset was found if the centre of the automatically segmented S2 sound was within this tolerance of the corresponding end-T-wave [6]. For this study, we evaluated the performance of Springer's algorithm using tolerances of 20, 40, 60, 80 and 100 ms, in order to test the sensitivity of the evaluation metrics to this tolerance.

4. Experiments

4.1 Training data

The heart sound recordings for training the classification model were from the PhysioNet/Computing in Cardiology Challenge 2016 (https://physionet.org/challenge/2016/), which included six databases from different data contributor in the world, containing a total of 3,240 heart sound recordings from 764 subjects/patients, lasting from 5 s to just over 120 s. The collected data included not only clean heart sounds but also very noisy recordings. They were recorded from both normal subjects and pathological patients, and from both children and adults. The recordings from the same patient did not appear in both the training and test datasets. The data were also recorded from different locations, depending on the individual protocols used for each database. However, they were generally recorded at the four common recording locations of aortic area, pulmonic area, tricuspid area and bicuspid area. All recordings were resampled to 2,000 Hz using an anti-alias filter and provided in a standard uncompressed (wav) format.

Heart sound recordings were divided into two types: normal and abnormal recordings. The normal recordings were from healthy subjects and the abnormal ones were from patients with a confirmed cardiac diagnosis. The patients were noted to suffer from a variety of illnesses but typically they are heart valve defects and CAD patients. Heart valve defects include mitral valve prolapse, mitral regurgitation, aortic regurgitation, aortic stenosis and valvular surgery. All the recordings from the patients were generally labeled as abnormal.

In clinical practice, the criteria adopted by the cardiologist to annotate the beginning and the ending of S1 and S2 sounds was defined as follows: the beginning of S1 is the start of the high frequency vibration due to mitral closure, the beginning of S2 is the start of the high frequency vibration due to aortic closure, and the endings of S1 and S2 are annotated by the end of the high frequency vibrations [20]. According to this criteria, manual annotations for the four heart

sound states (i.e., S1, systole, S2 and diastole) for each beat for the PhysioNet/CinC Challenge 2016 data were provided by the authors [2]. Some recordings, or some signal episodes in heart sound recordings, were so noisy that making the manual segmentation of the four heart sound states impossible. These recordings and episodes were manually annotated as 'noisy' and were excluded for the algorithm evaluation.

Table 1 summarizes the total numbers of patients, recordings (not including the noisy recordings) and manually annotated beats for each database, as well as the maximum (max), median and minimum (min) values of time lengths and manually annotated beats for each recording. As shown in Table 1, 409 heart sound recordings (totaling 14,559 beats) were used for training Springer's HSMM segmentation algorithm and the other independent 4,021 heart sound recordings from 951 healthy subjects/patients (totaling 102,306 beats) were used for testing.

Table 1. Summary of the heart sound data used in this study from the PhysioNet/CinC Challenge 2016.

Database	# patients	# recordings	Recording length (s)			# beats (manual annotation)			
			Min	Median	Max	Min	Median	Max	Total
training-a	121	409	9.3	35.6	36.5	12	37	78	14,559
training-b	106	490	5.3	8	8	4	9	15	3,353
training-c	31	31	9.6	44.4	122.0	15	67	143	1,808
training-d	38	55	6.6	12.3	48.5	6	14	72	853
training-e	356	2,054	8.1	21.1	101.7	4	27	174	59,593
training-f	112	114	29.4	31.7	59.6	7	39	75	4,259

Members:

Xingyao Wang
Master in Information Science and Engineering.
Primary clinical medical engineer.
Proficient in computer vision, embedded development and optimization and transplantation of the algorithm on the hardware platform.

Hongxiang Gao
Master student in Southeast University.
Major in Instrument Science and Technology.

Yao Zhang
Ph.D. student in Institute of Computing Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences.
Research interests include medical image analysis, deep learning, computer-aided diagnosis, etc.

Yang Zhang
PhD in Artificial Intelligence.
Major interest remains in the field of machine learning, deep learning, pattern recognition and medical data analysis.
Holding expertise and experience in designing AI systems, with a deep understanding of algorithms at each phase from feature extraction, optimization to decision making.

Chengyu Liu
Professor at Southeast University, Nanjing, China.
PhD in Biomedical Engineering.
Director of Southeast-Lenovo Wearable Heart-Sleep-Emotion Intelligent Monitoring Lab.
Federation Journal Committee Member of International Federation for Medical and Biological Engineering.
Research topics include: mHealth and intelligent monitoring, machine learning and big data processing for physiological signals, early detection of CADs, device development for CADs, sleep and emotion monitoring.