Heart Sound Classification based on Residual Shrinkage Networks

Lixian Zhu[1], *Student Member, IEEE*, Kun Qian*[1], *Senior Member, IEEE*, Zhihua Wang[3], Bin Hu*[1], *Senior Member, IEEE*, Yoshiharu Yamamoto[2], *Member, IEEE*, and Björn W. Schuller[4], *Fellow, IEEE*

***Abstract*— Heart sound classification is one of the non-invasive methods for early detection of the cardiovascular diseases (CVDs), the leading cause for deaths. In recent years, Computer Audition (CA) technology has become increasingly sophisticated, auxiliary diagnosis technology of heart disease based on CA has become a popular research area. This paper proposes a deep Convolutional Neural Network (CNN) model for heart sound classification. To improve the classification accuracy of heart sound, we design a classification algorithm combining classical Residual Network (ResNet) and Long Short-Term Memory (LSTM). The model performance is evaluated in the PhysioNet / CinC Challenges 2016 datasets using a 2 D time-frequency feature. We extract the four features from different filter-bank coefficients, including Filterbank (Fbank), Mel-Frequency Spectral Coefficients (MFSCs), and Mel-Frequency Cepstral Coefficients (MFCCs). The experimental results show the MFSCs feature outperforms the other features in the proposed CNN model. The proposed model performs well on the test set, particularly the F1 score of 84.3 % – the accuracy of 84.4 %, the sensitivity of 84.3 %, and the specificity of 85.6 %. Compared with the classical ResNet model, an accuracy of 4.9 % improvement is observed in the proposed model.**

I. Introduction

CVDs is the number one killer of people who die from the disease worldwide. According to the 2019 World Health Organization report, there are 9 million deaths worldwide because of heart disease, representing 16 % of all deaths [1]. While there are many ways to detect heart disease, patients usually need to be in the hospital for diagnosis. And people cannot make timely and effective diagnosis of heart health by themselves, which is an important reason for the high mortality rate of heart disease. Therefore, early aided diagnosis of heart disease is one of the most important ways to prevent heart disease.

The heart is the source of life-sustaining in the body and makes sound when it contracts. Heart sound contain a wealth of information about the state of heart health. Usually, people use a stethoscope to get a heart sound signal. Nevertheless, ordinary people cannot recognize abnormal heart sound signals because of the limited sensitivity of the human auditory system, and only experienced physicians can recognize heart sound. However, with the development of Computer Audition (CA) technology, heart sound classification based on artificial intelligence has become a popular research area. Currently, there are many artificial intelligence algorithms to classify heart sound maps, and there are two main categories of algorithms, machine learning and deep learning [2]–[6]. In the classical machine learning field, Goda et al. used the SVM method to classify the time-frequency features of heart tone signals [7]. Safara et al. used a combination method of wavelet and Support Vector Machines (SVM) for classification of heart sound [8]. In deep learning, Grzegorczyk et al. proposed a deep learning algorithm for classification of heart sound [9]. Humayun et al. designed a Convolutional Neural Network (CNN) model to identify heart sound using a time convolutional (tCONV) unit to simulate a Finite Impulse Response (FIR) filter [10]. Zhang et al. [11] designed a segmented CNN model, which uses two different designs to adjust the convolutional layers for cardiac abnormality detection. Muqing Deng et al. proposed the use of improved Mel-Frequency Spectral Coefficients (MFCCs) features combined with Rerrent Neural Network (RNN) model for heart sound classification [12]. The advantage of traditional machine learning is that the algorithm can achieve good classification results even with small sample size and features, yet the disadvantage is low anti-interference ability and poor robustness. Instead, deep learning can overfit or under fitting with fewer samples, and the explainable of deep learning is also a major drawback. However, it can full use feature information and achieve excellent results when the sample size is large enough, and the robustness is also better than machine learning under the same conditions. Therefore, deep learning has more potential for application in smart healthcare.

In this paper, we consider the time and frequency domain information of heart sound signals, and propose a joint network model for the classification of individual heart sound cycles, which combines the classical Residual Network (ResNet) [13] and Long Short-Term Memory (LSTM). In

This work was partially supported by the Ministry of Science and Technology of the People's Republic of China (2021ZD0201900), the BIT Teli Young Fellow Program from the Beijing Institute of Technology, China, and the Grants-in-Aid for Scientific Research (No. 20H00569) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. *Corresponding authors*: Kun Qian and Bin Hu.

[1]Kun Qian, Bin Hu, and Lixian Zhu are with the Laboratory on Brain Health Engineering (BHE), School of Medical Technology, Beijing Institute of Technology, No. 5 Zhongguancun South Street, Haidian District, Beijing 100081, China. {zhulx17, qian, bh}@bit.edu.cn

[2]Yoshiharu Yamamoto is with the Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. yamamoto@p.u-tokyo.ac.jp

[3]Zhihua Wang is with the School of Mechatronic Engineering, China University of Mining and Technology, Xuzhou, Jiangsu, China. w_z_hua@cumt.edu.cn

[4]Björn W. Schuller is with GLAM – the Group on Language, Audio & Music, Imperial College London, 180 Queens' Gate, Huxley Bldg., London SW7 2AZ, UK, and also with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Eichleitnerstr 30, Augsburg 86159, Germany. schuller@ieee.org

TABLE I
EXPERIMENTAL DATA

Subset	Abnormal Subject	Normal Subject	Total Subject
a	292	117	409
b	104	386	490
c	24	7	31
d	28	27	55
e	183	1958	2141
f	34	80	114
Total	665	2575	3240

TABLE II
SUMMARY OF DIFFERENT FEATURES

Name	Feature Description
Fbank	Filterbank energy features
MFSCs	Log-filterbank energy
MFCCs-26	MFCCs retaining all coefficients after DCT compression
MFCCs-13	MFCCs retaining 13 coefficients after DCT compression

this algorithm, we integrate the LSTM into the ResNet, firstly using the ResNet to extract the frequency domain information of the heart sound signal, and then put the frequency domain information into the LSTM to obtain the temporal information of the heart sound signal. In the end, a fully connected layer and a SoftMax layer are used for heart sound prediction. Meanwhile, we validated the performance of the proposed model using the publicly PhysioNet/CinC Challenge 2016 dataset, which is one of the well-known cardiology datasets available at present. In addition, we design two different CNN models to compare with the classical ResNet model, and study the performance of the proposed CNN model using different features.

II. Methods

A. Dataset

In this section, we use the Physionet/Cinc Challenge 2016 dataset to train and test the proposed model [14]. The dataset has 7 subsets, recorded by seven research groups using different devices in both clinical and non-clinical environment, with recording times ranging from several seconds to minutes [15]. Specifically, the dataset collected data from 3240 heart sound with 84 426 heartbeats, of which 2575 heart sound data were collected from the normal population and 665 data were collected from patients. Table I shows a brief description of the dataset.

B. Preprocessing

The preprocessing step of the heart sound is extremely important to eliminate the negative effects of different sampling rates and environmental noise for recognition of heart sound. In this section, there are two-stage for heart sound preprocessing. In the first step, a 3rd-order Butterworth bandpass with cut-off frequencies of 20 and 400 Hz [16] are used to filter out the noise. In step two, we adopt the method proposed by Schmidt [17] to eliminate spikes. It is worth noting heart sound preprocessing can standardises the dataset and removes the noise, such as lung sound, stethoscope fricatives, breathing sound and external environmental noise. Nevertheless, we do not focus on the heart sound preprocessing for the signal preprocessing techniques are well established.

C. Segmentation

There are two primary purposes for heart sound segmentation, expanding the dataset and extracting the entire cycle of individual heartbeat, which is essential for heart sound classification. Different methods of heart sound segmentation have been proposed. And the methods are generally classified into two types, one is the heart sound segmentation algorithm referenced to electrocardiogram (ECG) information annotation, and the other is direct segmentation with no referenced. In this section, Hidden Semi-Markov Models (HSMM) algorithm improved by Springer's [18] is used to segment the heart sound signal into individual heartbeat cycles, which does not require the ECG signal as a reference. In particular, the heart sound segmentation length is set to the longest heartbeat cycle length in the dataset [19], i.e., 2.5 s, and padding zeros are applied for data with a cardiac cycle of less than 2.5 s. This step effectively minimizes the effect of the imbalance of the dataset on the classification results.

D. Feature Extarction

MFCCs is a feature proposed based on the auditory characteristics of the human ear. It has an excellent performance in the area of acoustics and has been widely used in speech recognition systems [18]. Therefore, we adopt MFCCs features as input to test performance of the proposed model. In addition, we extract the other acoustic features based on the different filter-bank to validate the effect of various features on the proposed model, Table II shows the detailed information. In particular, Mel-Frequency Spectral Coefficients (MFSCs) is a special form of MFCCs, which omits the Discrete Cosine Transform (DCT) step with respect to MFCCs. Yet the MFSCs feature adds log operation compared to Fbank feature.

E. Residual Network Model

ResNet has been one of the hottest deep learning methods in the past five years [11]. Residual Block (RB) is the basic building block of the ResNet. As Figure. 1(a) depicted, the input data is performed in the RB unit with two Bach Normalization (BN), two convolution, and two Rectified Linear Unit (RELU), and then connected with an identity shortcut as the output of RB. It is worth noting that identity is a crucial step to deal with the explosion or disappearance of gradient, which is its merit in comparison with the traditional Convolutional Network (ConvNet). In the traditional ConvNet, the cross-entropy error is propagated backward layer by layer, while using identity allows the gradient to be passed to the nearer layers of the input layer earlier, thus updating the model parameters more efficiently.

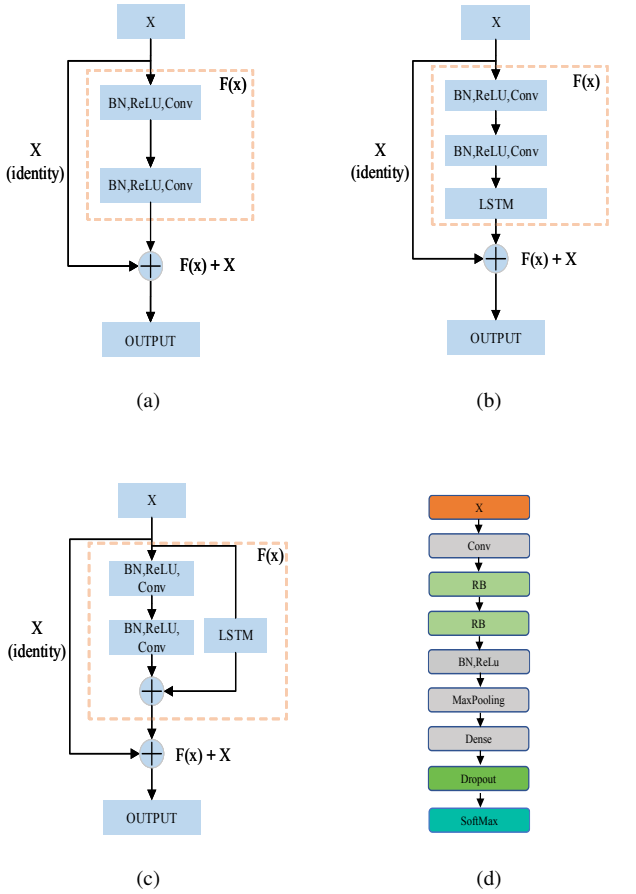

Fig. 1. (a) a RB unite of traditional ResNet, (b) a RB unite of RSN-SL, (c) a RB unite of RSN-PL, (d) a framework of RSN-SL.

In this paper, the idea for the model is derived from the classical ResNet, and we designed two types of residual shrinkage networks (RSN), the RSN series LSTM (RSN-SL) and the RSN parallel LSTM (RSN-PL), which are a variant of ResNet. Figure. 1(b) and Figure. 1(c) show the RB structures of RSN-SL and RSN-PL respectively, and they differ mainly in the location of the LSTM layers.

Figure. 1(d) illustrates the framework of RSN-SL, with the model included two RB units. Note that it adds a convolutional layer before the first RB. The motivation for adding convolutional layers is to increase the output feature map and thus integrate the different features into discriminative features. For instance, the model input is a MFCCs feature of shape 246 × 26 × 1, the output feature size is 246 × 26 × 8 after the convolutional layer. In addition, each RB adds the global average pooling (GAP) unit to match the LSTM layer, and incorporates L2 regularisation to prevent overfitting. And the kernel size in each RB unit is set to 3 × 3 and the number of output filters in the convolution is 8 and 16, respectively. Meanwhile, the stride of the first convolutional layer and the second convolutional layer are set to 2 and 1 respectively. Setting the stride to 2 reduces the width of the output features, the motivation for this step is to reduce the amount of calculation in the following layers.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT CLASSIFIERS ON THE TEST SET [%]

Method	F1 Score	Precision	Recall	Accuracy	Specificity
RSN	79.0	78.9	79.2	79.5	86.2
RSN-PL	80.3	80.3	80.2	80.5	82.9
RSN-SL	84.3	84.2	84.3	84.4	85.6

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT FEATURES USING THE RSN-SL MODEL [%]

Feature	F1 Score	Pprecision	Recall	Accuracy	Specificity
Fbank	82.5	82.9	82.1	81.1	84.3
MFSCs	84.3	84.2	84.3	84.4	85.6
MFCCs-26	79.2	79.3	79.2	79.5	78.0
MFCCs-13	83.6	83.46	83.7	83.7	85.7

III. Experimental and Results

A. Setup

We adopt TensorFlow (version–2.2.0) and Keras (version–2.3.1) to build our experimental environment, using a server configuration of i7-9700k and a GTX3080TI GPU. In the experiment, we divide the dataset into four folds randomly, each fold containing samples of "a-f" subset labels.In each subset, an individual subject contributes the data to both training an testing data. Further, a full training cycle carries out 100 epochs. The experimental results are calculated by the average of the 4 fold.

In this work, we first evaluated the performance of three models using MFCCs-13 features and deep learning indicators. These indicators include F1 score, accuarcy, recall, precision and specificity, which are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$Specificity = \frac{TN}{TN+FP} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

where TP indicates the number of true positives in abnormal samples, TN indicates the number of true negatives in normal samples, FP indicates the number of false positives in normal samples and FN indicates the number of false negatives in abnormal samples.

B. Results

Table III shows the performance of the different models when using MFSCs features. The RSN-SL model obtains the best performance, the F1 score and accuracy are 84.3 % and 84.4 %, respectively. Of notes, compared to RSN and RSN-PL, the F1 score of RSN-SL is 5.3 % and 4.0 % higher, respectively, while the accuracy of RSN-SL is 4.9 % and 3.9 % higher, respectively. Consequently, we chose the RSN-SL model to further investigate the effect of different features on the model.

Table IV illustrates the performance of the model with different features. The experimental results show that the MFSCs feature achieves the best performance with F1 score, precision, recall, accuracy and specificity of 84.3 %, 84.2%, 84.3%, 84.4%, and 85.6% respectively, which shows that Log operation of the feature can improve the performance of deep learning.

IV. Discussion

As can be seen from Table III, in the classical ResNet model, the F1 score and accuracy are 79.0 % and 79.5 %, respectively. Although the specificity of the RSN model is higher at 86.2 % than the RSN-SL at 85.6 %, it is lower than the RSN-SL in all other indicators. Moreover, the F1 score, the precision, the recall, the specificity, and the accuracy of RSN-PL are 80.3 %, 80.3 %, 80.2 %, 80.5 % and 82.9 % respectively, which clearly showed that the performance of RSN-PL is not noticeable improved compared to RSN. Obviously, we use ResNet to extract high-dimensional features in the frequency domain of the heart sound as CNN possesses a more reasonable feature representation capability. Meanwhile the LSTM takes the frequency domain features as input and extracts the heart sound time domain features. We combine ResNet with LSTM to effectively improve the heart sound recognition accuracy. Then, we discuss the effect of DCT on performance, with both MFCCs-26 and MFCCs-13 showing a decrease in performance relative to MFSCs, with MFCCs-26 showing a more noticeable decrease in performance. It is clear that the model does not to address the linear transformation in the DCT. In the theory, the deep ResNet can effectively deal with the effects of linear transformation, but the experimental results show the performance indicators of the model are dropped except for specifically, i.e. the model does not efficiently cope with the DCT operations.

V. Conclusion

In this study, we develope a RSN-SL model based on ResNet for the classification of individual heart sound cycles. We compare the model performance of three different networks, the experimental results show the proposed RSN-SL model can effectively accomplish the classification of heart sound signals. And the accuracy of RSN-SL improved by 4.9 % compared to the classical residual network. Furethmore, we evaluate the performance of the RSN-SL model by extracting four different features using the PhysioNet/CinC Challenge 2016 dataset. The experiments show the MFSCs features can achieve better classification results in the RSN-SL model. In summary, deep learning has considerable potential for applications in intelligent healthcare. In future research, we will consider designing end-to-end models that allow for more efficient heart sound recognition.

References

[1] W.H.O.News, "Who reveals leading causes of death and disability worldwide:2000-2019," Geneva,Switzerland.Accessed:Dec.9,2020, [Online].Available:https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019.

[2] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[3] J. Wu, Y. Ji, L. Zhao, M. Ji, Z. Ye, and S. Li, "A mass spectrometric analysis method based on ppca and svm for early detection of ovarian cancer," *Computational and Mathematical Methods in Medicine*, vol. 2016, 2016.

[4] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of Microbiological Methods*, vol. 43, no. 1, pp. 3–31, 2000.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[6] S. El Hihi and Y. Bengio, "Hierarchical recurrent neural networks for long-term dependencies," in *Advances in Neural Information Processing Systems*, 1996, pp. 493–499.

[7] M. A. Goda and P. Hajas, "Morphological determination of pathological pcg signals by time and frequency domain analysis," in *Proc. CinC, Vancouver, Canada*. IEEE, 2016, pp. 1133–1136.

[8] F. Safara, S. Doraisamy, A. Azman, A. Jantan, and A. R. A. Ramaiah, "Multi-level basis selection of wavelet packet decomposition tree for heart sound classification," *Computers in Biology and Medicine*, vol. 43, no. 10, pp. 1407–1414, 2013.

[9] I. Grzegorczyk, M. Soliński, M. Łepek, A. Perka, J. Rosiński, J. Rymko, K. Stępień, and J. Gierałtowski, "PCG classification using a neural network approach," in *Proc. CinC, Vancouver, Canada*. IEEE, 2016, pp. 1129–1132.

[10] A. I. Humayun, S. Ghaffarzadegan, M. I. Ansari, Z. Feng, and T. Hasan, "Towards domain invariant heart sound abnormality detection using learnable filterbanks," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2189–2198, 2020.

[11] Y. Zhang, S. Ayyar, L.-H. Chen, and E. J. Li, "Segmental convolutional neural networks for detection of cardiac abnormality with noisy heart sound recordings," *ArXiv Preprint ArXiv:1612.01943*, 2016.

[12] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, "Heart sound classification based on improved mfcc features and convolutional recurrent neural networks," *Neural Networks*, vol. 130, pp. 22–32, 2020.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR, Las Vegas, USA*, 2016, pp. 770–778.

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

[15] T. Koike, K. Qian, Q. Kong, M. D. Plumbley, B. W. Schuller, and Y. Yamamoto, "Audio for audio is better? An investigation on transfer learning models for heart sound classification," in *Proc. EMBC, Montreal, Canada*, 2020, pp. 74–77.

[16] K. M. Gaikwad and M. S. Chavan, "Removal of high frequency noise from ecg signal using digital iir butterworth filter," in *Proc. GCWCN, India*, 2014, pp. 121–124.

[17] 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 Measurement*, vol. 31, no. 4, p. 513, 2010.

[18] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan, "Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients," in *Proc. CinC, Vancouver, Canada*. IEEE, 2016, pp. 813–816.

[19] F. Binte Azam, I. Ansari, I. Mclane, T. Hasan *et al.*, "Heart sound classification considering additive noise and convolutional distortion," *arXiv e-prints*, pp. arXiv–2106, 2021.