Cutting Weights of Deep Learning Models for Heart Sound Classification: Introducing a Knowledge Distillation Approach

Zikai Song[1,2], Lixian Zhu[1,2], Yiyan Wang[1,2], Mengkai Sun[1,2], Kun Qian[*1,2], *Senior Member, IEEE*, Bin Hu[*1,2], *Fellow, IEEE*, Yoshiharu Yamamoto[3], *Member, IEEE*, and Björn W. Schuller[4,5], *Fellow, IEEE*

***Abstract*— Cardiovascular diseases (CVDs) are the number one cause of death worldwide. In recent years, intelligent auxiliary diagnosis of CVDs based on computer audition has become a popular research field, and intelligent diagnosis technology is increasingly mature. Neural networks used to monitor CVDs are becoming more complex, requiring more computing power and memory, and are difficult to deploy in wearable devices. This paper proposes a lightweight model for classifying heart sounds based on knowledge distillation, which can be deployed in wearable devices to monitor the heart sounds of wearers. The network model is designed based on Convolutional Neural Networks (CNNs). Model performance is evaluated by extracting Mel Frequency Cepstral Coefficients (MFCCs) features from the PhysioNet/CinC Challenge 2016 dataset. The experimental results show that knowledge distillation can improve a lightweight network's accuracy, and our model performs well on the test set. Especially, when the knowledge distillation temperature is 7 and the weight α is 0.1, the accuracy is 88.5 %, the recall is 83.8 %, and the specificity is 93.6 %.**

***Clinical relevance*— A lightweight model of heart sound classification based on knowledge distillation can be deployed on various hardware devices for timely monitoring and feedback of the physical condition of patients with CVDs for timely provision of medical advice. When the model is deployed on the medical instruments of the hospital, the condition of severe and hospitalised patients can be timely fed back and clinical treatment advice can be provided to the clinicians.**

I. INTRODUCTION

The annual death caused by cardiovascular diseases (CVDs) accounts for 45 % of all deaths in Europe [1]. Many of the sounds produced by the human body directly reflect information about our physiological and pathological status. In the case of CVDs, the initial diagnosis can be made by auscultation of heart sounds. However, the frequency and intensity of heart sounds are close to the lower limit of human hearing, and it takes a long period of professional training and clinical experience for clinicians to master auscultation skills [2]. Moreover, audio data and related computer audition (CA) technologies are non-invasive and ubiquitous [3]. Therefore, over the past decade, a growing number of researchers have focused on the use of computer-assisted heart sound analysis to assist physicians and patients in the diagnosis of CVDs.

Nowadays, there are many deep learning algorithms available for the classification of heartbeats. More prominently, Ren and Qian et al. proposed an attention-based deep representation learning method for heart sound classification [4]. Humayun et al. designed a Convolutional Neural Network (CNN) model that uses time-convolution (tCONV) units to simulate a finite impulse response filter to identify heart sound [5]. Ren et al. used a pre-trained CNN from large-scale image data for the classification of Phonocardiogram (PCG) signals by learning deep PCG representations [6]. Deng et al. applied improved Mel-scale Frequency Cepstral Coefficients (MFCCs) features combined with a Recurrent Neural Network (RNN) [7]. Due to the variety of CVDs and the complex acoustic environment of the heart [7], the models used for heart sound recognition are becoming increasingly large for accuracy improvement, which makes it difficult to deploy the models to wearable devices with low arithmetic power and small memory. To address this issue, we use a Knowledge distillation (KD) approach to generate a lightweight heart sound classification model.

Knowledge distillation aims to transfer knowledge from one large neural network (the teacher network) to another smaller neural network (the student network), which could compress the neural network model. In recent years, knowledge distillation has been widely used in different fields of artificial intelligence, including computer vision, natural language processing, speech recognition, and data privacy [8]. Three distillation schemes have been proposed, which are divided into three types according to whether the teacher model is updated during training, namely: offline distillation [9], online distillation [10], and self-distillation [11]. Offline distillation is mainly divided into two stages. First, the teacher model is trained on the training set separately, and then, the teacher model is used to extract logits to guide the

This work was partially supported by the Ministry of Science and Technology of the People's Republic of China with the STI2030-Major Projects (No. 2021ZD0201900), the National Natural Science Foundation of China (No. 62227807 and 62272044), the Teli Young Fellow Program from the Beijing Institute of Technology, China, the BIT Research and Innovation Promoting Project (Grant No. 2022YCXZ012), China, and the Grants-in-Aid for Scientific Research (No. 20H00569) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. *(Zikai Song and Lixian Zhu contributed equally to this work. Corresponding authors: K. Qian, and B. Hu.)*

[1,2]Kun Qian, Bin Hu, Zikai Song, Lixian Zhu, Yiyan Wang, Mengkai Sun are with Key Laboratory of Brain Health Intelligent Evaluation and Intervention, Ministry of Education (Beijing Institute of Technology), Beijing 100081, China, and also with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China. {songzk, zhulx17, yiyanwang, smk, qian, bh}@bit.edu.cn

[3]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

[4,5]Björn W. Schuller is with GLAM – the Group on Language, Audio, & Music, Imperial College London, 180 Queen's 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
DETAILS OF THE DATASET.

Subset	Abnormal	Normal	Total
Training-a	292	117	409
Training-b	104	386	490
Training-c	24	7	31
Training-d	28	27	55
Training-e	183	1 958	2 141
Training-f	34	80	114
Total	665	2 575	3 240

training of the student model. Offline distillation basically does not pay attention to the setting and updating of the teacher model in the first step, but focuses on improving the knowledge transfer part. However, in online distillation, the teacher model and student model are updated at the same time, and the whole knowledge distillation framework is end-to-end trainable [8]. Self-distillation uses the same network for both the teacher and student models; it could be seen as a special case of online distillation. Offline distillation is a one-way knowledge transfer that can be conducted using a large model that has been trained and performs well. The teacher network does not need to be updated with parameters [6], so here, we decided for offline distillation.

In this paper, we propose a lightweight heart sound classification model based on knowledge distillation. The main contribution of our work can be summarised as: (i) We use offline KD transfer knowledge from the teacher model to the student model; (ii) We evaluate its performance and robustness and compare the different models in terms of accuracy, Floating Point Operations (FLOPs), parameter count, and F1 score.

II. Materials and Methods

A. Dataset

The dataset used in this study is the Physionet/Cinc Challenge 2016 [12] data. Table I shows the details of this dataset. The dataset has six subsets recorded by six study groups in clinical and non-clinical settings using different devices, with recording times ranging from seconds to minutes [13]. The dataset contains 3 240 heart sounds from 84 426 heartbeats; among them, 2 575 data are from the normal case and 665 are from the abnormal one.

B. Preprocessing

Heart sound data are affected by sampling rate, noise, and other factors, which can make the accuracy of heart sound classification decrease, so it is exceedingly significant to preprocess the heart sound data. In this study, by standardising and preprocessing heart sounds, we remove noise such as lung sounds and breathing sounds, and external ambient noise. Heart sound preprocessing is divided into two steps. Firstly, we resample the signal with sampling rate fixed to 2 000Hz, and noise is removed using a third-order Butterworth bandpass filter with cutoff frequencies of 20 and 400 Hz [14]. Second, we applied the approach proposed by Schmidt et al. [15] to eliminate spikes.

C. Segmentation

To expand the dataset, we use Springer's improved algorithm of the Hidden Semi-Markov Model (HSMM) [16] for heart sound signal segmentation. We set the length of the heart sound signal segmentation as 2.5 seconds, which is the longest heartbeat cycle length in the dataset [17]. For the data with a heartbeat cycle less than 2.5 seconds, we pad with zeros. By this, we minimise the impact of non-uniform heartbeat cycles on the classification results.

D. Feature Extraction

For speech-recognition, the most commonly used speech feature is MFCCs. This feature has superior performance in acoustics [18]. In this experiment, we use MFCCs-13 features as input to construct the proposed model. MFCCs-13 is MFCCs with 13 coefficients after Discrete Cosine Transform (DCT) compression.

E. Lightweight Model

1) Knowledge Distillation: To compress the model size, knowledge distillation is employed. First, the 'teacher' network uses the training set for individual training, and then, in the environment of 'temperature' *T*, the knowledge of the teacher network is distilled into the student network.

In this section, the high-temperature distillation scheme is described in detail. The teacher network can produce class probabilities by using a softmax function, which are later used to calculate loss when training the 'student' network. However, in many cases, the probability of the correct category in this result is a very high value, while others are very close to zero, so that it could not provide a great deal of information for the student network. To solve this issue, Hinton et al. [9] introduced the concept of the "softmax temperature". As (1) shows:

$$q_i = \frac{exp(\frac{z_i}{T})}{\sum_j exp(\frac{z_j}{T})} \tag{1}$$

where q_i means the probability of class *i*, z_i denotes the logits of the model used, and *T* means distillation temperature. When *T* is set to 1, we get a normal softmax function. When we use a higher value for *T*, the probability distribution is softer than that generated by the softmax function.

As shown in Fig 1, soft labels are the probability distributions obtained by using the softmax function with temperature *T* for the predictions of the teacher model. Soft predictions represent the softmax distributions predicted by the student model with the same temperature *T*. Distillation loss (L_{soft}) indicates the cross entropy between the soft labels and soft predictions. Hard predictions denote the class probabilities predicted by the student model at the temperature of 1. Hard labels mean the value of the actual labels. Student loss (L_{hard}) indicates the cross entropy between hard labels and hard predictions. The objective function ($L(x, W)$) of knowledge distillation is determined by the distillation loss and the student loss together. As (2) shows:

$$L(x, W) = \alpha L_{hard} + \beta L_{soft} \tag{2}$$

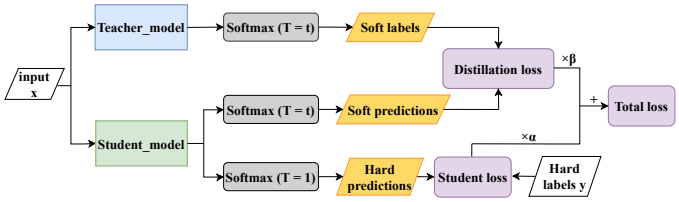

Fig. 1. Method for calculating the objective function of the knowledge distillation.

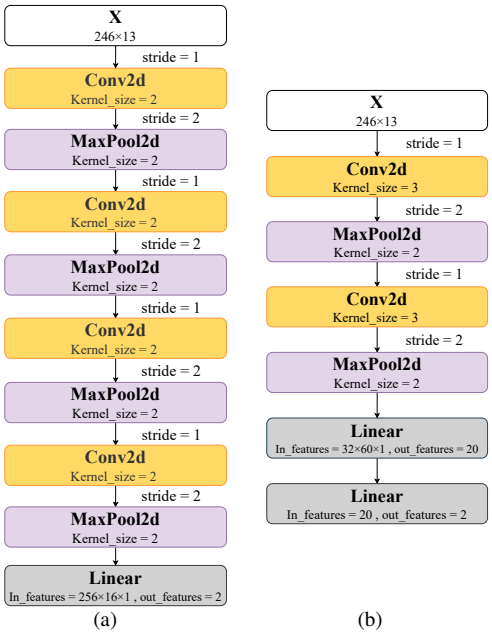

Fig. 2. Network structure of (a) the teacher network; (b) the student network.

By using a weighted average of two loss functions, i. e., $\beta = 1 - \alpha$, we obtain the final objective function.

2) Convolutional neural network: In this section, the teacher network and the student network are introduced. The teacher-student network architecture is shown in Fig 2. Both our teacher and student networks use a CNN. The teacher model consists of four convolutional layers and one linear layer. Each convolutional layer is followed by a ReLU function and a pooling layer. The student network is similar in structure to the teacher network, except that it consists of two convolutional layers and two linear layers. We use the student network to build our lightweight model.

III. EXPERIMENTAL RESULTS

A. Setup

We adopt Pytorch (version-1.7.1) to build our experimental environment, with i7-8750H CPU and GTX1050TI GPU. By segmenting the dataset, we obtain 37 631 data points. We divide the dataset into a training set and a test set by 10:1, each set containing samples of "a-f" subset labels.

In the experiment, we first train the teacher model and the student model separately, and record the training results. In [9], the authors set T ranging from 1 to 20, but it observe better results when T is between 2.5 and 4 and the value of α is low. Therefore, we set a series of temperature T values ranging from 1 to 9 and weight α from 0.1 to 0.7 for the experiments, and compare the results with the student model and the previous models. The evaluation indexes are recall, precision, accuracy, F1 score, and specificity, which are defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

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

$$Accurary = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

TP represents positive samples predicted by the model to be the positive class, TN represents negative samples predicted by the model to be the negative class, FP represents negative samples predicted by the model to be the positive class, FN represents positive samples predicted by the model to be the negative class.

B. Results

Table II shows classification performance of the teacher model, student model and KD model obtained under different hyperparameters. In the table, the teacher model is obtained by training the teacher network alone, and the same holds for the student model. The KD model represents the model obtained by knowledge distillation, T represents the distillation temperature, and α represents the weight of L_{hard}. The experimental results show that the performance of the KD model is the best (except for specificity) when choosing $T = 7, \alpha = 0.1$ with accuracy, recall, precision and F1 score of 88.5 % ($p < .001$ by one-tailed z-test), 83.8 %, 85.4 %, 84.5 %, respectively.

Table III shows the performance comparison between the previous models and the proposed KD model. The FLOPs of the KD model ($T = 7, \alpha = 0.1$) is 66.7 M and the parameter count is 43.3 K, being remarkably low values. The accuracy of 88.5 % is better than the previous models and the student model.

IV. DISCUSSION

As can be seen from Table II, The recall, precision, F1 score, accuracy, and specificity of the student model obtained from the student network trained separately are 79.4 %, 83.1 %, 81.0 %, 86.3 %, and 93.5 %, respectively. The performance of the KD model is often better than that of the student model. When $T = 7$ and $\alpha = 0.1$, all indexes are higher than those of the student model. The accuracy of the KD model ($T = 7$, $\alpha = 0.1$) is lower than that of the teacher model, but this appears acceptable as the KD model requires considerably less computing power and parameters, which is a valid trade-off in terms of required performance and model complexity.

As Table III shows, we reproduce the classical lightweight models, MobileNetV3-Small and MobileNetV3-Large [19],

TABLE II

PERFORMANCE OF DIFFERENT CLASSIFIERS ON THE TEST SETS. [%]

	Teacher model	**Student model**	**KD model** $T=1$ $\alpha=0.3$	**KD model** $T=3$ $\alpha=0.3$	**KD model** $T=5$ $\alpha=0.3$	**KD model** $T=7$ $\alpha=0.1$	**KD model** $T=7$ $\alpha=0.3$	**KD model** $T=7$ $\alpha=0.5$	**KD model** $T=7$ $\alpha=0.7$	**KD model** $T=9$ $\alpha=0.3$
Accuracy	90.5	86.3	87.4	87.2	87.0	**88.5**	87.6	86.4	85.5	87.0
Recall	86.6	79.4	82.6	80.5	81.2	**83.8**	83.4	81.2	78.5	81.7
Precision	87.9	83.1	83.9	84.5	83.0	**85.4**	83.9	82.5	81.9	83.4
F1 score	87.2	81.0	83.4	82.2	82.0	**84.5**	83.6	81.8	80.0	82.5
Specificity	94.3	93.5	94.2	94.2	92.5	93.6	92.1	91.8	90.8	93.5

TABLE III

COMPARISON BETWEEN PREVIOUS WORKS AND THE MODEL INVOLVED IN THE EXPERIMENT.

Model	**FLOPs**	**Params**	**Acc** [%]	**F1** [%]
MobileNetV3-Small	18.3 M	1.0 M	86.9	82.2
MobileNetV3-Large	71.0 M	2.6 M	88.5	84.7
CNN+FocalLoss [20]	-	4.3 K	85.5	-
CardioXNet [21]	-	0.7 M	86.6	88.0
Teacher_Model	1 009.8 M	254.6 K	90.5	87.2
Student_Model	66.7 M	43.3 K	86.3	81.0
KD model ($T=7, \alpha=0.1$)	**66.7 M**	**43.3 K**	**88.5**	**84.5**

for comparison with our model. The FLOPs and parameter count of the KD model ($T = 7$, $\alpha = 0.1$) are 4.3 M and 2.56 M less than those of MobileNetV3-Large, respectively. The parameter count of KD model is 99.7 K and 656.7 K less than that of MobileNetV3-Small and model in [21], respectively. However, the parameter count of KD model is 39.0 K higher than that of the model in [20]. For the classification results of heart sounds, the accuracy and F1 score of our model are 1.6 % and 2.3 % higher than those of MobileNetV3-Small. Nevertheless, these metrics are about equal to MobileNetV3-Large's. We cannot directly compare the accuracy and F1 score of the model in [20] and [21] with those of our KD model, due to the factor that the data preprocessing and partitioning share different strategies.

V. CONCLUSION

We proposed and implemented a lightweight model based on knowledge distillation for heart sound classification. This model performed well compared to both the student model and the previous models. The computing power and number of parameters required for the model were small. This model reached a trade-off between the performance and the model complexity, which made it suitable for deployment on mobile terminals.

REFERENCES

[1] E. Wilkins, L. Wilson, K. Wickramasinghe et al., "European Cardiovascular Disease Statistics 2017", *European Heart Network*, 2017.

[2] Y. Tan, Z. Wang, K. Qian, et al., "Heart Sound Classification based on Fractional Fourier Transformation Entropy". In *Proc. LifeTech, Osaka, Japan*. IEEE, 2022, pp. 588-589.

[3] K. Qian, Z. Zhang, Y. Yamamoto and B. W. Schuller, "Artificial Intelligence Internet of Things for the Elderly: From Assisted Living to Health-Care Monitoring", *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 78-88, 2021.

[4] R. Zhao, K. Qian, F. Dong et al. "Deep attention-based neural networks for explainable heart sound classification", *Machine Learning with Applications*, vol. 9, no. 100322, pp. 1-9,2022.

[5] A. I. Humayun, S. Ghaffarzadegan, M. I. Ansari et al., "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.

[6] Z. Ren, N. Cummins, V. Pandit et al., "Learning image-based representations for heart sound classification". In *Proc. ICDH, New York, USA*. ACM, 2018, pp.143-147.

[7] M. Deng, T. Meng, J. Cao et al., "Heart sound classification based on improved mfcc features and convolutional recurrent neural networks", *Neural Networks*, vol. 130, pp. 2232, 2020.

[8] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey", *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789-1819, 2021.

[9] G. Hinton, O. Vinyals and J. Dean, "Distilling the knowledge in a neural network". In *Proc. NIPS, Montreal, QC, Canada*, 2014.

[10] D. Chen, J. P. Mei, C. Wang, Y. Feng and C. Chen, "Online knowledge distillation with diverse peers". In *Proc. AAAI, New York, USA*. IEEE, 2020, vol. 34, no. 4, pp. 3430-3437.

[11] L. Zhang, J. Song, A. Gao et al., "Be your own teacher: Improve the performance of convolutional neural networks via self distillation". In *Proc. ICCV, Seoul, Korea (South)*. IEEE, 2019, pp. 3713-3722.

[12] C. Liu, D. Springer, Q. Li et al., "An open access database for the evaluation of heart sound algorithms", *Physiological Measurement*, vol. 37, no. 12, p. 2181, 2016.

[13] T. Koike, K. Qian, Q. Kong et al., "Audio for audio is better? An investigation on transfer learning models for heart sound classification". In *Proc. EMBC, Montreal, QC, Canada*. IEEE, 2020, pp. 74–77.

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

[15] 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, pp. 513, 2010.

[16] J. Rubin, R. Abreu, A. Ganguli et al., "Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients". In *Proc. CinC, Vancouver, BC, Canada*. IEEE, 2016, pp. 813–816.

[17] L. Zhu, K. Qian, Z. Wang et al., "Heart Sound Classification based on Residual Shrinkage Networks". In *Proc. EMBC, Glasgow, Scotland, UK*. IEEE, 2022, pp. 4469-4472.

[18] J. Rubin, R. Abreu, A. Ganguli et al., "Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients". In *Proc. CinC, Vancouver, BC, Canada*. IEEE, 2016, pp. 813–816.

[19] A. Howard, M. Sandler, G. Chu et al., "Searching for mobilenetv3". In *Proc. ICCV, Seoul, Korea*. IEEE/CVF 2019, pp. 1314-1324.

[20] T.Li, Y. Yin, K. Ma et al., "Lightweight End-to-End Neural Network Model for Automatic Heart Sound Classification", *Information*, 2021, vol. 12, no. 2, pp. 54.

[21] S. B. Shuvo, S. N. Ali, S. I. Swapnil et al., "CardioXNet: A Novel Lightweight Deep Learning Framework for Cardiovascular Disease Classification Using Heart Sound Recordings", *IEEE Access*, vol. 9, pp. 36955-36967, 2021.