AMNet: Introducing an Adaptive Mel-spectrogram End-to-End Neural Network for Heart Sound Classification

Yang Tan[1,2†], Zhihua Wang[3†], Kun Qian[1,2*], Zhihao Bao[1,2], Zheyu Cao[1,2],
Bin Hu[1,2*], Yoshiharu Yamamoto[3] and Björn W. Schuller[4,5]
1. Key Laboratory of Brain Health Intelligent Evaluation and Intervention, Ministry of Education, Beijing Institute of Technology, Beijing, China
2. School of Medical Technology, Beijing Institute of Technology, Beijing, China
3. Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo, Japan
4. GLAM – Group on Language, Audio, & Music, Imperial College London, London, UK
5. Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany
Corresponding author: {qian, bh}@bit.edu.cn

***Abstract*—The cardiovascular diseases (CVDs) cause tremendous deaths yearly. The Mel-spectrogram is widely used as a tool to analyse the heart sound, which facilitate a cheap and efficient diagnosis of CVDs. Nevertheless, the amplitude and frequency responses of the Mel filter banks remain constant, limiting its function to frequency selection. We propose an adaptive Mel-spectrogram end-to-end neural network (AMNet) for a better characterisation and classification of heart sound in the work. The core of the adaptive Mel-spectrograms (AMel) lies in an adaptive Mel filter banks whose frequency characteristics remain the same as the original Mel-spectrogram (OMel) and amplitude is learnt by the backropagation algorithm. The AMNet learns the raw audio representation directly and outputs the classification results. It reaches 43.5 % Unweighted Average Recall (UAR) and surpasses the model with the OMel and the baseline by 6 % UAR. It is demonstrated that the AMel characterises the heart sound more effectively.**

***Index Terms*—Adaptive Mel-spectrogram, Computer Audition, Heart Sounds, End-to-End, mHealth**

I. Introduction

Cardiovascular diseases (CVDs) are considered the number one killer of human life. According to the World Health Organisation (WHO), 17.9 million people died due to CVDs in 2019, accounting for 32 % of all deaths worldwide [1]. At present, magnetic resonance images and electronic computed tomography are the main methods for early diagnosis of cardiovascular disease. Yet, this is cumbersome, not portable, and costly for the patients. Heart sounds auscultation has been increasingly used by experts due to its advantages of rapidity, low cost and high efficiency. However, the diagnosis is not so reliable. On the one hand, the effective auscultation through heart sounds depends on the knowledge level and rich experience of experts as well as professional and sophisticated equipment [2], [3]. On the other hand, the characteristics of heart sound with low intensity and dominant frequency close to the lower limit of human hearing make accurate diagnosis more challenging for experts [4]–[6]. In addition, in order to avoid the effects of environment noise, doctors need a quiet space in which to conduct their auscultation sessions [7], [8]. And they also need to have the ability to filter the sounds from other organs around the heart of the patients, such as lung and breath sound.

The rapidly developing computer audition (CA) [9] technology is becoming a popular topic of digital medicine research in the search for new digital phenotypes [10]. Heart sounds, as an inexpensive, easy-to-collect, universal and noninvasive audio signal, have been demonstrated to be effective in the classification of cardiac abnormalities. Most previous work have been implemented in two stages, i. e., feature extraction and classification [11], [12]. In recent years, it has become increasingly popular to use lightweight end-to-end networks to learn the features of audio, which perform well [13], [14]. The end-to-end network is considered to be heuristic and capable of learning the characteristics of a sample autonomously [15], [16], just as the human ear acquires audio information. The Experiments about human auditory perception have shown that the human cochlea is equivalent to filter banks. According to this mechanism, the Mel filter banks (usually triangular filters) are used to simulate the perception of different frequencies of audio by the human ear [17]. The amplitude spectrum of the audio signal is filtered by a set of Mel filters to obtain the Mel-spectrogram of each frequency band which is mathematically the dot product of the amplitude spectrum and the frequency response of the Mel filter banks. Since the amplitude-frequency response of the Mel filter banks is con-

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 No 62272044), the National High-Level Young Talent Project, the Teli Young Fellow Program from the Beijing Institute of Technology, China, the JSPS KAKENHI (No. 20H00569), Japan, the JST Mirai Program (No. 21473074), Japan, the JST MOONSHOT Program (No. JPMJMS229B), Japan and the Japan and the China Scholarship Council (No. 202106420019), China.

Yang Tan and Zhihua Wang contributed equally to this work.

stant, it simply simulates the non-linear nature of frequencies in the human ear's hearing and does not have the ability to learn the amplitude spectrum from audio. To this end, we aim to construct an end-to-end model with an adaptive Mel-spectrogram (AMel) which is capable of learning heart sounds. It is able to keep the frequency characteristics of the Mel filter banks while adjusting the amplitude characteristics to the adaptive network to provide better discrimination over the heart sounds events. Two contributions are as follows.

- We design the AMel whose Mel filter amplitude is adaptively adjusted.
- We implement the AMel in an end-to-end network to simulate the human ear to learn heart sounds better.

The remainder of this paper is organised as follows: First, some related works are listed in Sec. II. Then, the database and method are presented in Sec. III. Subsequently, the results and discussion about our work are given in Sec. IV. Finally, we draw a conclusion for our work in Sec. V.

II. Related Works

There have been several innovative and high-performance approaches for the task of heart sound classification using Mel-spetrogram. Haq *et al.* [18] used the Mel-spectrogram as the feature map and achieved a more competitive result on a imbalanced dataset. Bae *et al.* [19] fine-turned the Inception V3 to extract the features from Mel-spectrogram and used an artificial neural network to distinguish systolic murmur from normal heart sounds. Yildirim *et al.* [20] developed hybrid model based on the Mel-spectrogram and achieved the accuracy of 99.63 % on their dataset. It can be seen that Mel-spectrogram could make an encouraging performance on the heart sound classification tasks. However, these Mel-spectrograms are generated by the invariant Mel filter banks. A growing body of researches have reported some works about building pre-filter learning modules into systems [21]–[23]. The method presented in the [21], [22] attempted to approximate or replace Mel filters. In contrast to the essence of these works, our model are improved base on a standard Mel filter bank entirely because of its superiority.

III. Database and Method

A. HSS Database

The database we use is the heart sounds Shenzhen Corpus (HSS) which was released in the INTERSPEECH 2018 ComParE challenge Heart Beats Sub-challenge [24]. The HSS was established and described by the Shenzhen University General Hospital [25]. All the heart sound audio recordings were sampled to 4 000 Hz and included 170 subjects (female: 55, male: 115) with various health conditions. Considering the subject-independence, three datasets, i. e., a training set, development (dev) set, and test set were split by the organisers. Further, three classes, i. e., normal, mild, and mod(erate) / sev(ere) were involved in the HSS, which were annotated by experienced cardiologists through the use of the gold standard. The detail of the data distribution information can be found in Table I.

TABLE I
Data distribution of the HSS database.

Class	Train	Dev	Test	$\sum$
Normal	84	32	28	144
Mild	276	98	91	465
Mod./Sev.	142	50	44	236
$\sum$	502	180	163	845

B. Method

The overall architecture of our proposed adaptive Mel-spectrogram end-to-end neural network (AMNet) is shown in Fig. 1. We present an end-to-end model, in which audio is directly feature-extracted and learnt, then classified. It is worth noting that we construct an adaptive Mel-spectrogram layer (AM layer) as the first layer in the network which is generated by adaptive Mel filter banks (AMFBs).

1) AMNet: In the AMNet architecture, raw heart sounds are fed into the network. Mel-spectrograms are learnt and Convolutional Neural Networks are used to discriminate features. There are six layers in the network, i. e., the AM layer, four convolutional layers, and the last dense layer. We configure 3×3 filters for the four convolutional layers. In order to speed up the convergence and prevent overfitting, each convolutional layer is followed by a batch normalisation layer. Subsequent to the last three convolutional layers activated by "ReLu", the feature maps are dimensionally reduced using a maxpooling layer with pool size of 2×2. During the training, the categorical crossentropy between the CNN output and the true class is minimised using stochastic gradient descent with the Adam optimiser. The AM layer is also learnt by the backpropagation algorithm. Finally, the dense layer with softmax activation function integrates the features and calculates the probability of the heart sound for each of the three classes.

2) Adaptive Mel-spectrogram layer: The core of the proposed AM layer is a set of Mel filter banks with trainable amplitude. It is well known that Mel filters are designed to simulate the hearing of the human ear which focuses only on certain specific frequency domain and not on the entire spectral envelope [26]. To this end, we keep the frequency selection characteristics of the individual Mel filters the same as the original ones and only optimise their amplitudes, which maintains the significance of the Mel filter banks originally used to simulate human hearing in CA.

The short time Fourier transformation (STFT) [27] is often used for processing non-stationary signals. The signal of a heart sound in the time domain is transformed by the STFT, and then we calculate the amplitude spectrum, defined as X.

Next, we design the AMFBs. At first, we obtain a transposed Mel filter banks which contains m filters. Each Mel filter is adjusted in amplitude at the original fixed frequency to suit the CNN model. The amplitude of each Mel filter passband is initialised by the original Mel filter banks accordingly and is defined as a trainable variable that can be adjusted separately

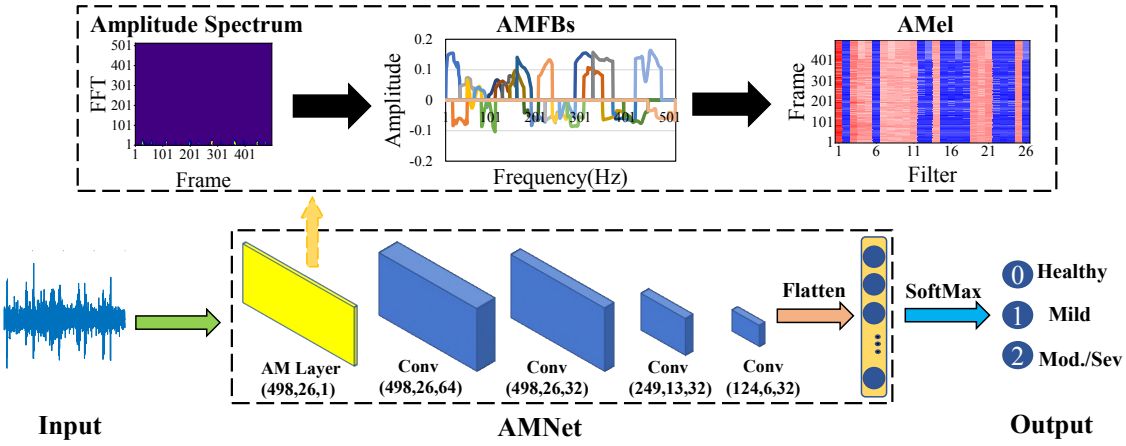


Fig. 1. The proposed AMNet architecture.

by the backpropagation algorithm. The $AMFBs$ is shown in (1).

$$AMFBs = \begin{bmatrix} \mathbf{A_1} & 0 & 0 & \dots & 0 \\ 0 & \mathbf{A_2} & 0 & \dots & 0 \\ 0 & 0 & \mathbf{A_3} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{A_m} \end{bmatrix}, \quad (1)$$

where $\mathbf{A_i}(i = 1 \dots m)$ are the trainable amplitude variables in the passband of each Mel filter, respectively. In this way, the amplitude of the passband portion of a single filter is adjusted without changing its frequency selection function. In the experiments, it is optimised in our end-to-end network.

After that, we apply the $AMFBs$ on the amplitude spectrum X to realise the Mel-spectrogram $AMel$ which is defined as the dot product of X and $AMFBs$ shown in (2).

$$AMel = X \cdot AMFBs. \quad (2)$$

At last, the AMel is transformed to decibel values as the output of the first layer in the network. The AM layer is set as the first layer of the end-to-end network. As the network is back-propagated each time, each filter is adjusted to learn the amplitude spectrum information to obtain a more optimised Mel-spectrogram.

IV. Results and Discussion

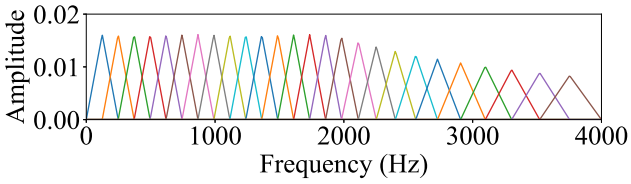
In this section, we implement our method and discuss the results achieved by our experiments.

A. Experiment Setup

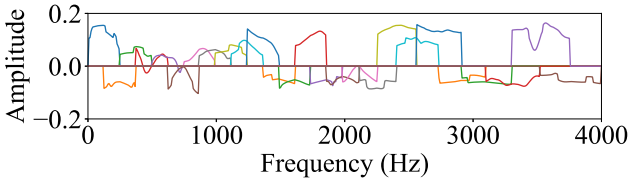
The networks are trained using "Adam" with a learning rate of 0.01 and batch size of 128. Each convolutional layer is maintained with the same size of the feature map, kernel-initialised by "he_normal", and kernel-regularised by "l1" with a weight_decay of 0.0005. In addition, we adopt a window length of 25 ms and hop length of 10 ms for audio framing. The number of the fast Fourier transformation points is 1 024 and the number of Mel filters is 26.

B. Results

In this section, we will show the results of the model with the original Mel-spectrogram (OMel) and the AMel. At first, Fig. 2 shows the original Mel filter banks (OMFBs) and the AMFBs, respectively. We can see that the passband characteristics of the corresponding individual filters of the two sets remain the same. It is worth noting that the amplitude of each of the latter filters is completely adjusted by the backpropagation algorithm. It is a set of adaptive and trainable Mel filter banks.



(a) OMFBs.



(b) AMFBs.

Fig. 2. The two different Mel filter banks. Their frequency selection characteristics are the same, and the amplitude of the latter is trained by the backpropagation algorithm for AMel in the network.

Further, we apply the two different filter banks to the amplitude spectrum to obtain two Mel-spectrograms for subsequent network learning. The output of the first Mel-spectrogram layer and the four two-dimensional convolutional layers generated by the best models of the first mild recording in the test set are shown in Fig. 3. It can be observed that the two Mel-spectrograms are visually distinctly different in Fig. 3 (a) and Fig. 3 (f).

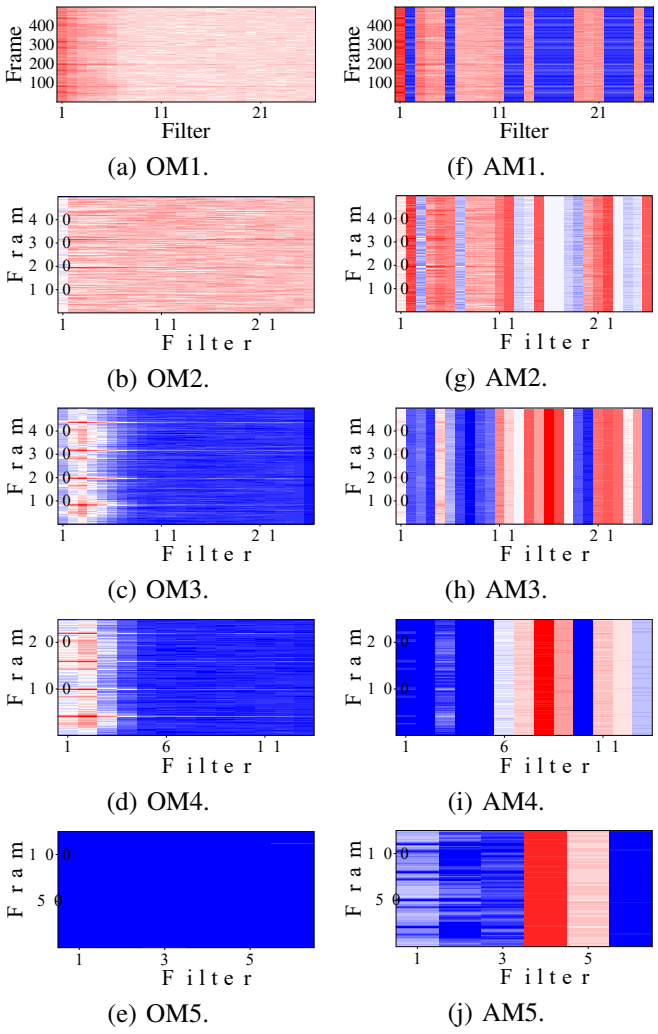


Fig. 3. The output of the first Mel-spectrogram layer and the four two-dimensional convolutional layers generated by the best models of the first mild heart sound recording in the test set. OM*i* (a-e) is the output of the *i*-th layer of the model with OMel. AM*i* (f-j) is the output of the *i*-th layer of the model with AMel. ($i = 1 \ldots 5$).

After fitting the network, the confusion matrices for the best model obtained on the test set for the two Mel-spectrogram are shown in Fig. 4. It should be noted that the parameters of the best model are fitted on the dev set and then applied to the test set. For this imbalanced dataset, it can be seen that the model with the OMel lacks the ability to fit the healthy class of the heart sounds. The model with the ability to train the Mel-spectrogram learns a better representation of the heart sounds and thus has a better ability to fit the three classes of heart sounds than the former. Although our network sacrifices in terms of better recognition of the Mod./Sev., the overall performance has increased.

In order to evaluate our results fairly, we choose the Unweighted Average Recall (UAR) [28] as the primary evaluation metrics and the accuracy as secondary. UAR - the official measure of the ComParE heart sound Challenge, is the average of the Recall for all classes, which is a more comprehensive

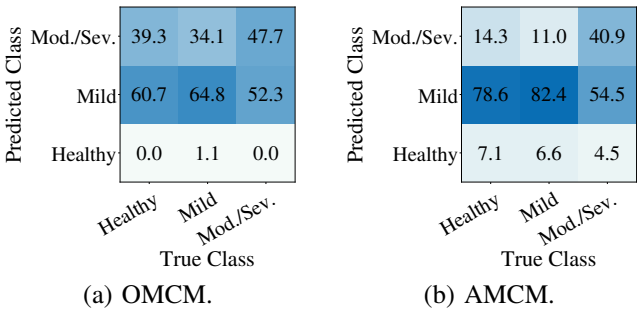


Fig. 4. The two normalised confusion matrixs achieved by the best models. OMCM: The confusion matrix generated by the OMel. AMCM: The confusion matrix generated by the AMel.

TABLE II
CLASSIFICATION RESULTS OF THE PROPOSED AND BASELINE MODELS. OM-MODEL: THE PROPOSED END-TO-END NETWORK WITH THE OMEL. AM-MODEL: THE PROPOSED END-TO-END NETWORK WITH THE AMEL. UAR: UNWEIGHTED AVERAGE RECALL. CHANCE LEVEL: 33.3 %.

Model	**UAR** [%]	**ACC** [%]
Baseline	37.7	\
OM-model	37.5	48.1
AM-model	**43.5**	**58.3**

evaluation index, especially for the imbalanced database we use. The UAR is defined as

$$UAR = \frac{\sum_{i=1}^{N_c} Recall_i}{N_c}, \tag{3}$$

where N_c is the number of classes.

In Table II, we give a result list to compare the best baseline [24] and our models which includes two models with the OMel and the AMel on the test set, respectively. It should be noted that these models are end-to-end structures. From the table, our models perform as well or even better than the baseline. In particular, the proposed model with the AMel shows excellent advantages in characterising heart sound features. It beats the baseline and achieves a relative improvement in the UAR of approximate 6.0 % compared to the other two models. In addition to this, the accuracy obtained by the AMel is 9.8 % higher than the OMel.

C. Discussion

All the blue Mel-spectrograms bands in Fig. 3 (f) are converted to negative Mel coefficient values and have the same value by the action of the Mel filters with negative amplitude (see Fig. 2 (b)), because the amplitude of the Mel filter banks is optimized by the backpropagation gradient algorithm in the end-to-end network. These same Mel coefficients are considered to be redundant. With the deepening of the convolutional network, the selected frequency bands are still distinct and the individual Mel coefficients are still quite different in the network with the AMel (see Fig. 3 (g-j)). However, the individual coefficients are finally convolved into a layer of almost the same colour in the original Mel-spectrogram network (see Fig. 3 (b-e)). This is the direct reason of the difference in the

classification performance of the two models. Therefore, the amplitude spectra of some frequencies are completely filtered due to the action of the AMFBs, while only the essential frequency band targeting the global goal is retained. In general, the proposed Mel filter banks are similar to the brain-inspired hearing mechanism, which can directly learn the amplitude spectrum characteristics of heart sound adaptively. The first layer of the network generates the AMel with a better time-frequency representations for classification of heart sound, which can also contribute to a more interpretable model. Our drawback is that the UAR achieved by the proposed model is limited. With this proposed adaptive Mel-spectrogram approach, combining a more robust and large model to tune the Mel filter banks could be considered in the future.

V. Conclusion

In this work, we presented the AMel which was used in our end-to-end networks. The end-to-end network autonomously and directly learnt the representation of raw heart sounds for classification. To maintain the frequency selection characteristics of Mel filters, we set the amplitudes of the passbands as trainable parameters to obtain the AMel. The paremeters of AMel were learnt and adjusted in the network by the backpropagation algorithm for the purpose of adaptive networking. Experimental results demonstrated that the AMFBs were able to filter out amplitude spectrum at unimportant frequencies and adjusted the effect of individual amplitudes on the classification target. The AMel achieved the UAR of 43.5 % and outperformed the original one by 6 % in UAR. Finally, the AMel approach could be extended beyond heart sound classification, allowing it to effectively adapt to a diverse range of classification tasks and acquire task-specific knowledge. Future efforts could consider further learning of parameters of established expert features from data rather than entirely attempting to learn the representation.

References

[1] WHO, "Cardiovascular diseases (cvds)," 2021, accessed: 2021-06-11. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/

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

[3] A. Zubair and G. Irabor, "Engineering assisted medical training: Development of an auscultation simulator," *J Cardiol Curr Res*, vol. 15, no. 2, pp. 61–66, 2022.

[4] I. Hossain and Z. Moussavi, "An overview of heart-noise reduction of lung sound using wavelet transform based filter," in *Proc. EMBC*, vol. 1. Cancun, Mexico: IEEE, 2003, pp. 458–461.

[5] 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.

[6] M. Altuve, L. Suárez, and J. Ardila, "Fundamental heart sounds analysis using improved complete ensemble emd with adaptive noise," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 426–439, 2020.

[7] H. Wang *et al.*, "Continuous intro-operative respiratory auscultation in anesthesia," in *Proc. SENSORS*, vol. 2. Toronto, ON, Canada: IEEE, 2003, pp. 1002–1005.

[8] F. B. Azam, M. Ansari, I. Mclane, T. Hasan *et al.*, "Heart sound classification considering additive noise and convolutional distortion," *arXiv preprint arXiv:2106.01865*, 2021.

[9] K. Qian, X. Li, H. Li, S. Li, W. Li, Z. Ning, S. Yu, L. Hou, G. Tang, J. Lu *et al.*, "Computer audition for healthcare: Opportunities and challenges," *Frontiers in Digital Health*, vol. 2, p. 5, 2020.

[10] 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.

[11] Y. Tan, Z. Wang, K. Qian, B. Hu, S. Zhao, B. W. Schuller, and Y. Yamamoto, "Heart sound classification based on fractional fourier transformation entropy," in *Proc. LifeTech*, Osaka, Japan, 2022, pp. 588–589.

[12] Z. Wang, Z. Bao, K. Qian, B. Hu, B. W. Schuller, and Y. Yamamoto, "Learning optimal time-frequency representations for heart sound: A comparative study," in *Proc. CSMT*. Hangzhou, China: Springer, 2023, pp. 93–104.

[13] M. Gjoreski, A. Gradišek, B. Budna, M. Gams, and G. Poglajen, "Machine learning and end-to-end deep learning for the detection of chronic heart failure from heart sounds," *IEEE Access*, vol. 8, pp. 20 313–20 324, 2020.

[14] S. B. Shuvo, S. N. Ali, S. I. Swapnil, M. S. Al-Rakhami, and A. Gumaei, "Cardioxnet: A novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings," *IEEE Access*, vol. 9, pp. 36 955–36 967, 2021.

[15] C. Cummins, P. Petoumenos, Z. Wang, and H. Leather, "End-to-end deep learning of optimization heuristics," in *Proc. PACT*, Portland, OR, USA, 2017, pp. 219–232.

[16] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

[17] A. Biswas, P. K. Sahu, and M. Chandra, "Admissible wavelet packet features based on human inner ear frequency response for hindi consonant recognition," *Computers & Electrical Engineering*, vol. 40, no. 4, pp. 1111–1122, 2014.

[18] H. F. D. U. Haq, R. Ismail, S. Ismail, S. R. Purnama, B. Warsito, J. D. Setiawan, and A. Wibowo, "Efficientnet optimization on heartbeats sound classification," in *Proc. ICICoS*. Semarang, Indonesia: IEEE, 2021, pp. 216–221.

[19] J. Bae, M. Kim, and J. S. Lim, "Feature extraction model based on inception v3 to distinguish normal heart sound from systolic murmur," in *Proc. ICTC*. Jeju, Korea (South): IEEE, 2020, pp. 460–463.

[20] M. Yildirim, "Diagnosis of heart diseases using heart sound signals with the developed interpolation, cnn, and relief based model." *Traitement du Signal*, vol. 39, no. 3, 2022.

[21] N. Zeghidour, O. Teboul, F. d. C. Quitry, and M. Tagliasacchi, "Leaf: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[22] H. Seki, K. Yamamoto, and S. Nakagawa, "A deep neural network integrated with filterbank learning for speech recognition," in *Proc. ICASSP*. New Orleans, LA, USA: IEEE, 2017, pp. 5480–5484.

[23] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification," in *Proc. Interspeech*, vol. 8, Stockholm, Sweden, 2017, p. 9.

[24] B. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny *et al.*, "The interspeech 2018 computational paralinguistics challenge: Atypical and self-assessed affect, crying and heart beats," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 122–126.

[25] F. Dong, K. Qian, Z. Ren, A. Baird, X. Li, Z. Dai, B. Dong, F. Metze, Y. Yamamoto, and B. W. Schuller, "Machine listening for heart status monitoring: Introducing and benchmarking hss—the heart sounds shenzhen corpus," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 2082–2092, 2020.

[26] H. Xu, X. Zhang, and L. Jia, "The extraction and simulation of mel frequency cepstrum speech parameters," in *Proc. ICSAI*, Yantai, China, 2012, pp. 1765–1768.

[27] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[28] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 312–315.