CNN-Based Heart Sound Classification with an Imbalance-Compensating Weighted Loss Function

Zishen Li[1], Yi Chang[1] and Björn W. Schuller[1,2]

Abstract— **Heart sound auscultation is an effective method for early-stage diagnosis of heart disease. The application of deep neural networks is gaining increasing attention in automated heart sound classification. This paper proposes deep Convolutional Neural Networks (CNNs) to classify normal/abnormal heart sounds, which takes two-dimensional Mel-scale features as input, including Mel frequency cepstral coefficients (MFCCs) and the Log Mel spectrum. We employ two weighted loss functions during the training to mitigate the class imbalance issue. The model was developed on the public PhysioNet/Computing in Cardiology Challenge (CinC) 2016 heart sound database. On the considered test set, the proposed model with Log Mel spectrum as features achieves an Unweighted Average Recall (UAR) of** 89.6 %**, with sensitivity and specificity being** 89.5 % **and** 89.7 % **respectively.**

Clinical relevance— **This work proposes a CNN-based model to enable automated heart sound classification, which can provide auxiliary assistance for heart auscultation and has the potential to screen for heart pathologies in clinical applications at a relatively low cost.**

I. INTRODUCTION

Cardiovascular diseases remain a leading threat to heart health; thus, accurate heart sound auscultation is in great demand to help diagnose heart diseases. However, due to the lack of experienced clinicians and low patient-to-doctor ratio, accurate heart sound diagnosis is hard to obtain [1]. Nowadays, the rapid developments in computer-aided heart sound analysis provide a potential solution to this problem. Various computational methods with signal processing and machine learning techniques are gaining increasing attention in automatic heart sound diagnosis [2] [3].

In the early works, heart sound classification mainly relied on low-level human-designed features, typically including durations, amplitudes, frequency components, and energy measurements of the heart sound signal [4]. In recent years, deep learning techniques have reported high performance in heart sound classification [5] [6] [7]. Extracted features such as signal energy distribution [5], frequency bands [6], and wavelet features [7] can be input to neural networks, which allows the learning of more complex representations from the primitive features. Particularly, inspired by the remarkable performance of Convolutional Neural Networks (CNNs) in image classification tasks, the application of CNNs in heart sound classification is most frequently discussed in recent studies [8] [9] [10] [11]. CNNs are usually combined with two-dimensional time-frequency representations of the heart sound signal, in which power spectrum, wavelet spectrogram, and Mel-scale features are widely used [8]. Potes et al. combined CNNs and AdaBoost with time-frequency features extracted from nine frequency bands for heart sound classification. They achieved 86.0 % UAR on the randomly sampled test set, which was the first place in the PhysioNet Challenge 2016 [9]. Tschannen et al. proposed a wavelet-based CNN feature extractor to obtain wavelet features and combined the features with an SVM classifier to classify the heart sounds [12]. Nilanon et al. visualised the Power Spectral Density (PSD) features as one channel images and employed a CNN model to classify heart sounds [13]. Mel-scale features have shown high performance in heart sound classification. Rubin et al. visualised the Mel frequency cepstral coefficients (MFCCs) as heatmaps and input the heatmap images to a 2D-CNN model for heart sound classification [10]. Dong's study on Heart Sound Shenzhen database [14] showed that the Log Mel spectrogram achieved better results than other low-level descriptors. Noman et al. explored the utilisation of a 1D-CNN and a 2D-CNN with MFCC features in heart sound classification [11].

Various studies have shown the predictive power of CNNs for heart sound classification. However, only one feature set is studied in most studies [8] [10], and more importantly, the problem of class imbalance has been overlooked in most studies [9] [13] [15]. In this paper, we propose two deep CNNs for heart sound classification with two Mel scale acoustic features considering the class imbalance problem. Specifically, We extract MFCC and the Log Mel spectrum from the heart sound signal as the input feature to the CNN and compare the models' performance associated with these two features. Moreover, to alleviate the data imbalance issue in the heart sound database, we apply two weighted loss functions in the classification stage, which leads to a more accurate model for detecting abnormal heart sounds.

II. METHODS

A. Feature Extraction

Mel-scale features are widely applied in speech signal processing [16]. We extract MFCCs and the Log Mel spectrum as two-dimensional representations of the heart sound signal. Besides, we found that adding additional features as the first-order and second-order derivatives of the MFCCs brings no performance improvement, probably because the linear derivatives can be approximated by the linear layers in the CNN, which is consistent with recent findings [15] [17].

[1]Z. Li, Y. Chang, and B. Schuller are with the Department of Computing, Imperial College London, UK zishen.li20@imperial.ac.uk, y.chang20@imperial.ac.uk, and bjoern.schuller@imperial.ac.uk

[2]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

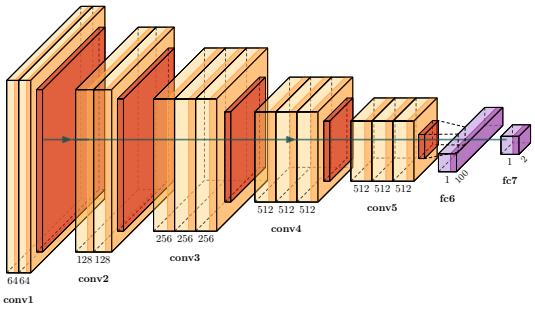

Fig. 1. Illustration of the VGG architecture

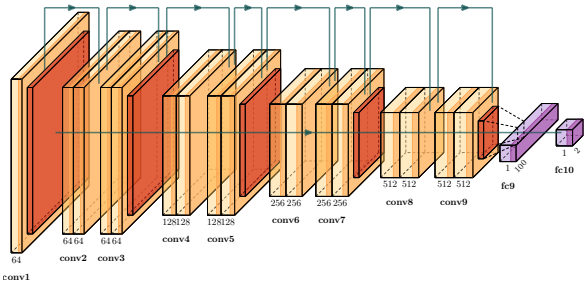

Fig. 2. Illustration of the ResNet architecture

B. Convolutional Neural Network

The CNN is implemented to extract high-level spatial features from the named two-dimensional MFCCs and Log Mel spectrum. We propose two architectures that mimic the building blocks of VGG [18] and a residual neural network (ResNet) [19] to classify the heart sounds, and compare the classification results of these two models.

The network modified from VGG contains the basic block as a stack of 3×3 convolutional kernels, followed by a 2×2 max-pooling layer to reduce the size of the feature map. The convolutional layer is followed by a BatchNorm layer and a ReLu activation function. Fully connected layers are employed at the end to output the classification result.

The ResNet architecture has similar convolutional blocks as the VGG, which contains a stack of 3×3 convolutional kernels, followed by a 2×2 max-pooling layer. The residual network also consists of residual connections of identity mapping before the activation function to add expressiveness. Similar to the VGG architecture, BatchNorm layers and ReLu activation are used, and fully connected layers are connected to perform classification. Figure 1 and figure 2 illustrate the architectures of the proposed models.

C. Loss Function

Cross entropy loss is commonly used in previous studies [10] [13] [15]. However, it ignores the class imbalance problem, which is common in heart sound databases since subjects with heart disease are relatively rare. In this paper, we considered two weighted loss functions, balanced cross-entropy loss and focal loss [20], as the algorithm-level approach to address the imbalanced dataset problem.

The balanced cross-entropy loss assigns larger weights to the minority class, which gives more penalty in computing the loss if the model misclassifies the abnormal heart sounds. However, it does not consider the hard or easily misclassified samples. Particularly, signals that are corrupted by noise are easily misclassified in this task. Focal loss handles this problem by introducing a modulation term. Focal loss is defined as follows:

$$FL(p_t) = -a_t(1-p_t)^{\gamma}\log(p_t), \tag{1}$$

where a_t denotes the class weights and p_t denotes the predicted probability of the sample being in the ground truth class. The class weights term a_t handles the class imbalance problem, similar to that in balanced cross-entropy loss. Additionally, the modulation term $(1-p_t)^{\gamma}$ adds larger weights to the sample if it is misclassified with high probability. In this way, focal loss focuses more on hard-to-classified samples, such as the noisy heart sound recordings in the database.

In this paper, the class weight is set as 0.2 for normal heart sounds and 0.8 for abnormal ones according to the distribution of the classes. And the modulation factor γ is chosen to be 1 after the hyperparameter tuning.

III. Experimental Results

A. Dataset

The heart sound database used in this paper was released by the PhysioNet/Computing in Cardiology challenge in 2016 [21], which aimed to provide the largest heart sound database for the development of algorithms. A total of 3240 heart sound recordings gathered from independent research centres were categorised as normal and abnormal by expert labelling. While 2575 heart sounds collected from healthy subjects were labelled as normal, 665 heart sounds from subjects with confirmed heart disease were labelled as abnormal.

B. Experimental Setup

1) Preprocessing: Since the heart sounds are collected from non-standard environments, preprocessing is applied to normalise the signal. We use a Butterworth filter ($f1 = 25Hz, f2 = 400Hz$) to suppress the noise while preserving the morphology of clean signals. The original sampling rate of the heart sound signal is $2kHz$ and is downsampled to $1kHz$ to reduce the computational cost. The signals are normalised by dividing the maximum value to ensure the amplitudes of the signals are limited to the scale of $[-1,1]$.

2) Segmentation: The original heart sound recordings are segmented into short intervals of cardiac cycles to increase the number of heart sound samples and ensure that all heart sounds are of the same length. We implement Springer's algorithm for heart sound segmentation, which utilises a trained hidden semi-Markov model with logistic regression to identify the states of heart sounds [22]. The heart sound is segmented at the beginning of each cardiac cycle with a length of 3 seconds. A total of 40640 segmented heart sound samples were obtained after segmentation, consisting of 31882 normal samples and 8758 abnormal ones. Each heart sound segment is aligned at the beginning of the cardiac cycle.

TABLE I

SENSITIVITY (SE), SPECIFICITY (SP) AND UAR ON CROSS-VALIDATION

Loss function	Feature	Model	Se / Sp / UAR (%)
Cross-Entropy Loss	Log Mel	VGG	81.2 / 90.1 / 85.7
	Log Mel	ResNet	77.0 / 93.0 / 85.0
	MFCC	VGG	73.5 / 96.1 / 84.8
	MFCC	ResNet	77.6 / 91.7 / 84.7
Balanced Cross-Entropy Loss	Log Mel	VGG	88.7 / 86.8 / 87.8
	Log Mel	ResNet	88.0 / 86.6 / 87.3
	MFCC	VGG	85.7 / 88.5 / 87.1
	MFCC	ResNet	88.7 / 86.6 / 87.7
Focal loss	Log Mel	VGG	88.0 / 91.0 / **89.5**
	Log Mel	ResNet	88.0 / 89.3 / 88.7
	MFCC	VGG	91.0 / 87.6 / 89.3
	MFCC	ResNet	87.5 / 83.2 / 85.4

TABLE II

CLASSIFICATION RESULTS ON HIDDEN TEST SET

Loss function	Feature	Model	Se / Sp / UAR (%)
Focal loss	Log Mel	VGG	89.5 / 89.7 / **89.6**

3) Training setup: Since the official test set of the challenge is not published, for a fair comparison with the related work [8] [11] [13] [15], we keep a 20% test set by stratified sampling for evaluating the model. The 80% training set is then partitioned into a 5-fold cross-validation set to choose the optimal model. The data splitting process is operated on the level of original heart sound recordings rather than segments to ensure that segments from the same recordings cannot occur in both the training and test sets.

The extracted features, Log Mel spectrum and MFCCs, are trained on two network architectures with three loss functions for comparison. In the training stage, the batch size is set to be 64, and a learning rate decay strategy is used to decay the learning rate by 50% every 20 epoch. Early stopping is used to prevent overfitting that ends training if the loss on the validation set does not decrease after 50 epochs.

Since the heart sound recordings are segmented into short frames before the training process, the classification result is associated with segments rather than original recordings. Therefore, we consider a score-level fusion with a majority voting strategy on the segments that belong to the same heart sound to determine the final predictions of the original heart sound recording.

4) Evaluation Metrics: For evaluation of the predicted result, the Unweighted Average Recall (UAR) is utilised in this paper. UAR is defined as the average value of sensitivity and specificity, which takes into account the recall of both positive samples and negative samples; thus, it is more reasonable with a highly imbalanced data distribution [23].

C. Results

The extracted features, Log Mel spectrum and MFCCs, are trained on VGG and the ResNet model, respectively. Table I shows the classification performances of the models on the validation set with three loss functions.

Comparing the results, we notice that the classification performance is comparable between the proposed VGG and ResNet models. The model with Log Mel spectrum tends to perform slightly better than that with MFCCs by less than 1%. Moreover, it is noticeable that there is a high trade-off between sensitivity and specificity with the plain cross-entropy loss, which means the model performs much worse in classifying abnormal heart sounds. By applying the weighted loss function, the models achieve better classification results and considerable improvements in performance when classifying abnormal heart sounds. With balanced cross-entropy loss, the proposed ResNet model achieves a maximum improvement of 3% with the MFCCs features. Furthermore, the models with focal loss outperform those with balanced cross-entropy loss. With the focal loss, the proposed VGG model outperforms the plain cross-entropy by a maximum increase in UAR of 4.5% with MFCCs feature. The results on the cross-validation set show that the proposed VGG model achieves the best performance using the Log Mel spectrum features with the focal loss function. Next, we evaluate the best model on the hidden considered test set. As shown in table II, the model achieves a UAR of 89.6% with relatively balanced sensitivity (89.5%) and specificity (89.7%).

In the PhysioNet/CinC Challenge, Potes et al. achieved the UAR of 86%, which was accepted as the top model [9]. There are models proposed after the challenge that used the same dataset. Bozkurt et al. achieved 81.5% UAR [14] and Zhang et al. achieved 90% UAR [24] on 10-fold cross validation, respectively. Maknickas's model obtained 84.1% UAR [8] and Noman's model achieved 88.2% UAR [13] tested on a 20% held-out test set. Our proposed model achieves the UAR of 89.6% on the 20% held-out test set, which is comparable to current state-of-the-art models.

D. Discussion

We demonstrated that the deep CNN model is effective in learning high-level representations from the acoustic features for heart sound classification. Comparing the VGG and ResNet architectures, the performances are relatively similar. ResNet was proposed to solve the problem of degraded model performance in very deep neural networks and has reported better performance than VGG in classification tasks [19]; however, the performance of the VGG and the ResNet models in this task was comparable. This appears reasonable since the network used in this study did not involve too deep an architecture, therefore, the advantage of the residual network is not evident in this task.

For comparison between each feature, the Log Mel spectrum performs slightly better than MFCCs. The Log Mel spectrum was extracted without the discrete cosine transform (DCT), while MFCCs were obtained after DCT to decorrelate the Mel filter bank coefficients. However, DCT is unnecessary for application in a neural network-context since it is not easily affected by correlated inputs. Moreover, since DCT is a linear transformation, applying DCT will lose non-linearly related information which may carry

important information about heart pathology. This is worth noticing, since most related studies prefer MFCCs as the feature [8] [10] [11] while fewer studies have explored the use of the Log Mel spectrum. In this paper, results show that the Log Mel spectrum leads to a better performance than MFCCs, which is worth exploring in the future.

The comparison between the loss functions is of importance. With plain cross-entropy loss, the model had a high trade-off between sensitivity and specificity, showing that the model performs much worse in detecting abnormal heart sounds. This performance can be troublesome in heart disease detection since false negatives could be dangerous and need more attention. By applying balanced cross-entropy loss and focal loss, the model achieved better overall performance and improved performance in classifying abnormal heart sounds. When comparing these two weighted loss functions, focal loss outperforms the balanced cross-entropy loss, probably because focal loss assigns larger weights to the hard-to-classify samples (e. g., those affected by noise).

IV. CONCLUSIONS

We proposed VGG and ResNet CNN architectures that take two-dimensional Mel-scale features as input for heart sound classification. We compared the performance of Log Mel spectrum and MFCCs features together with three types of loss functions. Two weighted loss functions, balanced cross-entropy loss and focal loss, were implemented to address the problem of imbalanced data distribution. Evaluations on the PhysioNet/CinC dataset have demonstrated improvements in model performance with weighted loss functions, among which focal loss leads to the best classification results. Among all the experiments, the Log Mel spectrum trained on the VGG network with focal loss function presented the best performance by cross-validation, which achieved 89.5% sensitivity, 89.7% specificity, and overall UAR of 89.6% on the considered test set.

Future works could consider some up-sampling and down-sampling methods to further address the data imbalance issue, such as using Generative Adversarial Networks (GANs) for data augmentation. Moreover, end-to-end architectures such as Recurrent Neural Networks (RNNs) can be explored in future works to extract the temporal dependencies from the signal, bypassing the segmentation step.

REFERENCES

[1] G. D. Clifford, C. Liu, B. E. Moody, J. M. Roig, S. E. Schmidt, Q. Li, I. Silva, and R. G. Mark, "Recent advances in heart sound analysis," Physiological measurement, vol. 38, pp. E10–E25, 2017.

[2] S. Ismail, I. Siddiqi, and U. Akram, "Localization and classification of heart beats in phonocardiography signals—a comprehensive review," EURASIP Journal on Advances in Signal Processing, vol. 2018, no. 1, pp. 1–27, 2018.

[3] Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, and B. Schuller, "Learning image-based representations for heart sound classification," in Proc. ICDH, Lyon, France, 2018, p. 143–147.

[4] P. Wang, C. S. Lim, S. Chauhan, J. Y. A. Foo, and V. Anantharaman, "Phonocardiographic signal analysis method using a modified hidden markov model," Annals of Biomedical Engineering, vol. 35, no. 3, pp. 367–374, 2007.

[5] S. R. Bhatikar, C. DeGroff, and R. L. Mahajan, "A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics," Artificial intelligence in medicine, vol. 33, no. 3, pp. 251–260, 2005.

[6] A. A. Sepehri, J. Hancq, T. Dutoit, A. Gharehbaghi, A. Kocharian, and A. Kiani, "Computerized screening of children congenital heart diseases," Computer methods and programs in biomedicine, vol. 92, no. 2, pp. 186–192, 2008.

[7] H. Uğuz, "A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases," Journal of medical systems, vol. 36, no. 1, pp. 61–72, 2012.

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

[9] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy, "Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds," in Proc. CinC, Vancouver, Canada, pp. 621–624.

[10] 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, pp. 813–816.

[11] F. Noman, C.-M. Ting, S.-H. Salleh, and H. Ombao, "Short-segment heart sound classification using an ensemble of deep convolutional neural networks," in Proc. ICASSP, Brighton, UK.

[12] M. Tschannen, T. Kramer, G. Marti, M. Heinzmann, and T. Wiatowski, "Heart sound classification using deep structured features," in Proc. CinC, Vancouver, Canada.

[13] T. Nilanon, J. Yao, J. Hao, S. Purushotham, and Y. Liu, "Normal/abnormal heart sound recordings classification using convolutional neural network," in Proc. CinC, Vancouver, Canada.

[14] 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, 2019.

[15] B. Bozkurt, I. Germanakis, and Y. Stylianou, "A study of time-frequency features for cnn-based automatic heart sound classification for pathology detection," Computers in biology and medicine, vol. 100, pp. 132–143, 2018.

[16] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," Journal of Computer science and Technology, vol. 16, no. 6, pp. 582–589, 2001.

[17] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghami, and C. Fookes, "A robust interpretable deep learning classifier for heart anomaly detection without segmentation," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 6, pp. 2162–2171, 2020.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, Las Vegas, Nevada.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. ICCV, Venice, Italy.

[21] 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.

[22] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-hsmm-based heart sound segmentation," IEEE Transactions on Biomedical Engineering, vol. 63, no. 4, pp. 822–832, 2015.

[23] B. Schuller and A. Batliner, Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing, 1st ed. Wiley Publishing, 2013.

[24] W. Zhang, J. Han, and S. Deng, "Heart sound classification based on scaled spectrogram and tensor decomposition," Expert Systems with Applications, vol. 84, pp. 220–231, 2017.