IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 18, 2020, accepted August 18, 2020, date of publication August 28, 2020, date of current version September 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3020099

A CRNN System for Sound Event Detection Based on Gastrointestinal Sound Dataset Collected by Wearable Auscultation Devices

XUE ZHENG[1], CHUN ZHANG[1], (Member, IEEE), PING CHEN[2], KANG ZHAO[1], HANJUN JIANG[1], (Senior Member, IEEE), ZHIWEI JIANG[3], HUAFENG PAN[3], ZHIHUA WANG[1], (Fellow, IEEE), AND WEN JIA[4]

[1]Department of Microelectronics and Nanoelectronics, Tsinghua University, Beijing 100084, China
[2]Beijing Yiemed Medical Technology Company Ltd., Beijing 100016, China
[3]Department of General Surgery, Affiliated Hospital, Nanjing University of Chinese Medicine, Nanjing 210000, China
[4]Guangdong Engr. Research Center on ICs for Wireless Healthcare, Research Institute, Tsinghua University, Shenzhen 518055, China

Corresponding authors: Hanjun Jiang (jianghanjun@tsinghua.edu.cn) and Zhiwei Jiang (surgery1@aliyun.com)

This work was supported in part by the National Key Research and Development Plan of China under Grant 2016YFC0105603, and in part by the Science, Technology, and Innovation Commission of Shenzhen Municipality Fund under Grant JCYJ20180306170435280 and Grant JCYJ20190807142805490.

ABSTRACT In this article, we set up a novel audio dataset named Gastrointestinal (GI) Sound Set which includes 6 kinds of body sounds Bowel sound, Speech, Snore, Cough, Groan, and Rub. We do sound event detection (SED) based on it, and can accurately detect 6 types of sound events. First, the GI Sound Set is collected by wearable auscultation devices. To ensure generalization, patients from five different hospital departments are recruited for data collection, along with a group of healthy subjects. GI Sound Set refers to Google AudioSet in data format but varies in audio length and sampling rate. Second, we extract Mel-filter features from the recordings and investigate the performance of different activation functions and neural network architectures for detecting sound events. We use data augmentation, class balance to deal with the problem of quantitative imbalance between classes on the dataset. We apply multiple instances learning(MIL) to give out not only bag-level results but also frame-level results. In this work, GI Sound Set is the largest body sound dataset to date, and our approach shows state-of-the-art performance with an average score of F1=81.06% evaluated on the test set. Due to its simple network and conventional processing method, our CRNN system has high universality, which can be used in other audio datasets, such as respiratory sound and heart sound.

INDEX TERMS Gastrointestinal (GI) sound set, sound event detection(SED), convolutional recurrent neural network (CRNN), multiple instance learning(MIL).

I. INTRODUCTION

In recent years, with the development of artificial intelligence technology and wearable medical devices, a lot of AI-assisted diagnoses using medical imaging and electronic medical record data have been proved to be effective in reducing workload for doctors. By contrast, the absence of stereophonic data has prevented the auscultation process from being digitized, with only a small number of datasets collected by electronic stethoscopes for research, such as ICBHI [1], Physionet [2] and the Noisy Guts project [3]. [3] explored the use of bowel sounds to characterize IBS with a view to diagnostic use using a diagnostic case-control study, and Independent testing demonstrated 87% sensitivity and 87% specificity for IBS diagnosis using the 15 IBS and 15 healthy participants. In connection with the work of this article, Our original intention is to mine information about our body and provide helpful information for doctors. As is mentioned in [3], the interpretation of bowel sounds (BS) provides a convenient and non-invasive technique to aid in the diagnosis of gastrointestinal (GI) conditions. However, this approach is limited by the variation between BS and its irregular occurrence. In a few cases, manual auscultation can make judgments, but there is plenty of subjectivity and uncertainties. A longer recording

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Shen.

has the potential to unlock additional understanding of GI physiology and clinical utility. That is the point why we set up our GI Sound Set.

According to [4], the models which have been used to model sound events are classified into four groups: hidden Markov models (HMMs) [7]–[10], a multi-pass decoding procedure [11], non-negative matrix factorization (NMF) [12], convolutional neural networks (CNNs) [13]–[18], recurrent neural network (RNNs) [19]–[21] and their combination convolutional and recurrent neural networks (CRNNs) [22]. As is known, HMMs can not handle polyphonic audio while HMMs with multi-pass decoding figure it out, but can not explain and simulate how overlapping sound events affect the acoustic characteristics of each individual sound event. NMF has the ability to handle overlapping sounds but has no ability to take advantage of the temporal context thus losing information. CRNNs show state-of-the-art performance on sound event detection by combining the ability of CNNs to learn locally invariant filters and the power of RNNs to event detection.

Our main contributions and results are:

1) We set up GI Sound Set with 6 sound events that cover common sound events related to body sounds established by ourselves.
2) We do contrast experiments on Activation Functions, Network Size, Input Features, and try to find the best way to do sound event detection.
3) We explore effective data augmentation methods based on a quantitative imbalance between classes.
4) We show the state-of-the-art performance on GI Sound Set.

The paper is structured as follows: In Section II, we introduce the collection method and processing procedures of GI Sound Set. In Section III, CNNs, BiGRU, CRNN is discussed. In particular, we introduce multiple instance learning and some data augmentation methods. In Section IV, we introduce our network architecture, feature extraction, and experimental results. Finally, we discuss our findings in Section V and conclude the paper in Section VI.

II. GI SOUND SET

To our knowledge, GI Sound Set is the largest dataset about body sounds. We introduce it from four aspects: Collection Instrument, Collection Method, Dataset Annotation, GI Sound Set Distribution, and Medical Significance.

A. COLLECTION INSTRUMENT

A small and portable device named Continuous Auscultation Recorder (Type: YM-TYJL-01) is used to collect body sounds. It is an innovative medical instrument designed by Beijing Yiemed Medical Technology Co. Ltd. It consists of an auscultation patch, Bluetooth receiver, and software running on the computer. Each patch with a high sensitivity sensor and a high-performance processor can collect different types of body sounds for 24 hours by sticking to the subject's abdomen. The wireless audio signal will transmit to the computer via Bluetooth and Ethernet through the receiver inward. As shown in Fig.1, simultaneous remote and continuous auscultation of 100 patients can be achieved [16].

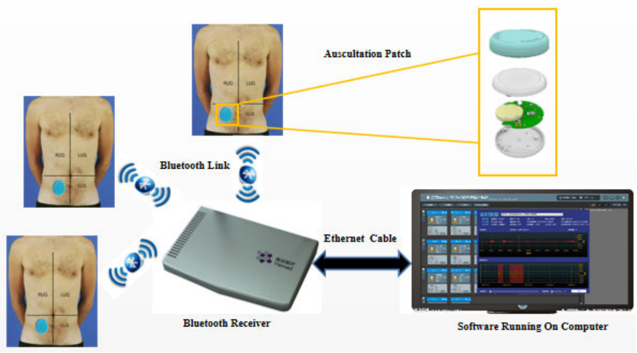


FIGURE 1. Continuous Auscultation Recorder.

B. COLLECTION METHOD

GI Sound Set signals are recorded by attaching Continuous Auscultation Recorder to the right lower Qu (RLQ) of the subject's abdomen, where we are able to acquire the most bowel sound. In order to ensure the generalization of GI Sound Set, we make a careful arrangement and selection from the three aspects: collection sites, subjects, collection time, meanwhile count and record all the subjects' information in detail.

1) COLLECTION SITES

Collection sites contain five departments of four hospitals in different districts and in-home visits.

We select clinical and non-clinical (such as in-home visits) environment for data diversity. These addresses respectively are Jishuitan Hospital Orthopedics Department, Jiangsu Provincial Hospital of Traditional Chinese Medicine General Surgery Department, Shanxi Provincial People's Hospital General Surgery Department, Liaocheng People's Hospital Hepatobiliary Surgery Department, Intensive Care Unit of Liaocheng People's Hospital and Yiemed Medical Technology Co. Ltd which is the only non-clinical environment. All our data collection work has passed the ethical review of the hospital ethics committee.

All recordings are performed in an uncontrolled environment, which is either performed during in-home visits or in the hospital. This results in many recordings being corrupted by various sources of noise, such as white noise. Other noise sources include the friction sound between the patch and the clothes, the current noise, and so on.

2) SUBJECTS

20 subjects are selected at each collection site including 10 males and 10 females, a total of 120 subjects. Their ages range from 20 to 80, with an average age of 55.91; Their BMIs range from 13.96 to 39.13, with an average BMI of 23.70.

TABLE 1. **Subject Type and Number.**

Subjects Type	Disease	Number
The healthy	None	20
Patient	Gastric Cancer	12
	Rectal Cancer	10
	Colon Cancer	11
	Gastric Perforation	1
	Tumor Thrombus of Inferior Vena Cava	1
	Ostomy	5
	Cholecystolithiasis with Cholecystitis	1
	Cholecystectomy	19
	Total Knee Arthroplasty	7
	Total Hip Arthroplasty	13
	Acute Myocardial Infarction	1
	Intestinal Obstruction	1
	Pulmonary Infection	5
	Coronary Heart Disease Respiratory Failure	1
	Right Femoral Hip Impingement Syndrome	2
	Brain Atrophy	4
	Copd	4
	Pancreatic Cancer	2
Total	--	120

Subjects are generally divided into two categories based on their health status: the healthy and patient, and Table 1 details the number of people in each category.

3) COLLECTION TIME

The data is collected continuously for 24 hours. When tagging, to reduce the workload, we choose a time point every 4 hours based on food intake from the 24 hours a day, respectively 00:00, 04:00, 08:00, 12:00, 16:00, 20:00. One subject contributes 5-minute recordings at each time point.

C. DATASET ANNOTATION

We divide all the recordings into segments which last 5 seconds and do annotation on a coarse-grained and fine-grained level.

1) DIVIDING INTO 5-SECONDS RECORDINGS

Considering that the frequency of most body sounds is relatively low, around 1KHz, the signals are sampled at 4KHz with 16-bit quantization. The total duration of GI sound set is 60 hours. We divide the signals into segments of 5s for the subsequent marking process and get 43200 segments. Each segment is stored as a.wav file, and we refer one segment as one recording.

2) WEAK LABEL AND STRONG LABEL

The weak label means giving the type(s) of the sound event(s) occurring in an audio recording, without time boundaries. While strong label gives not only the types but also the onset and offset of the sound event occurring in the recording. Our task has two subtasks, both of them require dataset with the weak label when training. For testing, Task A requires a dataset with a weak label while Task B requires a dataset with a strong label. Therefore, we conduct weak markings in all the training data, while the test set has both a strong label and a weak label.

TABLE 2. **Classes and quantities of GI sound set.**

IDX	Class	No. of Recordings
01	Bowel sounds	17605
02	Speech	12828
03	Rub	10025
04	White noise	38042
05	Snore	4430
06	Groan	1470
07	Cough	560

Note that the audio collection process is definitely accompanied by white noise, we counted the amount of white noise but there was no white noise in sound event detection because it was ineffective.

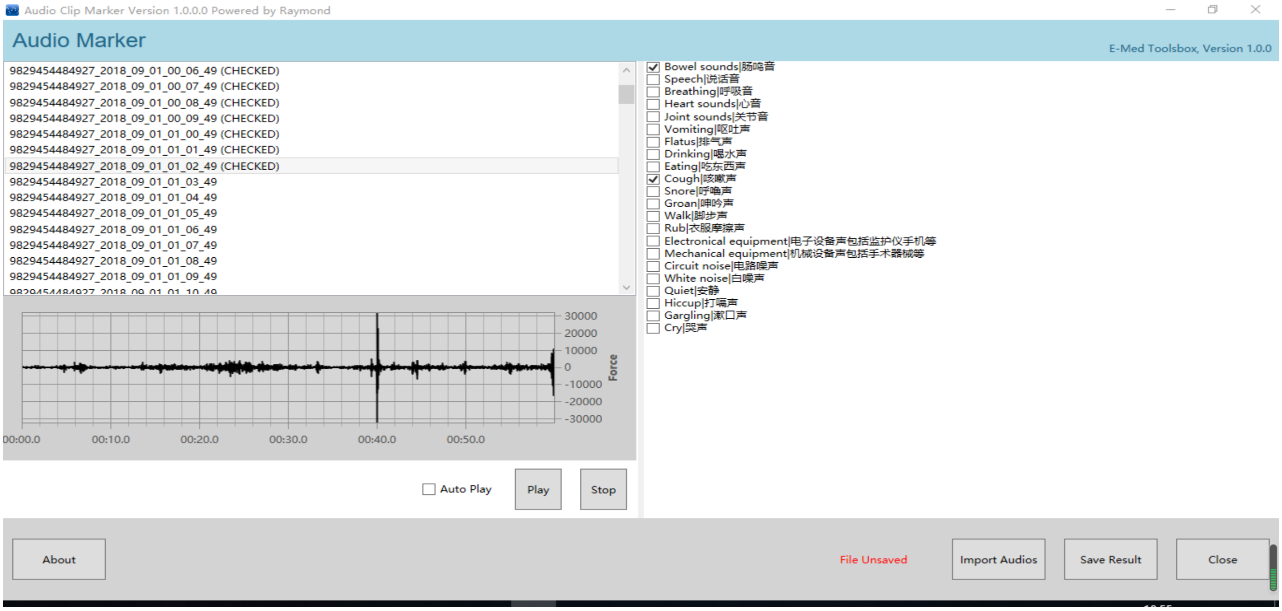
3) RATING METHODS

Our GI sound set learns from human rating methods of Audio set [25] constructed by Google. When marking weak labels, Audio Marker software presents a 5-second segment to human raters as shown in Fig.2. For each segment, raters are asked to independently rate the presence of one or more labels. The possible ratings are "present", "not present". Each segment is rated by three raters and a majority vote is required to record an overall rating. When marking strong labels, we divide a 5-second recording into 50 frames with 0.1 seconds for each frame considering the limited discrimination of the human ear and its practical application. Fig.3 shows the software for fine marking.

D. GI SOUND DISTRIBUTION AND MEDICAL SIGNIFICANCE

We make statistics on the types and quantities of sound events contained in the dataset, and present them in Table 2. Table 3. lists the audio quantity and other information of GI Sound Set, and compares it with other datasets.

Next, we illustrate the reasons for selecting the six sound event classes for GI Sound Set. Bowel sounds can effectively screen out functional Gastrointestinal disease such as irritable bowel syndrome (IBS) [33], monitor whether postoperative intestinal paralysis POI and paralytic intestinal obstruction will occur [34], and it is possible to fine guide the feeding time and dose of enhanced recovery after surgery(ERAS). The patient's pain has always been an important observation index after the operation. Currently, the subjective method such as the VAS score is mainly adopted. The detection of the patient's moaning can help assess the pain level of the patient, so as to use analgesics reasonably. Cough can cause pain and tear of surgical wounds. It is also the key to postoperative care in order to avoid postoperative cough as much as possible. Cough sound detection can help doctors assess the respiratory



9829454484927_2018_09_01_00_06_49 (CHECKED)
9829454484927_2018_09_01_00_07_49 (CHECKED)
9829454484927_2018_09_01_00_08_49 (CHECKED)
9829454484927_2018_09_01_00_09_49 (CHECKED)
9829454484927_2018_09_01_01_00_49 (CHECKED)
9829454484927_2018_09_01_01_01_49 (CHECKED)
9829454484927_2018_09_01_01_02_49 (CHECKED)
9829454484927_2018_09_01_01_03_49
9829454484927_2018_09_01_01_04_49
9829454484927_2018_09_01_01_05_49
9829454484927_2018_09_01_01_06_49
9829454484927_2018_09_01_01_07_49
9829454484927_2018_09_01_01_08_49
9829454484927_2018_09_01_01_09_49
9829454484927_2018_09_01_01_10_49

- ☑ Bowel sounds|肠鸣音
- ☐ Speech|说话音
- ☐ Breathing|呼吸音
- ☐ Heart sounds|心音
- ☐ Joint sounds|关节音
- ☐ Vomiting|呕吐声
- ☐ Flatus|排气声
- ☐ Drinking|喝水声
- ☐ Eating|吃东西声
- ☑ Cough|咳嗽声
- ☐ Snore|呼噜声
- ☐ Groan|呻吟声
- ☐ Walk|脚步声
- ☐ Rub|衣服摩擦声
- ☐ Electronical equipment|电子设备声包括监护仪手机等
- ☐ Mechanical equipment|机械设备声包括手术器械等
- ☐ Circuit noise|电路噪声
- ☐ White noise|白噪声
- ☐ Quiet|安静
- ☐ Hiccup|打嗝声
- ☐ Gargling|漱口声
- ☐ Cry|哭声

FIGURE 2. The usage diagram of the audio maker for coarse-grained marking.

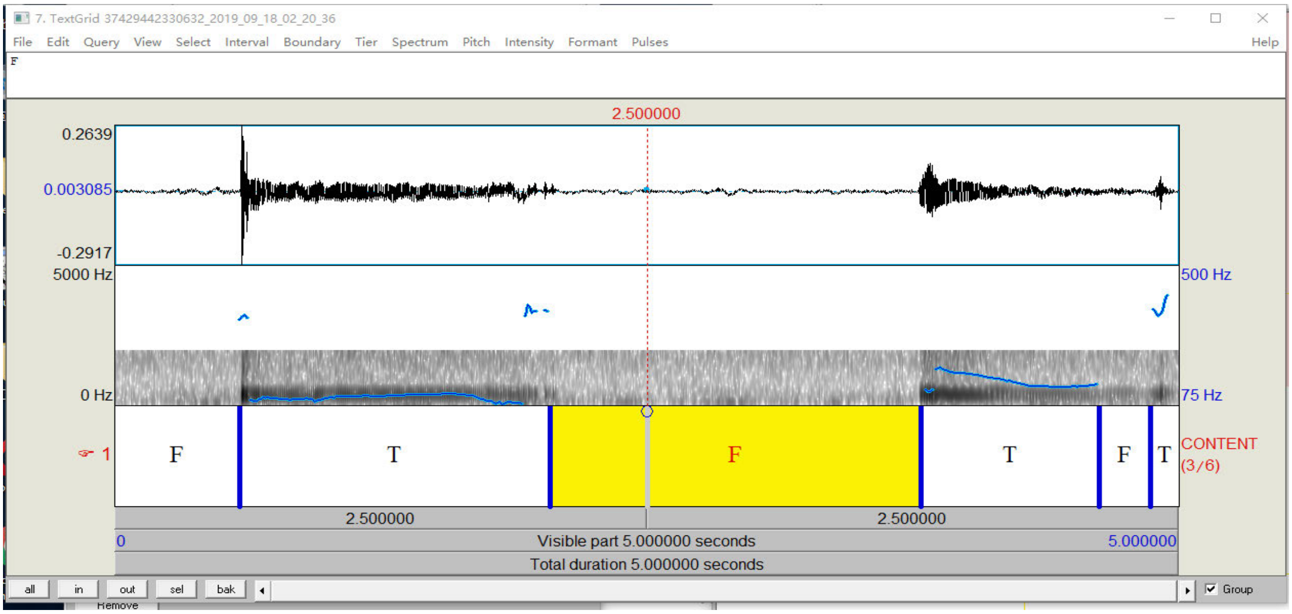


FIGURE 3. The usage diagram of software for fine-grained marking.

condition. Sleep quality is also important for patients, and snoring can be used as an indicator of sleep monitoring, especially for apnea OSA. In addition, frequent talking and turning over can also reflect the postoperative recovery status and comfort of patients, so the detection of voice and friction can also play a role.

III. CONVOLUTIONAL RECURRENT NEURAL NETWORK ARCHITECTURES

As is known, neural networks show great superiority over traditional machine-learning methods for a number of pattern recognition tasks, thus become the most popular method especially in image and speech recognition. The following we focus on CNN, Bidirectional Gated Recurrent Neural Networks(BiGRU) which are used in our work and introduce a special case of machine learning-MIL.

A. CONVOLUTIONAL NEURAL NETWORK

CNN contains the convolution layer, pooling layer, activation function, and fully connected layer basic components. Among them, convolutional layers are always interweaved with pooling layers. The data is passed from layer to layer in the form of a three-dimensional tensors, where each slice is called a feature map. If the input tensor of l − thlayer

TABLE 3. Comparison of GI sound set and others.

Dataset Name	No. of Recordings	Avg.Rec. Duration	Total Duration	No. SE Types	No. SE Instances	No. of Subjects
ESC-50 [38]	2,000	5s	2.8h	50	2,000	--
UrbanSound [39]	1,302	75s	27.0h	10	3,075	--
DCASE2017 [40]	52,763	10s	146.6h	17	N/A	--
TUT-SED2009 [41]	103	660s	18.9h	61	10,278	--
Google Audio Set [29]	2.1million	10s	8 months	527	N/A	--
ICBHI [1]	920	21.5s	5.5h	2	2750	126
GI Sound Set	**43,200**	**5s**	**60h**	**6**	**84,645**	**120**

is $x1 \in R^{H^l \times W^l \times D^l}$, the convolution kernel of the layer is$f^l \in R^{H \times W \times D^l}$, and the number of convolution kernel is D, the convolution operation can be expressed as (1):

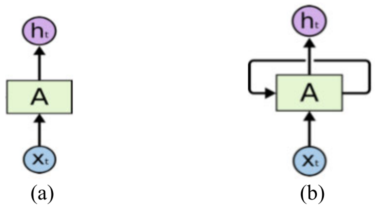
$$y_{i^{l+1},j^{l+1},d} = \sum_{i=0}^{H}\sum_{j=0}^{W}\sum_{d^l=0}^{D^l} f_{i,j,d^l,d} \times x^l_{i^{l+1}+i,j^{l+1}+j,d^l} \tag{1}$$

where $\left(i^{l+1}, j^{l+1}\right)$is the position coordinates of the convolution result.

The introduction of the pooling layer is to reduce the dimension which in other words is sampling and abstract the visual input object by imitating the human visual system. Pooling operation commonly used includes max-pooling and average-pooling.

CNN has been widely used in image processing because of its strong feature extraction ability. When processing audio, we can extract spectrograms or filter bank outputs as input, and the input is a 2-dimensional feature map whose axes are time and frequency, then it can be treated in the same way as an image.

According to [16], convolutional neural networks (CNN) are, in principle, very well suited to the problem of environmental sound classification. Here are two reasons. Firstly, when applied to spectrogram-like inputs, CNNs have a remarkable ability to capture energy modulation patterns across time and frequency, which has been shown to be an important feature in distinguishing different sound events, often noise-like sounds (such as engines and jackhammers [23]). Secondly, the net with convolution kernel and small receptive field is capable to learn and recognize the spectro-temporal pattern, which represents different sound classes.



(a) (b)

FIGURE 4. (a) The structure of a feed-forward neural network (b) the structure of RNN.

B. BIGRU

As shown in Fig. 4(a) [40], a time sequence is an input into the feed-forward network, which can only be independently processed frame by frame. It means that the output at time *t* is only based on the input, without taking advantage of any context information, which results in the loss and waste of information for machine learning tasks with sequence input. Splicing the input feature of several consecutive frames alleviates this problem to some extent, but it only provides limited contextual information. A more rational way is to make use of recurrent neural networks. The structure of RNN is shown in Fig. 4(b) [40]. They are competent for processing variable-length sequential input and learning temporal dependencies within the data.

Long Short-Time Memory (LSTM) networks are an evolution of RNNs, which can solve the gradient vanishing and gradient explosion problem in dealing with long sequence training. It performs better than RNN for long sequences. LSTM relies on the input gate, forget gate, and output gate to selectively affect the state information of RNN at a certain time. Its structure is shown in Fig. 5(a) [40].

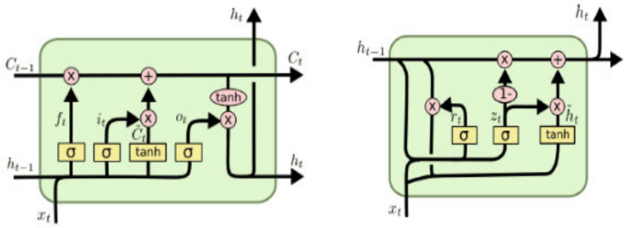


FIGURE 5. (a) The structure of LSTM (b) the structure of GRU.

Gated Recurrent Neural Networks(GRU) can be seen as a variant of LSTM. GRU replaces forget and input gates in LSTM with update doors, which is shown in Fig. 5(b) [40]. Combining cell state and hidden state, the method of calculating new information at the current time is different from LSTM.

In the classical cyclic neural network, the state is transmitted from front to back uniaxially. However, in some cases, the output at the current moment is not only constrained by the previous state but also related to the latter. In order to deal with this kind of problem, bidirectional RNN is needed, so is BiGRU.

C. MULTIPLE INSTANCE LEARNING

Multiple instance learning (MIL) is a special case of machine learning, which is the fourth framework in parallel with supervised learning, unsupervised learning, and reinforcement learning. It is proposed by Dietterich in 1997, in the context of a study of molecular activity: a molecule has many isomers chemically, yet some are effective in treating disease and others are ineffective. Based on this background, the training data contains no conformational information but only the validity of each molecule is provided, and the effectiveness of molecules is predicted by the training data [24]. Recently, multiple instance learning has been gradually applied to histopathological imaging for cancer detection based on the machine learning framework, which is a new method emerging in the field of computer-aided diagnosis.

Here are two concepts: bags and instance. Bags are made up of multiple instances, with a class label on the package and no label on the instance. In the molecular activity experiment, the molecules are bags, and their conformations are instances. Link to the work of this article, the whole recordings can be seen as bags while frames can be seen as instances.

The relationship between the bag label and the instance labels can be complex. Under binary labels, we adopt the standard multiple instance (SMI) assumption: a bag is positive if it contains at least one positive instance, and negative if it only contains negative instances. Our work in this article follows the SMI assumption. In other words, we regard a recording as a positive sample if there is one frame or more positive, while negative if all the frames are negative. Under MIL, we may not only determine recording as positive but also give out which frames are positive based on recording-level labels, this is why we adopt MIL to our work.

D. DATA AUGMENTATION

A classic assumption in machine learning is that the number of samples is equal for each class. However, actual tasks are difficult to satisfy this assumption, so is our GI Sound Set. In order to deal with this problem and make our model more general, there are usually data-level and algorithm-level processing methods.

We briefly introduce the processing method at the algorithm-level. Unbalanced samples cause the ''under learning'' of a small sample size class. The cost-sensitive method is often used to deal with this problem, which means increasing the penalty cost of small data samples misclassification and reflecting it in the objective function, then we can adjust the model's focus on small samples by optimizing our objective function. Since we make a data generator to decrease the unbalance of different classes, the cost-sensitive method nearly does not improve our metrics.

We employ three methods, including data sampling, class-balance sampling, and audio transformation at the data-level.

1) DATA SAMPLING

It is the most commonly used method from the data level. It contains over-sampling (or up-sampling), under-sampling (or down-sampling). In general, up-sampling is suitable for a class with a small sample size, which means that the audio of that class is copied to the scale of the maximum data volume class. Down-sampling can be used for the classes with a larger sample size. Note that the preciousness of data in deep learning, we can not simply discard some data, otherwise the diversity of training data will be reduced and the generalization ability of the model will be affected. The scientific method of down-sampling is to strictly control the number of samples of the specific class with a bigger sample size when randomly selected from each batch during batch training.

2) CLASS BALANCED SAMPLING

It focuses on classes, where the strategy is to categorize the samples by class, generating a list of samples for each class. When training, one or more classes are randomly selected, and then samples are randomly selected from the list of samples corresponding to each class. This ensures the opportunities are balanced for each class to participate in the training. We use this method on our GI Sound Set to deal with the imbalance of classes.

3) AUDIO TRANSFORMATION

It is also a good way to do data augmentation. To be specific, there are basically three methods: Time Shift Augmentation, Noise Augmentation, and Pitch Shift Augmentation. Timeshift augmentation means accelerating or decelerating the audio sample on the timeline while pitch keeps constant; Similarly, pitch shift augmentation requires increasing or decreasing the pitch of the audio sample on the frequency axis without changing the time axis. Noise augmentation means mixing the sample with another recording containing background sounds from different types of acoustic scenes.

IV. SOUND EVENT DETECTION(SED)-EXPERIMENTS

A. FEATURE EXTRACTION

Mel-Frequency Cepstral Coefficients and filter banks are two main popular features in the audio processing field. In this section, we discuss the differences between them, then describe the filter banks' extraction process in detail.

The calculation process is similar when getting MFCCs and filter banks, both of them calculate the filter bank, and MFCCs need to get some more steps on this basis. In a nutshell, a signal goes through a pre-emphasis filter; then gets overlapping frames and a window function is applied to each frame; afterward, do a Short-Time Fourier Transform on each frame and calculate the power spectrum; and subsequently, compute the filter banks. To obtain MFCCs, a Discrete Cosine Transform (DCT) is applied to the filter banks retaining a number of the resulting coefficients while the rest are discarded. A final step in both cases is mean normalization. Mel-filter is one of filter banks features.

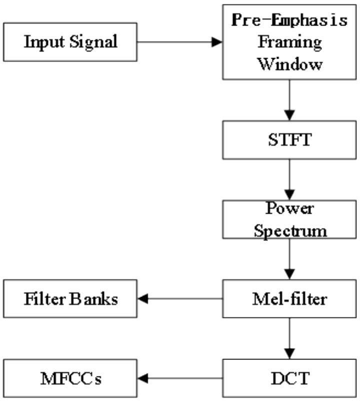


FIGURE 6. **A flow diagram for obtaining MFCCs and Filter banks.**

B. CRNN NETWORK ARCHITECTURE

We use the LibROSA toolkit [27] to extract the Mel-filter feature, where the parameters clip length= 5s, Mel-bins=64, and frame size=12.5ms.Under this parameter, the feature dimension extracted from a piece of audio is 400∗64. The CRNN network is made up of a 5-layer CNN network, a BiGRU network, and a fully connected layer as shown in Fig. 6. We implement the network based on the PyTorch toolkit [28] and send the extracted features of 400∗64 dimensions into the network as input. After passing through the convolution layers and pooling layers, a BiGRU is an input, and then the prediction probability of each frame of the sound event is obtained after passing through a fully connected layer composed of 6 neurons and the sigmoid function. The prediction at the frame-level is aggregated to output the prediction at the recording-level by pooling function. The loss function is calculated by comparing the predictions and labels at the recording-level. We use Adam algorithm to minimize the average cross-entropy, with an initial learning rate of 3e-4 and a batch size of 100 recordings. The threshold values are calculated from the probability of frame-level and recording-level through threshold calculation, which is used to generate output for evaluation. The optimal threshold value depends on our evaluation metrics F1.

C. EVALUATION METRICS

Sound Event Detection always includes two subtasks: classification and localization. Classification is to determine the types of sound events that occur in the recording, while localization gives timestamps after the determination of sound events.

Our task is a multi-classification task based on GI Sound Set, which is divided into two subtasks according to the result types. Task A is to do sound event detection on recording level and give recording-level labels, while Task B is to do sound event detection on 1s-level and give 1s-level labels. The following we introduce the corresponding evaluation metrics according to different sub-tasks [30].

1) TASK A: EVALUATION METRICS ON RECORDING LEVEL

The goal of Task A is to give coarse-grained prediction results, and it is evaluated by the micro-average on the recording-level which is defined as the harmonic average of the precision and recall [29] as shown in (2):

$$F_1 = \frac{2}{\left(\frac{TP}{TP+FP}\right)^{-1} + \left(\frac{TP}{TP+FN}\right)^{-1}} = \frac{2TP}{2TP+FP+FN} \tag{2}$$

TP: the number of correctly predicted sound events;
FN: the number of missed sound events;
FP: the number of spuriously predicted sound events;
Note that TP, FN, FP all three metrics are accumulated over all recordings and sound event types.

2) TASK B: EVALUATION METRICS ON 1S-LEVEL

Task B is to give fine-grained prediction results, and it is evaluated by the micro-average F1 and micro-average error rate(ER) on one-second segments. F1 is defined in a similar way as Eq.(2) in Task A, but with TP, FN, and FP counted at the segment level.

Error rate measures the number of errors in terms of *insertions* (I), *deletions* (D), and *substitutions* (S) which were often used in Automatic Speech Recognition(ASR) evaluation metric Word Error Rate(WER). To calculate the segment-based error rate, errors are counted segment by segment. In a segment m, the number of substitution errors $S(m)$ is the number of reference events for which a correct event was not output, yet something else was. This is obtained by pairing false positives and false negatives, without designating which erroneous event substitutes which. The remaining events are insertions and deletions: $D(m)$—the number of reference events that are not correctly identified (false negatives after substitutions are accounted for) and $I(m)$—the number of events in system output that are not correct (false positives after substitutions are accounted for) [30]. This leads to the following formula:

$$\begin{aligned} S(m) &= \min(FN(m), FP(m)) \\ D(m) &= \max(0, FN(m) - FP(m)) \\ I(m) &= \max(0, FP(m) - FN(m)) \end{aligned} \tag{3}$$

ER is calculated by (4), Note that S, D, I, and N are all calculated on 1-second segments.

$$ER = \frac{S+D+1}{N} = \frac{FN+FP-S}{TP+FN} \tag{4}$$

S: the minimum of false negatives and false positive;
D: the number of extra false negative;
I: the number of extra false positive predictions;
N: the total number of true occurrences of sound events.

3) OTHER METRICS FOR SED MULTI-LABEL TASK

There are MAP, MAUC, and d-prime three metrics, which are widely used in multi-label tasks according to the literature [25].

These metrics are designed to measure how well a system separates positive and negative recordings, and they do not require calculating thresholds compared with F1 and ER. The calculation is as follows: the system generates a sorted list of evaluation recordings for each sound event class, sorted in descending order according to the probability of the event. For each positive recording in the list, we set the threshold below its probability of occurrence and calculate a precision score. We define the mean of all sound event classes [35]. In order to amplify small changes, the d-prime metric can be calculated by wrapping MAUC making use of the following equation (5):

$$d' = \sqrt{2}\Phi^{-1}(AUC) \tag{5}$$

where Φis the accumulative density function of the standard normal distribution. Note that all of MAP, MAUC, and d-prime are the larger the better.

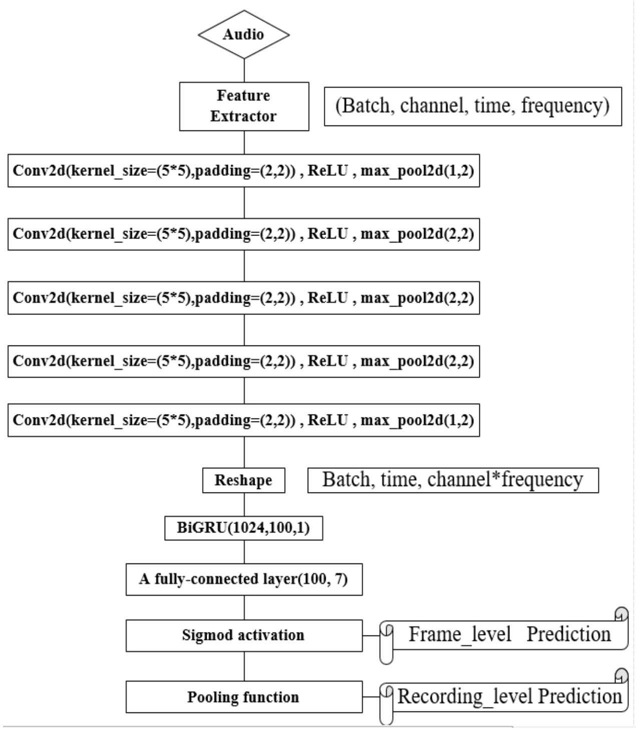

Audio
Feature Extractor
(Batch, channel, time, frequency)
Conv2d(kernel_size=(5*5),padding=(2,2)) , ReLU , max_pool2d(1,2)
Conv2d(kernel_size=(5*5),padding=(2,2)) , ReLU , max_pool2d(2,2)
Conv2d(kernel_size=(5*5),padding=(2,2)) , ReLU , max_pool2d(2,2)
Conv2d(kernel_size=(5*5),padding=(2,2)) , ReLU , max_pool2d(2,2)
Conv2d(kernel_size=(5*5),padding=(2,2)) , ReLU , max_pool2d(1,2)
Reshape
Batch, time, channel*frequency
BiGRU(1024,100,1)
A fully-connected layer(100, 7)
Sigmod activation
Frame_level Prediction
Pooling function
Recording_level Prediction

FIGURE 7. CRNN network structure.

D. EXPERIMENTS AND RESULTS

1) EXPERIMENTS SET UP

We divide GI Sound Set into a training set (40,530 recordings) and a test set (446 recordings). It should be noted that the training set and test set are disjoint and independent. The test set is another set of samples collected under the same conditions as the training set. The test set is strongly labeled so they can be evaluated for both subtasks, but the training set only comes with presence/absence labeling. Meanwhile, the test set has balanced numbers of events, but the training set is unbalanced. We set aside 2000 recordings from the training set to make a balanced validation set, and use the remaining 38,530 recordings for training.

TABLE 4. Comparison of CRNN Activation Functions.

		Max Pooling	Ave. Pooling	Lin. Softmax	Exp. Softmax	Attention
TaskA	TP	501	496	502	484	495
	FN	111	116	110	128	117
	FP	155	1405	166	166	139
	Precision	76.37	26.09	75.15	74.46	78.08
	Recall	81.86	81.05	82.03	79.08	80.88
	F1	**79.02**	**39.47**	**78.44**	**76.70**	**79.45**
TaskB	TP	752	1015	844	1056	895
	FN	523	260	431	219	380
	FP	626	8092	804	1940	1288
	Precision	54.57	11.15	51.21	32.25	41.00
	Recall	58.98	79.61	66.20	82.82	70.20
	F1	**56.69**	**19.55**	**57.75**	**49.45**	**51.76**
	Sub.	82	260	126	131	181
	Del.	441	0	305	88	199
	Ins.	544	7832	678	1809	1107
	ER	**83.69**	**634.67**	**86.98**	**159.06**	**116.63**

Since our training set is unbalanced, we make a data generator for each sound event class to overcome the negative effects of data imbalance. We number classes and extract features according to classes. When training, the data generators cycle through all the feature files with a certain class and sample uniformly to form a training batch.

2) COMPARISON OF CRNN ACTIVATION FUNCTIONS

We initiate our experiments with finding an appropriate and high performing Activation Function (linear/attention/max /average/exponential) by using Mel-filter features. Table 4 lists the results of the max pooling, average pooling, linear softmax, exponential softmax, and attention systems on both subtasks of GI Sound Set. Attention systems outperform than others in terms of F1 for Task A. However, Max systems show the best performance in terms of F1 and error rate for Task B. Combining two subtasks, we choose max pooling function as the prime one.

3) COMPARISON OF CRNN NETWORK SIZE

Based on the CRNN network with 5 Conv layers following by 1 GRU layer as Fig.7, we try different layers and different neurons to find an optimal network taking metrics and speed into account. Table 5 lists the results, and Fig. 8 shows the results with varying numbers of neurons per hidden layer and varying numbers of hidden layers per model.

4) COMPARISON OF DIFFERENT NETWORKS

After a series of basic operations on CRNN, we try to find which of the three networks performs the best: BiRNN, BiGRU, and LSTM. We have such prior knowledge: LSTM and GRU are variants of RNN. Due to gradient vanishing,

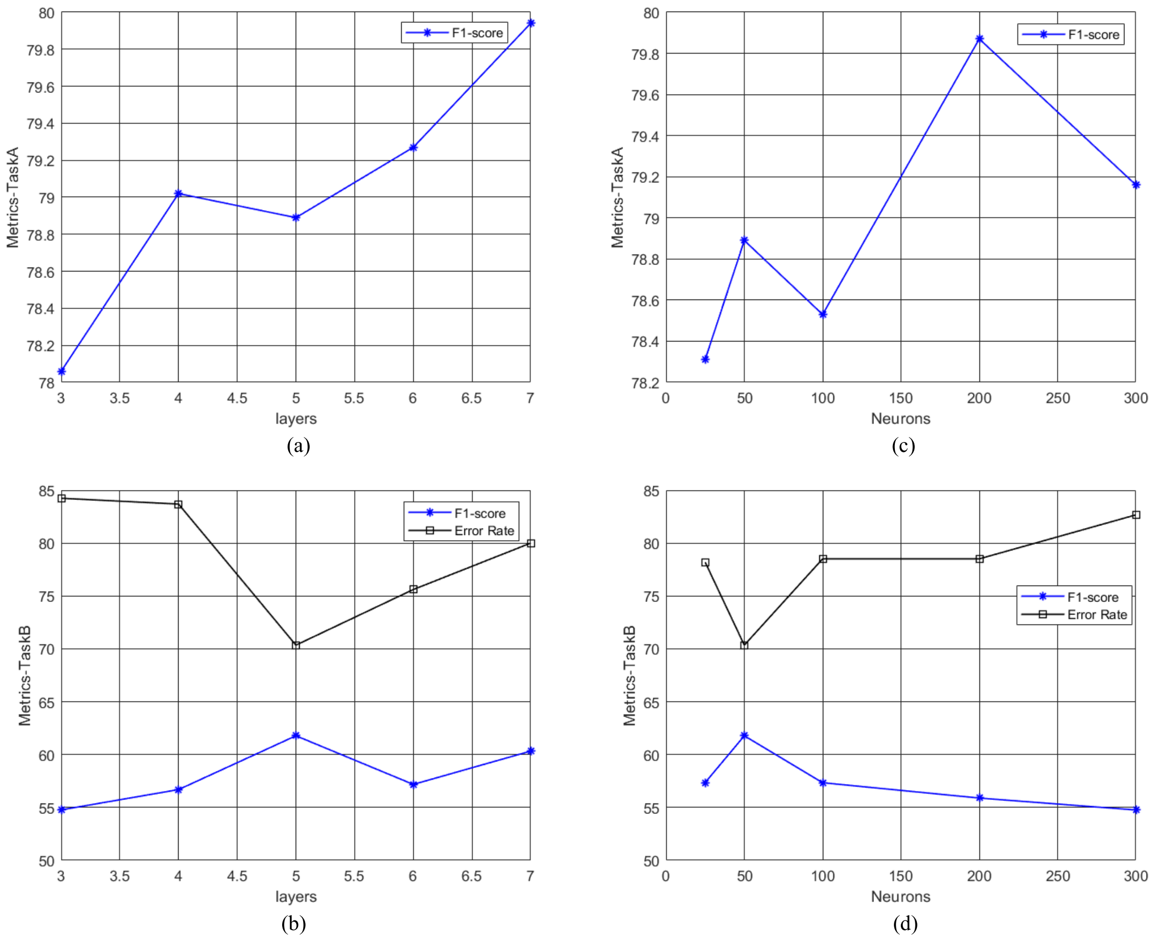


FIGURE 8. Comparison of CRNN network size. (a)(b) shows the F1-scores for Task A and F1-scores, Error Rate for Task B for a CRNN using {3,..., 7} CNN layer. (c)(d)shows F1-scores for Task A and F1-scores, Error Rate for Task B for a CRNN with 5 CNN layers using {25, 50,..., 300} neurons per layer.

TABLE 5. Comparison of CRNN Network size.

CRNN		Task A	Task B	
layers	GRU neurons	F1	F1	Error Rate
3	50	78.06	54.77	84.24
4		79.02	56.69	83.69
5		**78.89**	**61.78**	**70.35**
6		79.27	57.18	75.61
7		79.94	60.33	80.00
5	25	78.31	57.35	78.20
	50	**78.89**	**61.78**	**70.35**
	100	78.53	57.34	78.51
	200	79.87	55.89	78.51
	300	79.16	54.76	82.67

RNN can only have short-term memory. The LSTM network solves the problem of gradient vanishing by combining short-term memory with long-term memory. Compared with LSTM, GRU is easier to converge. Table 6 shows the results.

TABLE 6. Comparison of Different Networks.

	Task A	Task B	
Network Architecture	**F1**	**F1**	**Error Rate**
CNN+BiGRU	**78.89**	**61.78**	**70.35**
CNN+BiRNN	77.58	55.48	85.65
CNN+BiLSTM	77.73	60.40	72.39

5) COMPARISON OF INPUT FEATURES

Table 7 shows the results for CRNN with single input feature such as MFCCs, spectrograms, Mel-filters, transfer learning features extracted by VGGISH, and with combined features. The Mel-filter combined with the spectrogram feature shows promising results. Note that we do not apply data augmentation in those contrast experiments.

6) EVALUATION OF THE FINAL SETUP ON THE TEST SET

Table 8 shows the optimal results, experiment setup (i.e. CRNN,5 CNN layers, 50 neurons, data augmentations, max-pooling functions, input fusion features combined with

TABLE 7. Comparison of Different Input features.

	Task A	Task B	
Input Features	**F1**	**F1**	**Error Rate**
Mel-filter	76.39	57.98	75.12
MFCC	70.48	51.91	81.88
Spectrograms	66.61	56.19	75.29
Vggish embedding	**79.65**	**60.51**	**72.16**
Mel-filter + MFCC	77.86	61.11	76.94
Mel-filter + Spectrograms	78.29	57.87	84.94
Mel-filter+MFCC+Spectrograms	**81.06**	**63.40**	**70.13**

Mel-filter, MFCC, and Spectrograms) and metrics for both Task A and Task B evaluated on the test set.

7) RESULTS ANALYSIS FOR CLASSES

We use confusion matrices to visualize the confusion between the 6 sound event types on our GI Sound Set, both on the recording level and the 1-second segment level. The recording or the second segment may have multiple ground truth labels, as well as multiple predicted labels considering the existence of polyphonic sounds. To deal with that

the problem, we divide the recordings into fractional recordings with equal weights, So it is with segments. Fig. 9. shows the confusion matrices of the optimal results on the test data for both Task A and Task B.

8) COMPARISON WITH RELATED WORKS

Reference [41] presents an algorithm for automatic detection of cough events from acoustic signals. The algorithm uses three spectral features with a logistic regression model to separate sound segments into cough and non-cough events. Its dataset includes a total of 43 real-world environment recordings, with a length of over 32 minutes, from different patients. In these recordings, cough sound locations are manually marked to assess the performance of the proposed cough detection method. As a result, a total of 980 events are labeled as cough sounds whilst more than 1000 separate non-cough sound events are also identified. Non-cough sound events include speech, laughter, machine noises, and other types of background noise. The algorithm achieves a high F1 of 88.07.

Reference [42] proposes such a method and achieves a very high F1 of 94.93. The centerpiece of the proposed method is a recurrent neural network for modeling of sequential data with variable length. Clinical sleep recordings from 20 subjects are recorded by a microphone, which are used to assess the performance of the proposed method. Mel-frequency cepstral coefficients, which is extracted from snoring and non-snoring sound events, are used as inputs to the proposed network.

Reference [43] constructs a wearable bowel sound monitoring system using off-the-shelf components. It proposes a new algorithm for quantitative bowel sounds classification and gets F1 of 77.62.

Our experiment is a multi-classification task. Based on GI Sound Set, we extract Mel-filter features and use CRNN architecture to output labels. As a result, we achieve a micro average F1 of 81.06. Focusing on each sound event, we achieve F1 of 65.67 on Bowel Sound, F1 of 91.09 on Cough, high F1 of 97.51 on Snore.

Compared with [41]–[43], there are some differences:

① Task: [41]–[43] focus on one specific type of sound event while our intention is to do multi-classification. For example, [42] only outputs cough and non-cough events, our experiment outputs the results of all those 6 sound events.

② Dataset: The dataset is labeled on the event-level in [41]–[43] while our GI Sound Set is labeled on the recording-level. In terms of generalization and the data volume, GI Sound Set is superior to them.

③ Method: Under the same conditions, the performance of neural networks excel machine learning methods. From results in Table 9, CRNN architecture is more capable to learn features from audios than traditional methods.

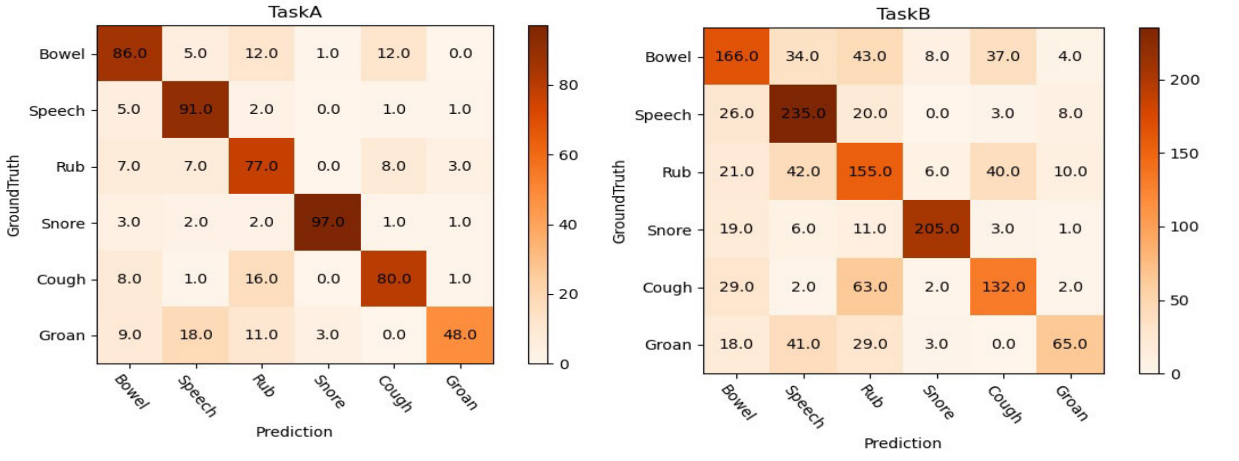
④ Results: For snore and cough, our experiment performs better than [41], [42]. However, it performs less than [43] when it comes to bowel sound. We speculate that this is mainly due to the strong characteristics of cough and snore that are different from other sound events. The frequency of bowel sound is lower, which is easily confused with other sound events and leads to poor performance.

Comparison between our experiment and [41]–[43] shows that CRNN architecture has the ability to learn most of the features in audio and detect those 6 sound events. Our system achieves better performance, which can be extended to other audio sounds.

V. DISCUSSION

In our experiments, we compare activation functions, and we find max Pooling is the optimal one combining performance on both tasks. Then we test different CNN layers and neurons of BiGRU hidden layer on two subtasks performance. The increase of CNN layers does not result in a huge improvement in metrics, but slows down the training speed and occupies too much memory, so does the increasing neurons of the hidden layer. We measure metrics of BiGRU, RNN, LSTM along with the CNN layer forward, and find BiGRU performs better than the other two. Note that the experiments mentioned above are all based on the GI sound set with data augmentation and extract Mel-filter features as an input of the network.

In the subsequent experiment, we extract different features based on the raw dataset which means without data augmentation. Results show that Mel-filter performs the best among the three of Mel-filter, MFCC, and Spectrograms when inputting a single feature. Transfer learning shows the absolute advantage over regular audio features. We discover



(a) (b)

FIGURE 9. Confusion Matrices.

TABLE 8. Final Set-Up for Evaluation.

Network Architecture	CNN layers	Neurons	Pooling Functions	Data Augmentation(Y/N)	Features	Task A	Task B	
						F1	**F1**	**Error Rate**
CRNN	5	50	Max	Y	Mel-filter+ MFCC+ Spectrograms	81.06	63.40	70.13

TABLE 9. Comparison with related works.

Experiment	Dataset			Task	Feature	Method	Sound Event	Recording-level F1
	#Class	#Subjects	Total Duration					
[43]	2	N/A	32 mins	Binary: cough or not	Spectra	Logistic Regression	Cough	88.70
[44]	2	20	N/A	Binary: snore or not	MFCC	RNN	Snore	94.93
[45]	2	17	N/A	Binary: bowel sound or not	Spectra	SVM	Bowel Sound	77.62
Ours	6	120	60 h	Multi-classification	Mel-filter	CRNN	Cough	**91.09**
							Snore	**97.51**
							Bowel Sound	65.67

that the combination of conventional audio features for early fusion input into the network improves the performance.

The following conclusions can be interpreted from the confusion matrices:

a. There is a correlation between the results on the recording level and the 1-second segment level. A good result on task A will likely lead to good results on task B such as Speech and Snore, and vice versa.
b. Speech and Snore are the two best-learned sound event types. Snore is seldom recognized as other events, nor are other events often recognized as it, while Speech has a little confusion with Groan.
c. Groan has the lowest performance. Because the Groan training data is the least on our GI Sound Set, and there is some confusion between Speech and Groan which agrees with intuition.
d. Bowel sound is the one that we pay more attention to. However, Its performance is not up to snuff. The frequency of Bowel sound is 1KHz approximately. We guess that a little low frequency may lead to it. Even worse Bowel sounds have some confusion with other classes such as Rub and Cough. This fits our perception: the collection process of Bowel Sound is often accompanied by more or less Rub.

This method can be extended to various other categories of body-related acoustic events, including breath sounds, heart sounds, vomiting, laughing, crying, walking, and so on. Combined with 24-hour long-term monitoring of more clinical trials, the postoperative recovery of patients with different types of surgery can be analyzed, and the gastrointestinal activity of functional gastrointestinal patients can be recorded, thus providing doctors with help in patient monitoring and auxiliary diagnosis.

VI. CONCLUSION

In this article, we set up GI Sound Set collected and labeled by ourselves, with the best environmental generalization of time, place, and disease, which is the largest data set for gastrointestinal sounds to date. Based on our GI Sound Set, we introduce a CRNN system to do sound event detection both recording level and segment level only based on a weakly labeled training set. We get 81.06% F1 on Task A, 63.40% F1 and 70.13% Error Rate on Task B. Our experimental methods can be implemented on a larger dataset with more kinds of sound events, and give the results on frame level, which can be used to diagnose the patient's disease and determine the recovery of the patient. The proposed system offers great possibilities for future intelligent monitoring of patient health conditions, Meanwhile how to improve F1 further and decrease Error Rate of the recording and frame-level requires more exploration and experimentation.

ACKNOWLEDGMENT

The authors acknowledge Xianglong Duan of the People's Hospital of Shanxi Province, Yixin Zhou, and Hao Tang of Beijing Ji Shui Tan Hospital, Tiejun Wu, Zongwang Zhang and Huili Kan of the People's Hospital in Liao Cheng City, for providing great help in collecting data and doing research.

REFERENCES

[1] B. M. Rocha *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiol. Meas.*, vol. 40, 2019, Art. no. 035001.

[2] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000. [Online]. Available: http://circ.ahajournals.org/content/101/23/e215.full

[3] X. Du, G. Allwood, K. Webberley, A. Osseiran, and B. Marshall, "Bowel sounds identification and migrating motor complex detection with low-cost piezoelectric acoustic sensing device," *Sensors*, vol. 18, no. 12, p. 4240, Dec. 2018.

[4] Y. Wang, "Polyphonic sound event detection with weak labeling," Ph.D. dissertation, Google Inc.Google Scholar, 2017.

[5] M. H. Swartz, "The physical examination," in *Textbook of Physical Diagnosis: History and Examination*, M. H. Swartz, Ed., 7th ed. Philadelphia, PA, USA: Elsevier Saunders, 2014, ch. 4.

[6] R. Malkin, D. Macho, A. Temko, and C. Nadeu, "First evaluation of acoustic event classification systems in CHIL project," in *Proc. Joint Workshop Hands-Free Speech Commun. Microphone Arrays (HSCMA)*, 2005, pp. 1–2.

[7] C. Zieger, "An HMM based system for acoustic event detection," in *Multimodal Technologies for Perception of Humans*. Berlin, Germany: Springer, 2007, pp. 338–344.

[8] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. Huang, "HMM-based acoustic event detection with AdaBoost feature selection," in *Multimodal Technologies for Perception of Humans*. Berlin, Germany: Springer, 2008, pp. 345–353.

[9] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2010, pp. 1267–1271.

[10] A. Waibel and R. Stiefelhagen, *Computers in the Human Interaction Loop*. London, U.K.: Springer-Verlag, 2009.

[11] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2013, no. 1, p. 1, Dec. 2013.

[12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2001, pp. 556–562.

[13] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 559–563.

[14] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," 2016, *arXiv:1604.06338*. [Online]. Available: http://arxiv.org/abs/1604.06338

[15] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2015, pp. 1–6.

[16] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017, doi: 10.1109/LSP.2017.2657381.

[17] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," 2016, *arXiv:1604.07160*. [Online]. Available: http://arxiv.org/abs/1604.07160

[18] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2721–2725.

[19] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6440–6444.

[20] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Proc. Workshop Detection Classification Acoustic Scenes Events (DCASE)*, 2016, pp. 6–10.

[21] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection," in *Proc. Workshop Detection Classification Acoustic Scenes Events (DCASE)*, 2016, pp. 35–39.

[22] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.

[23] J. Salamon and J. P. Bello, "Feature learning with deep scattering for urban sound analysis," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2015, pp. 724–728.

[24] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, Jan. 1997.

[25] D. E. Rumelhart and J. L. McClelland, "Learning internal representations by error propagation," in *Neurocomputing: Foundations of Research*, J. A. Anderson and E. Rosenfeld, Eds. Cambridge, MA, USA: MIT Press, 1988, pp. 673–695.

[26] Y. Yin, H. Jiang, S. Feng, J. Liu, P. Chen, B. Zhu, and Z. Wang, "Bowel sound recognition using SVM classification in a wearable health monitoring system," *Sci. China Inf. Sci.*, vol. 61, no. 8, Aug. 2018, Art. no. 084301.

[27] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 776–780, doi: 10.1109/ICASSP.2017.7952261.

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS Workshop*, 2017, pp. 1–4.

[29] B. McFee, C. Rael, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.

[30] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, p. 162, May 2016.

[31] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.

[32] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, "Chime-home: A dataset for sound source recognition in a domestic environment," in *Proc. Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[33] X. Du, G. Allwood, K. Webberley, A.-J. Inderjeeth, A. Osseiran, and B. Marshall, "Noninvasive diagnosis of irritable bowel syndrome via bowel sound features: Proof of concept," *Clin. Transl. Gastroenterol.*, vol. 10, no. 3, p. e00017, 2019, 10.14309/ctg.0000000000000017.

[34] J. Nors *et al.*, "Postoperative paralytic ileus after cytoreductive surgery combined with heated intraperitoneal chemotherapy," *Pleura Peritoneum*, vol. 5, no. 1, 2019.

[35] R. Wang and K. Tang, "Feature selection for MAUC-oriented classification systems," *Neurocomputing*, vol. 89, pp. 39–54, Jul. 2012, doi: 10.1016/j.neucom.2012.01.013.

[36] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018.

[37] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 1041–1044.

[38] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. Detection Classification Acoust. Scenes Events Workshop (DCASE)*, 2017, pp. 1–9.

[39] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio context recognition using audio event histograms," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2010, pp. 1272–1276.

[40] C. Olah. (Aug. 27, 2015). *Understanding LSTM Networks*. Accessed: Dec. 14, 2017. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LS

[41] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "Automatic cough detection in acoustic signal using spectral features," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Berlin, Germany, Jul. 2019, pp. 7153–7156.

[42] B. Arsenali, J. van Dijk, O. Ouweltjes, B. den Brinker, D. Pevernagie, R. Krijn, M. van Gilst, and S. Overeem, "Recurrent neural network for classification of snoring and non-snoring sound events," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Honolulu, HI, USA, Jul. 2018, pp. 328–331.

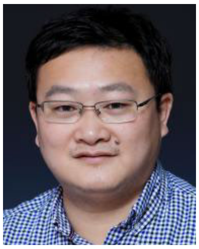
[43] Y. Yin, H. Jiang, S. Feng, J. Liu, P. Chen, B. Zhu, and Z. Wang, "Bowel sound recognition using SVM classification in a wearable health monitoring system," *Sci. China Inf. Sci.*, vol. 61, no. 8, pp. 222–224, Aug. 2018.

XUE ZHENG received the B.S. degree from the College of Optoelectronic Engineering, Chongqing University, China, in 2018. She is currently pursuing the M.S. degree with the Institute of Microelectronics, Tsinghua University, China. Her research interests include deep learning and audio processing.

CHUN ZHANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 1995, 1997, and 2000, respectively. In 2000, he became a Faculty Member with Tsinghua University, where he has been an Associate Professor, since 2005, and the Director of the Integrated Circuit and System Division, Institute of Microelectronics, since 2008. He has published over 60 articles and four books. He holds 14 patents. His current research interests include signal processing, CMOS integrated circuit design, RFID, and biomedical applications. His ongoing projects include RFID, SERDES, SSD, low-power wireless transceivers, and implanted stimulators.

PING CHEN received the B.S. degree from the University of Science and Technology, Beijing, in 2012, and the M.S. degree from Tsinghua University, in 2015. From 2014 to 2015, he was a Research Assistant with the Computational Intelligence Laboratory, Lincoln University, U.K. Since 2015, he has been an Engineer with Beijing Yiemed Medical Technology Company Ltd., focus on medical instrument development.

KANG ZHAO received the B.S. degree in microelectronics from Beijing Jiao Tong University, Beijing, China, in 2014. He is currently pursuing the Ph.D. degree with the Microelectronic Engineering Department, Tsinghua University, Beijing. His research interests include signal processing and relevant circuits design of human body sounds.



HANJUN JIANG (Senior Member, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2001, and the Ph.D. degree in electrical engineering from Iowa State University, Ames, IA, USA, in 2005. From 2005 to 2006, he was with Texas Instruments, Dallas. He is currently an Associate Professor with Tsinghua University. He has authored more than 80 peer-reviewed journals and conference papers. His current research interests include the area of low-power circuit design and system-level integration for wireless medical and healthcare applications. He is also an Associate Editor of the IEEE Transactions on Biomedical Circuits and Systems.

ZHIWEI JIANG was born in Jiangsu, China, in 1969. He is currently the Doctor of Surgery, the Chief Physician title, and a second-class Professor, specializing in the various minimally invasive treatments of gastrointestinal tumors. He is also the Vice President of the Affiliated Hospital, Nanjing University of Chinese Medicine, and the Director of general surgery.

HUAFENG PAN was born in Jiangsu, China, in 1986. He is currently the Doctor of Surgery and an Attending Physician, specializing in the various minimally invasive treatments of gastrointestinal tumors and relevant clinical research.

ZHIHUA WANG (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 1983, 1985, and 1990, respectively. From 1992 to 1994, he was a Visiting Scholar with the Carnegie-Mellon University, Pittsburgh, PA, and Katholieke University, Leuven, Belgium. His research interests include the design methodology of integrated circuits and systems, lower power analog and RF ICs for medical and communication, and high-speed real-time signal processing. In his more than 24 years of academic experience, he has published over 170 academic articles, three books, filed more than 25 patents, and accomplished over 15 research projects. He is currently an Official Member of the China National Commission of URSI, since 2000. He is also the Founder of the IEEE Solid-State Circuit Society Beijing Chapter and served as the Chapter Chairman since 2000. He has served as a Technologies Program Member of International Solid-State Circuit Conference (ISSCC) for the years of 2005, 2006, 2007, and 2008. He is also the Deputy Chairman of the Beijing Semiconductor Industries Association and ASIC Society of Chinese Institute of Communication, and the Deputy Secretary-General of Integrated Circuit Society in China Semiconductor Industries Association.

WEN JIA received the B.S. degree from the Department of Electronic Science and Technology, Huazhong University of Science and Technology, Wuhan, China, in 2008, and the M.S. degree from the Institute of Microelectronics, Tsinghua University, Beijing, China, in 2011. She is currently with the Research Institute, Tsinghua University, Shenzhen, Guangdong, China. Her research interests include circuits and systems for wireless medical and healthcare applications, reconfigurable computing and multimedia processing, modeling, and simulation.