A Study on Sound Analysis Algorithm for Heart Sounds using YOLO Deep Learning Model

Hiroki Tamura1, Praveen Nuwantha Gunaratne2, Hiromu Takeguchi3, Hiroyuki Fukumoto4 and Yoshifumi Hashiguchi5

1,2,3University of Miyazaki, Japan, 4Kobayashi Municipal Hospital, Japan
5Desan Co.Ltd.
1htamura@cc.miyazaki-u.ac.jp, 2praveenwuwantha@gmail.com,
3hi19025@student.miyazaki-u.ac.jp, 4k_hosp3@city.kobayashi.lg.jp,
5yhashiguchi@densan-soft.co.jp
*Corresponding Author

Received 09 August 2022; accepted 29 August 2022

Abstract
As Japan’s population ages, the demand for home medical care is increasing. And, as the demand for home medical care increases, the burden on medical personnel becomes a problem. One of the ways to promote home medical care is to spread the use of medical equipment in households. However, since some medical knowledge is required to use basic medical equipment, it is considered difficult to spread the use of medical equipment among general households. Therefore, it demands the necessity to develop a medical device that can give the same decision as a medical doctor by using an algorithm. In this paper, we study the construction of an algorithm for an AI stethoscope that can make the same decisions as a medical doctor. We combine a frequency analysis method using three features and an image processing method using an image that represents the frequency features by wavelet transform. Using the results of each of these methods, we aim to improve the identification rate through machine learning techniques. The Random Forest training yields an identification rate of 94.68% on the dataset of this paper.

Keywords: Heart sounds, Wavelet transform, YOLOv2tiny, Deep learning, Support vector machine

1. Introduction
We predict that the medical devices will become widespread in ordinary households in near future. Using these devices, anyone can make fairly the same decisions as medical doctors, and home medical care is a way to reduce the burden on medical doctors. Currently, the operation of such basic medical devices requires the judgment of a medical doctor in order to pursue appropriate diagnosis and treatment procedures. In this way, since medical knowledge is a must to use medical equipment, the popularization of home medical equipment has been a difficult task for ordinary households. Therefore, it is necessary to develop medical devices that make the medical doctor's judgment algorithmically and allow anyone to make the same judgment as a medical doctor [1].
Here, we used machine learning techniques aimed at correctly classifying the audio data into the classification categories derived from judgment of a medical doctor. The target of this paper is heart sounds. The audio was recorded for shunts, both before and after stress. The wavelet transform [2][3] was used to image, shunt sounds before and after stress, and we trained YOLOv2tiny [4][5] to capture the features. In this paper, we found that YOLOv2tiny alone did not improve the identification rate. Therefore, we performed multivariate analysis using the multivariate calculated through the judgment of YOLOv2tiny. Here, we compare the identification rate of the result of YOLOv2tiny alone with the result of combination of YOLOv2tiny and multivariate analysis, and we explain how to achieve a higher identification rate.

2. Proposed method

Here, we have hypothesized that it is possible to diagnose using the heart sound and the linked judgment by a medical doctor. First, the audio data was converted into images using wavelet transform, and YOLOv2tiny was trained with the transformed images. Next, we applied the trained YOLOv2tiny to all available data. The identification rates were calculated under two classifications: normal and abnormal. In order to further improve the identification rate, we used multiple regression analysis, decision tree, random forest, and SVM as learning methods.

2.1. Image using wavelet transform

Digitize the acquired audio file in ‘.wav’ file format. Peak detection process is performed on the digitized data, and the peak is calculated. The analysis data is from first of the calculated peak to a 10 second data set. The sampling frequency is set to 2000 [Hz]. Wavelet transform is performed on the recorded data. The mother wavelet uses the Morlet function and this process outputs an image as shown below. The Fig. 1 depicts an example of the actual results. The vertical axis represents the frequency (Hz), horizontal axis represents the time slot (seconds) of experimental data, and the color bar represents amplitude (spectrum).

![Output of wavelet transform](image)

Figure 1. Output of wavelet transform. Left figure: Normal heart sound data, Right figure: Abnormal heart sound data
2.2. **YOLOv2tiny**

The YOLOv2 is one of the object detection method that introduces a convolutional network, consisting of 3 layers of convolutional layers and 5 layers of maxpool layers. YOLOv2tiny calculates the normal reliability score, which indicates the characteristics of healthy subjects, and the abnormal reliability score, which indicates the characteristics of sick patients, on the identification image. Based on the result, image clustering is performed. The Fig. 2 shows an example of the results, which is clustered with normal, and a reliability score of 0.998.

![Image of YOLOv2tiny output](image)

Figure 2. Output of YOLOv2tiny

Here, an open database in PhysioNet (healthy: 540, sick: 540) was used. The images after normalization of healthy person: 371 data, sick person: 371 data, were used as training data. The images after normalization of healthy person: 169 data, sick person: 169 data, were used as test data. System was learned through YOLOv2tiny method using the training data, and test data were set for identification. The results of this chapter are shown in Table 1. We succeeded in constructing an algorithm that exceeds the target identification rate of 95%.

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Normal (90.5%)</th>
<th>Abnormal (9.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>normal</td>
<td>153 (90.5%)</td>
<td>16 (9.5%)</td>
</tr>
<tr>
<td>abnormal</td>
<td>0</td>
<td>169 (100%)</td>
</tr>
</tbody>
</table>

2.3. **Method using image data of Wavelet Transform result and result of Fast Fourier Transform**

Although wavelet transform can obtain the characteristics of temporal changes, there is a possibility that the spectrum and power information at the frequency is insufficient. Therefore, here we propose to add analysis of sound data with Fast Fourier Transform for processing. We use three types of neural networks (input layer-hidden layer-output layer: Model1 (128-10-1), Model2(128-10-1),Model3(128-20-3-1)) where the input data is FFT results. The flow of this process is shown in Fig. 3. The FFT and neural
networks processing flow is represented under “Frequency Analysis”, and the process explained in section 2.2 is the “Image Recognition”. Frequency Analysis outputs three from three model neural network and Image Recognition outputs two reliability scores. The output was determined by fusing them with “Machine Learning”. In Machine Learning part Decision tree, Random forest and SVMs were used[6][7][8][9].

![Figure 3. The flow of purpose method](image)

3. Experiment

In this paper, we used not only the data from a database but also data measured from the actual patients. Fig. 4 shows the auscultation sites used for data acquisition. Auscultation at the site 1 and 2, as per Fig. 4 was performed, and an electronic stethoscope was applied to the site where heart sound could hear from either of them and auscultation was performed for 30 seconds. Total number of data is 3122. This protocol was approved by the Ethics Committee of the University of Miyazaki, Japan (Protocol code O-1097).

![Figure 4. The auscultation site of heart sound](image)

For the YOLOv2tiny model and neural networks, the ratio of training data to test data was 7:3. The identification rate was calculated as the percentage of correct answers in the test data. The labels used were: good (i.e., normal; n =1672) and need for treatment (i.e., abnormal; n = 1450). Decision tree, random forests, and SVMs were all trained using MATLAB by MathWorks, and the identification rates were verified by single-tailed cross-validation. For decision trees, random forests, and SVMs, the identification rate was calculated as the average of 100 cross-validation trials. Table 2 shows the environment of the computer used.
Table 2. The environment of the computer

<table>
<thead>
<tr>
<th>Learning environment</th>
<th>Execution Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>32GB</td>
</tr>
<tr>
<td>CPU</td>
<td>Corei9 – 10850K</td>
</tr>
<tr>
<td>GPU</td>
<td>GeForce RTX 2080 SUPER (VRAM:8GM)</td>
</tr>
<tr>
<td>CUDA</td>
<td>11.0</td>
</tr>
</tbody>
</table>

4. Results

Table 3 illustrates the results of the proposed method under Section 2.3. There are two types of voice data measurements, 5 seconds and 10 seconds. As per the results in Table 3, the method using random forest with a measurement time of 10 seconds gives the highest identification rate of 94.6%. The F1 score under this condition was 0.942. The confusion matrix is shown in Fig. 5. Our proposed method was able to obtain more than 90% of the results for actual data. For same experimental condition, the YOLOv2tiny method alone was 88.7%.

Table 3. Identification rate of the purposed methods

<table>
<thead>
<tr>
<th></th>
<th>tree</th>
<th>Random forest</th>
<th>Linear SVM</th>
<th>non-linear SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>5s</td>
<td>90.7%</td>
<td>91.9%</td>
<td>88.8%</td>
<td>90.4%</td>
</tr>
<tr>
<td>10s</td>
<td>92.6%</td>
<td>94.6%</td>
<td>88.7%</td>
<td>92.1%</td>
</tr>
</tbody>
</table>

Figure 5. The confusion matrix of proposed method using random forest and 10 seconds sound measurement

In addition, the experiment was conducted by changing the model of YOLO. The results of YOLOv3 are shown in Table 4. As per the results in Table 4, the identification rate is 90.9%, which is lower than that of YOLOv2tiny. Further, it was verified with YOLOv4[10] and the identification rate did not exceed 90%. We think that it is compatible with the number of parameters of the model and the number of training data.

Table 4. Identification rate of proposed method using YOLOv3

<table>
<thead>
<tr>
<th>Identification rate</th>
<th>normal</th>
<th>abnormal</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>5s</td>
<td>90.9%</td>
<td>93.5%</td>
<td>88.1%</td>
</tr>
<tr>
<td>10s</td>
<td>88.4%</td>
<td>89.4%</td>
<td>87.2%</td>
</tr>
</tbody>
</table>
5. Conclusion

In this paper, we proposed the construction of an algorithm for an AI stethoscope that can make fairly the same decisions as a medical doctor. Our proposal was based on the method of using YOLOv2tiny for the wavelet transform results, and extends it to improve the performance using machine learning techniques. Rather than classify wavelet transform images solely with YOLOv2tiny, we improved the identification rate of heart sounds by performing random forest on the variables derived from the YOLOv2tiny and neural networks decision. Overall, from our experiments we found that the random forest is better, eliminated by the combination method and recording for 10 seconds. Future work will consider adding and utilizing the label of gray.

Acknowledgments:
In this experiment, audio data was collected with the cooperation of the Miyazaki AI Stethoscope Project. We appreciate the support from Miyazaki AI Stethoscope Project.

References

10. Alexey Bochkovskiy,Chien-Yao Wang,Hong-Yuan Mark Liao "YOLOv4: Optimal Speed and Accuracy of Object Detection" Computer Vision and Pattern Recognition, 23 April. 2020