Anomaly detection of lung sounds using DAGMM

Ryosuke Wakamoto
Graduate School of Sciences and Technology for Innovation, Yamaguchi University,
2-16-1 Tokiwadai, Ube, Yamaguchi 755-8611, Japan

Shingo Mabu
Graduate School of Sciences and Technology for Innovation, Yamaguchi University,
2-16-1 Tokiwadai, Ube, Yamaguchi 755-8611, Japan

Shoji Kido
Graduate School of Medicine, Osaka University,
2-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Takashi Kuremoto
Graduate School of Sciences and Technology for Innovation, Yamaguchi University,
2-16-1 Tokiwadai, Ube, Yamaguchi 755-8611, Japan
E-mail: b093vg@yamaguchi-u.ac.jp, mabu@yamaguchi-u.ac.jp
kido@radiol.med.osaka-u.ac.jp, wu@yamaguchi-u.ac.jp

Abstract

There are only a few small-scale benchmark datasets of lung sounds that are annotated for the training of machine learning. Therefore, we aim to build an anomaly detection system that only uses normal data that can be obtained more than abnormal data. We propose an algorithm that improves the Deep Autoencoding Gaussian Mixture Model (DAGMM), where various types of neural networks are applied to DAGMM as the compression networks. Experimental results show that the proposed methods obtain effective classification performance.

Keywords: lung sounds, deep learning, anomaly detection, DAGMM

1. Introduction

In recent years, deep learning algorithms based on large amounts of data have been proposed and applied in various fields[1]. In the field of image recognition, in particular, various applications have been made using convolutional neural networks (CNNs). Because of its availability, CNNs have been applied to the medical field where high accuracy is required.

For example, various approaches have been developed for medical images, such as feature extraction specific to a lesion or 3D modeling of an organ. Most of the research on computer-aided diagnosis using deep learning has been attempted for image diagnosis, which has greatly contributed to improve the accuracy of medical image diagnosis such as lung diseases[2] which are dealt with in this study. However, diagnosis with only images is not sufficient because, in addition to the limitations of qualitative diagnosis in imaging, there are multiple ways to diagnose lung diseases. In the diagnosis of lung diseases, the patient is interviewed and physically examined, and the condition of the disease is estimated to some extent, and respiratory function tests, image diagnosis, electrocardiography, and gait assessment are performed[3]. Because of this, the burden on the physicians making the diagnosis is heavy, requiring precise treatment and enough experiences.

As mentioned above, because doctors need to conduct various tests for diagnosis, it is necessary to develop diagnostic support techniques for not only visual information such as images, but also for other senses such as hearing. Therefore, we consider applying deep learning-based speech recognition to auscultation which is one of the diagnostic methods for lung diseases. If we can construct a system for discriminating sounds, we can improve the accuracy of diagnosis by combining it with other examination results such as imaging.

Although deep learning requires a large amount of sound data to learn the features of lung sounds, there are few small benchmark datasets with annotation (class labeling), which is necessary for machine learning. In addition, it takes a great deal of effort and time for each medical facility to acquire a sufficient amount of sound data. When training is performed with a small amount of data, it may not obtain generalized performance because it is difficult to capture essential features that are important for the diagnosis due to the large individual differences of lung sounds. There have been some research related to lung sounds classification, such as the analysis using histogram statistics[4] and the lung sounds classification using deep learning[5], however, the number of data to generate classification models is limited in both methods. Therefore, this study focuses on normal data, which is easier to obtain than abnormal data, and constructs a system for detecting abnormalities of pulmonary auscultation sounds using deep learning with the aim of capturing generalized features even with small data.

In this paper, we propose a deep learning algorithm for feature extraction of lung sounds that improves the Deep Autoencoding Gaussian Mixture Model (DAGMM)[6], which is an anomaly detection algorithm capable of learning feature extraction and clustering at the same time. Specifically, it uses Mel-Frequency Cepstral Coefficients (MFCC)[7] for the feature extraction and three types of networks (auto-encoders) based on: 1) CNN, 2) Long short term memory (LSTM), 3) Convolutional LSTM (C-LSTM) are applied to DAGMM as compression networks to design an algorithm for efficient feature extraction. C-LSTM has features of both CNN and LSTM, so it can consider the features of both peripheral and time series information. These three types of networks have been shown to be useful in discriminating lung sounds in previous studies, and since the standard Multi-Layer Perceptron (MLP) is used as a compression network in the conventional DAGMM, the above three networks are expected to show better performance in DAGMM. In our experiments, we compare the conventional DAGMM and the proposed methods, and better classification performance is obtained by improving the compression networks in DAGMM.

The paper is organized as follows. Section 2 describes the related work and Section 3 presents the proposed method. Section 4 presents the experimental conditions and results, and Section 5 presents conclusions and future work.

2. Related work

There are two types of research related to sound and speech: speech analysis and speech recognition. Speech analysis is a process of extracting features from the sampled sound data for speech recognition. Speech recognition is the process of identifying the target sound or speech based on the features obtained by the speech analysis. In addition to these processes, this section also describes the anomaly detection.

2.1. *Speech Analysis*

Speech analysis improves recognition accuracy by eliminating unnecessary information as noise as well as extracting only the features necessary for speech recognition. In this study, we use MFCC, a feature extraction method to extract the cepstrum by applying a filter based on the human hearing (Mel-filter bank), which separates the formant frequencies, which is necessary for speech recognition, from the pitch frequencies, which contains individual differences. Since Formant frequencies of MFCC appear in the low-frequency regions, important information for sound classification can be obtained by extracting the lower dimensional components of MFCC. Fig. 1 shows an example of MFCC that shows the 20-dimensional features of MFCC for 5-second of lung sounds. The horizontal axis of Fig. 1 represents the 20 segments of the 5-second data and the vertical axis represents the dimensions of MFCC.

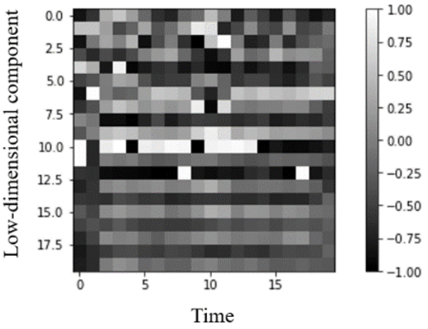
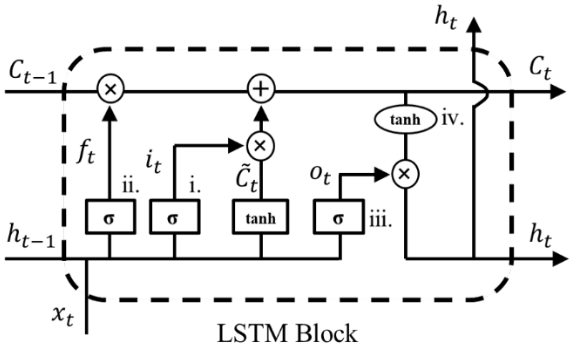


Fig. 1. MFCC with 20 dimensions

2.2. *Speech Recognition*

Before deep learning is actively studied, dynamic programming (DP) matching[8] and hidden Markov models (HMM)[9] were used for speech recognition, however, once deep learning was applied to speech recognition, Recurrent Neural Network (RNN) were applied and false recognition rate of speech was significantly decreased. However, it was difficult to learn long data stably due to the problem of gradient loss. Then, LSTM shown in Fig. 2 were used. LSTM has three gates: an input gate (i. in Fig. 2), a forget gate (ii. in Fig. 2), an output gate (iii. in Fig. 2), and a cell that holds past information. The problem of gradient loss is solved by using gates to select the information in the cells.



σ: Sigmoid function
C: Cell memory
h: Cell output
f: Forget gate
i: Input gate
o: Output gate

Fig. 2. Structure of LSTM

In addition, image recognition methods such as CNNs are often utilized in speech recognition[10]. CNNs are widely used in image recognition, but they also have excellent performance as a local feature extractor in speech recognition. Both CNNs and LSTMs are useful for feature extraction of lung sounds, and in the previous study, classification models using C-LSTM that possesses both CNN and LSTM features have shown higher accuracy[11].

2.3. *Anomaly Detection*

Anomaly detection is a method that creates a distribution of normal data by learning only normal data, then, regards the data that do not belong to the distribution of normal as abnormal data. In this section, we describe the outlier (anomaly) detection methods using machine learning. In general, unsupervised outlier detection methods execute feature extraction and distribution creation separately. For example, auto-encoder is first used to extract features from the input data, then clustering methods such as k-means and gaussian mixture model (GMM) are applied to create distributions. One of the problems in such learning structures is that the learning of feature extraction and clustering are executed separately. The features extracted independently from the clustering process may not be useful. In order to accurately detect outliers, a coherent feature extraction is required for clustering.

The structure of DAGMM is shown in Fig. 3. DAGMM is an outlier detection method that performs feature extraction and clustering simultaneously. The feature extraction is realized by a compression network using auto-encoder and the clustering is realized by combining an estimation network and GMM. The compression network outputs the encoded feature and the reconstructed data of the inputs, and the estimation network outputs the probability that the input x belongs to each cluster. In the compression network, input x is encoded to feature Z_c. Furthermore, the feature Z_c is decoded to obtain the reconstructed image x'. In the compression network, auto-encoder learns that x and x' become the same, where the reconstruction error Z_r is represented as follows.

$$Z_r = (d_1, d_2) = \left(\frac{\|x-x'\|_2}{\|x\|_2}, \frac{x \cdot x'}{\|x\|_2 \|x'\|_2}\right). \qquad (1)$$

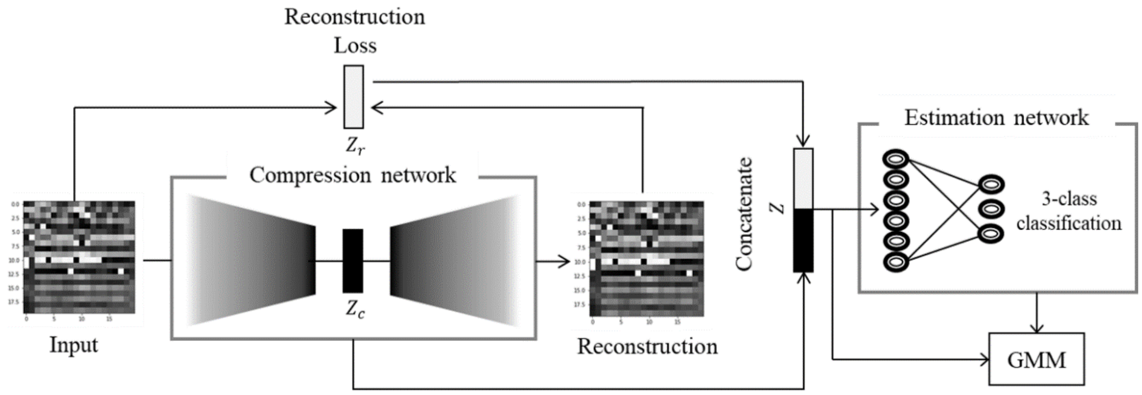


Fig. 3. Structure of DAGMM

By using the two measures d_1 and d_2, we can measure the reconstruction error from different perspectives. Then, the concatenated feature of Z_c and error Z_r generated by the compression network forms a new feature Z. With this as the input to the estimation network, the estimation network then outputs the affiliation probability $\hat{\gamma}$ of how well Z is matched to each cluster. Then, GMM is generated by the feature Z and the affiliation probability $\hat{\gamma}$. GMM requires three parameters: the mixture ratio phi, the mean matrix mu, and the covariance matrix $sigma$. The formula for each is as follows.

$$phi_k = \sum_{i=1}^{N} \frac{\hat{\gamma}_{ik}}{N}. \tag{2}$$

$$mu_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik} z_i}{\sum_{i=1}^{N} \hat{\gamma}_{ik}}. \tag{3}$$

$$sigma_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik}(z_i - mu_k)(z_i - mu_k)^T}{\sum_{i=1}^{N} \hat{\gamma}_{ik}}. \tag{4}$$

DAGMM calculates the energy of the feature Z after the generation of the GMM: when a data is located in the center of the distribution of the GMM, the energy of Z is small, and the energy of the off-centered Z is large. The energy function is represented as follows.

$$E_{(Z)} = -log\left(\sum_{k=1}^{K} phi_k \frac{exp\left(-\frac{1}{2}(z-mu_k)^T \sum_k^{-1}(z-mu_k)\right)}{\sqrt{|2\pi sigma_k|}}\right). \tag{5}$$

DAGMM can learn feature extraction suitable for clustering by the above learning structure. In addition, when outlier detection is performed, the energy of abnormal data is expected to be larger than that of normal data.

3. Proposed methods

In this study, we deal with an anomaly detection algorithm for lung sounds. Since previous studies in lung sounds have been concerned with the lack of data, an anomaly detection algorithm is suitable because normal data is easier to obtain than abnormal data and obtains generalized features even with a small number of data. In this study, the conventional DAGMM is regarded as the conventional method that realizes anomaly detection performing both clustering and feature extraction simultaneously.

From the next subsection, we propose a deep learning algorithm that improves the feature extraction of DAGMM for lung sounds. Specifically, we replace the compression network in DAGMM in Fig. 3 with various types of auto-encoders based on CNN, LSTM and C-LSTM. Since the three auto-encoders based on CNN, LSTM and C-LSTM have different structures and characteristics, the details of each structure are described one by one.

3.1. *Convolutional auto-encoder (CAE)*

The structure of a CAE is shown in Fig. 4. The input data is converted to an image as shown in Fig. 1 by applying MFCC to 5-second unlabeled sound data. Therefore, it is possible to perform convolutional processing on sound data as the same way as on images. CAE encodes the input by an Encoder consisting of two convolutional and two pooling layers, and then reconstructs the input by a Decoder consisting of two deconvolutional and two upsampling layers. The structure of the compression network with convolution and pooling captures the local features of MFCC making it easier to ignore the noise that is likely to have an influence on speech recognition.

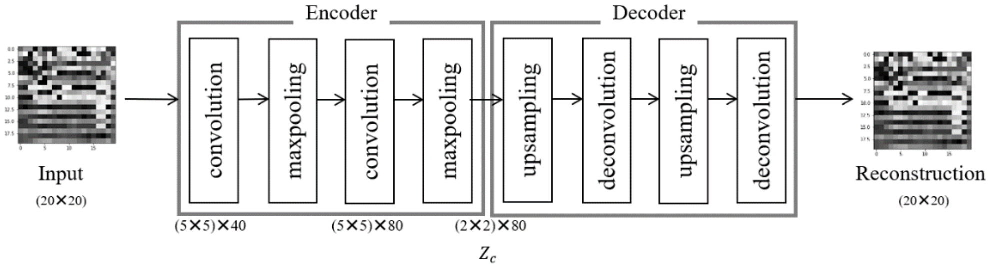


Fig. 4. Structure of CAE

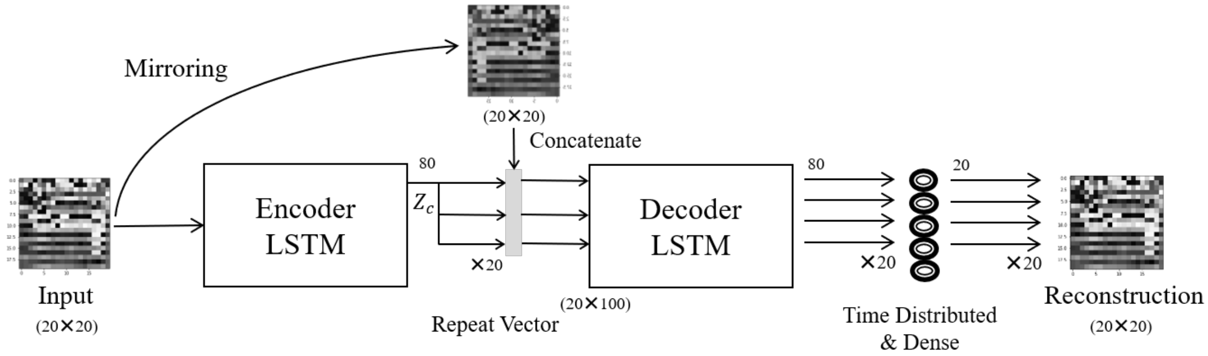


Fig. 5. Structure of LSTM-AE

3.2. *LSTM-based auto-encoder (LSTM-AE)*

The structure of a LSTM-AE is shown in Fig. 5 The LSTM-AE dealt with in this study is based on the structure proposed in the literature[12]. As in the case of CNN, we use unlabeled data to train Encoder LSTM and Decoder LSTM. In this case, the details of the network structure are as follows. First, Encoder LSTM outputs 80 features from the hidden layer of LSTM. Then, the input data to the Encoder LSTM (20×20) is mirrored and added to the generated features. This process is carried out by duplicating the generated one-dimensional features (80) 20 times to form a two-dimensional features (20×80) as shown in Fig. 5, and the generated two-dimensional features is combined with the mirroring input (20×20). This yields a new feature (20×100). Here, we explain the meaning of combining the generated features and mirroring input. Since the features (80) obtained by the Encoder LSTM has lost the information on the time axis, it is difficult to reconstruct the input image at the output layer. Therefore, the mirroring input serves as a flag for reconstructing the input from the features, and also has the effect of giving information on the inverse time axis. The combined features described above (20×100) are input to the Decoder LSTM row by row in time order (1×100), and the Decoder LSTM outputs the features (80) corresponding to each time. The output of one-dimensional features (80) is converted to a size (20) equivalent to the number of columns, i.e., 20 time segments of the input data, where the conversion is implemented by MLP. Here, the features (20) generated at all the time (20) are joined in sequence to obtain data with the same size as the input data (20×20). The purpose of an auto-encoder using LSTM is to make the output be the same as the input.

3.3. *C-LSTM-based auto-encoder (C-LSTM-AE)*

C-LSTM is a neural network that uses convolution layers when inputs to the sigmoid function of each gate of LSTM (i., ii., and iii. in Fig. 2) are calculated, and when the input to the tanh layer (iv. in Fig. 2) is calculated. Thus, LSTM can implement data processing considering peripheral information. The output shape is the same as the CNN, which is (the size of the feature map after

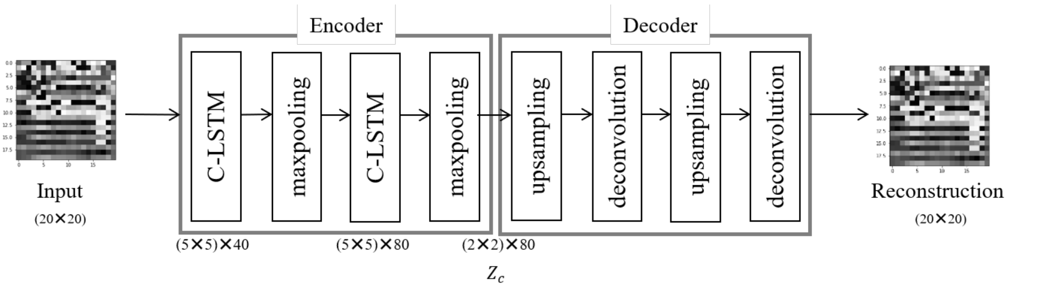


Fig. 6. Structure of C-LSTM-AE with convolutional decoder

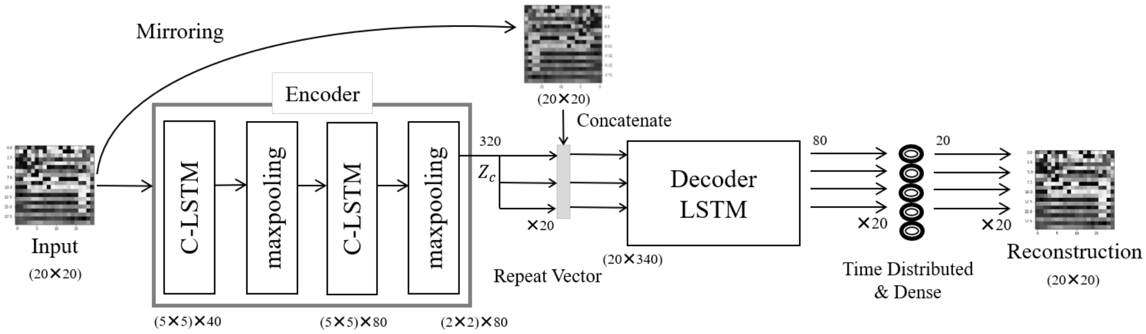


Fig. 7. Structure of C-LSTM-AE with LSTM decoder

convolution) x (the number of feature maps). We propose two structures of C-LSTM-AEs: the first is the C-LSTM-AE with convolutional decoder and the second is the C-LSTM-AE with LSTM decoder. The former structure is shown in Fig. 6 and the latter structure is shown in Fig. 7. The difference between the two is the structures of the decoders, which is described in detail in the next sub-subsection.

3.3.1. *Convolutional decoder*

In Fig. 6, the decoder consists of two deconvolutional layers and two upsampling layers, which is similar to the decoder used in the CAE. As the feature map of C-LSTM is two-dimension like CNN, the network structure is similar to that of CAE using deconvolution layers in the decoder. This structure is expected to capture local features while taking into account time-series information.

3.3.2. *LSTM-Decoder*

In Fig. 7, the decoder has the same structure as that used in the LSTM-AE, which converts the features (2×2×80) obtained by the C-LSTM's encoder into a one-dimensional features (320). In addition, the mirroring input is generated for combining the one-dimensional features and the mirroring input. Then, Decoder LSTM outputs a one-dimensional features (80) every time step, and it is converted by MLP into the array with the size (20) which is the same as the number of columns, i.e., 20 time segments of the input data. Therefore, the shape of the final output becomes the same as the input. Since we use C-LSTM and pooling layers in the Encoder, it is expected to learn the features considering time series information while mitigating the effects of noises in the input.

4. Experiments and results

4.1. *Data overview*

In this experiment, we used lung sounds data provided by Yamaguchi University Hospital, Japan to discriminate between discontinuous rale[3] (abnormal sounds) and normal sounds. The data used in this study are 5-second data, which were cut out from the data recorded during auscultation in a private room with the subjects in a seated position, and judged to be normal, fine crackle and coase crackle by the doctor. Here, the 12 areas were used for auscultation, that is, the chest area was firstly divided into four areas: upper and lower portions of both left and right lungs, respectively, and furthermore, the three areas : anterior, lateral, and posterior portions of the lungs are considered. Therefore, auscultation is done at totally 12 areas of lung, and the lung sounds of each area were recorded for more than three respiratory phases (inhalation plus exhalation) and more than about 15 seconds. Although the data were acquired from the three areas in the front, side, and back, the number of channels was 1 because the diagnosis was made independently from each area. The sampling rate was set to 11 kHz and the digital stethoscope (Power stethoscope, Starkey Japan) was connected to a voice recorder (ICD-MS1, Sony) and the sounds were recorded in a flash memory in 16-bit WAV file format. It is noted that noises caused during data acquisition depend on the way of breathing, subtle shifts in the stethoscope's position, and differences in the loudness of the sound. If an user is not skilled in handling the stethoscope, a lot of noises are picked up. In this study, the doctor who has enough experiences of auscultation recoded the lung sounds. The discontinuous rale in this study are discontinuous rales of short duration, and can be classified into two types: fine crackle and coarse crackle. Table 1 shows the characteristics of each auscultation sound including the normal sound. The characteristics of the abnormal intermittent sounds show sudden appearance, high sound pressure level, and very short duration. In addition to the paucity of data, the frequency bands of normal and abnormal sounds are overlapped, and the individual differences are large, thus, frequency analysis is difficult for the classification. In this study, to deal with the anomaly detection algorithm, two types of discontinuous rale are grouped together as one abnormal class.

4.2. *Experimental conditions*

The anomaly detection method was applied to two classes of auscultation sounds: abnormal and normal. The number of data of each class, the number of patients including their sex are shown in Table 2. 130 of the normal data were extracted as training data, and 10 normal data and 79 abnormal data that were not included in the training data were extracted as test data. In this study, we conducted an experiment with 14-fold cross validation. The evaluation index is Area under the Curve (AUC) which is area under the ROC curve, and the thresholds of the classification boundary of normal and abnormal when calculating ROC is determined by dividing the energy range of normal data into 11 ranges. In order to achieve high performance with this evaluation method, it is necessary to give lower energy to normal data and higher energy to abnormal data. In other words, it is necessary to obtain generalized features of normal data in training.

Table 1. Characteristics of lung sounds[4]

Class	Abnormal		Normal
	Coarse	Fine	
Frequency [Hz]	250-500	200-500 700-1000	150-600
Duration [ms]	10-15	Less than 5	-
Estimated diseases	Bronchitis, Pneumonia, Pulmonary tuberculosis	Interstitial pneumonia, Pulmonary fibrosis	-

Table 2. Overview of data of each class

Class	Abnormal	Normal
Number of data	79	140
Number of patients (male, female)	24 (16, 6)	12 (12, 0)

*Abnormal includes two of unknown sex patients.

For the input data, pre-processing of data sampling (sampling frequency of 2000 Hz), calculation of MFCC, and normalization (mean 0, variance 1) were applied prior to the training of each method. Because lung sounds are diagnosed mostly at frequencies below 2000 Hz[13] clinically, the sampling rate was set at 2000 Hz.

4.3. *Results*

Table 3 shows the AUC scores obtained by the conventional and proposed methods after 14-fold cross validation. From Table 3, we can see that the mean AUC

Table 3. AUC scores of each method obtained by 14-fold cross validation

	Methods				
Folds	DAGMM	DAGMM with CAE	DAGMM with LSTM-AE	DAGMM with C-LSTM-AE (conv)	DAGMM with C-LSTM-AE (LSTM)
1	0.9316	0.9791	0.8949	0.9911	**0.9924**
2	0.8665	0.8722	0.8930	0.8962	**0.9038**
3	0.9386	0.9633	0.9032	0.9741	**0.9766**
4	0.8633	0.8399	0.7241	0.8437	**0.8918**
5	0.8506	0.9101	0.6987	**0.9380**	0.9323
6	0.7949	0.8310	0.7633	0.8715	**0.8892**
7	0.8519	0.8696	0.8886	0.8930	**0.8949**
8	0.8962	0.9557	0.9570	**0.9899**	0.9861
9	0.9665	0.9804	0.9873	0.9880	**0.9930**
10	0.9146	**0.9867**	0.9608	0.9684	0.9633
11	0.8842	0.9006	0.9386	0.9342	**0.9462**
12	0.8709	0.9595	0.8842	**0.9728**	0.9646
13	0.8506	0.9177	0.8842	0.9449	**0.9532**
14	0.7411	0.9044	0.8823	**0.9399**	0.9272
Mean	0.8730	0.9193	0.8757	0.9390	**0.9439**

(conv): convolutional decoder, (LSTM): LSTM decoder

Table 4. Standard deviation of each methods by 14-folds cross validation

Method	DAGMM	DAGMM with CAE	DAGMM with LSTM-AE	DAGMM with C-LSTM-AE (conv)	DAGMM with C-LSTM-AE (LSTM)
Standard deviation	0.0560	0.0508	0.0841	0.0452	**0.0365**

score (0.9439) obtained by C-LSTM-AE with LSTM decoder is the best. The standard deviation obtained by each method is shown in Table 4, where C-LSTM-AE with LSTM decoder is also show the lowest standard deviation, therefore, the AUC scores are not deviated by a wide margin compared to other models in each validation.

Here, we discuss the usefulness of the proposed method. While the average AUC score of the conventional DAGMM without the proposed method is 0.8730, the average AUC scores of all the proposed methods are higher than that of the conventional DAGMM.

The results of the t-test of all methods are shown in Table 5. We can see from Table 5 that there is significant differences between the proposed methods and the conventional DAGMM, except for LSTM-AE, where the significance level is 5% (<0.05). It can be said that the performance of DAGMM can be improved by improving the compression network. Since the difference is not significant for LSTM but significant for the other proposed methods, it can be said that the peripheral information considered by convolution layers and the mitigation of the influence of noises by the pooling layers are related to the superiority of the performance in the feature extraction of lung sounds. When comparing C-LSTM-AE(conv) and LSTM-AE, C-LSTM-AE shows the better result with a significant difference, thus, the combination of both the peripheral information obtained by convolution and the time series information obtained by LSTM is useful in the feature extraction of lung sounds. Although time series information includes the important features in the lung sounds, it is susceptible to noises, thus it is necessary to use convolution and pooling together. When comparing C-LSTM-AE(LSTM) and C-LSTM-AE(conv), LSTM decoder is better because it recovers time series information more accurately.

Table 5. P-values of the results of t-test obtained by each method

	DAGMM	DAGMM with CAE	DAGMM with LSTM-AE	DAGMM with C-LSTM-AE (conv)	DAGMM with C-LSTM-AE (LSTM)
DAGMM	-	1.06×10^{-3}	4.48×10^{-1}	1.58×10^{-4}	1.43×10^{-5}
DAGMM with CAE	-	-	1.43×10^{-2}	2.29×10^{-4}	3.39×10^{-4}
DAGMM with LSTM-AE	-	-	-	1.76×10^{-3}	1.41×10^{-3}
DAGMM with C-LSTM-AE (conv)	-	-	-	-	1.20×10^{-1}
DAGMM with C-LSTM-AE (LSTM)	-	-	-	-	-

5. Conclusions

In this paper, various types of compression networks are proposed and evaluated by the AUC scores. The experimental results showed that utilizing convolution and pooling was effective in learning lung sounds, and utilizing time-series information together improved the performance of the model. In the future, we would like to compare these methods with other anomaly detection methods and improve the generalization abilities.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP19K12120.

References

1. Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature* **521**(7553), 2015, pp. 436-444
2. M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network, *IEEE Transactions on Medical Imaging* **35**(5), 2016, pp. 1207-1216
3. T. Ishihara, T. Kawashiro, T. Abe, K. Kikumaru, and M. Yonemaru, Auditory training by CD, Nankodo, 1993 (in Japanese)
4. Y. Igami, H. Shouno, and S. Kido, Identification of pulmonary auscultation data using statistics of adjacent waveform intervals, *The 9th IEEE Hiroshima Student Symposium (HISS2007)* **B53**, 2007 (in Japanese)
5. D. Bardou, K. Zhang, and S. M. Ahmad, Lung sounds classification using convolutional neural networks, *Artificial Intelligence in Medicine* **88**, 2018, pp. 58-69
6. B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, H. Chen, Deep autoencoding gaussian mixture model for unsupervised anomaly detection, *International Conference on Learning Representations*, 2018
7. B. Logan, Mel Frequency Cepstral Coefficients for Music Modeling, *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* , 2000
8. H. Sakoe, Two-level DP-matching-A dynamic programming-based pattern matching algorithm for connected word recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **27**(6), 1979, pp. 588-595
9. A. P. Varga and R. K. Moore, Hidden Markov model decomposition of speech and noise, *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing* **2**, 1990, pp. 845-848
10. R. Narasimhan, X. Z. Fern, and R. Raich, Simultaneous segmentation and classification of bird song using CNN, *Proceedings of the 2017 International Conference on Acoustics, Speech, and Signal Processing*, 2017, pp. 146-150
11. R. Wakamoto, S. Mabu, S. Kido, and T. Kuremoto, Lung Sound Classification Using Deep Neural Networks with Pre-training - Comparison of the Performance between CNN, LSTM and Convolutional LSTM -, *IEEJ Transactions on Electronics, Information and Systems* **140**(12), 2020, pp. 1402-1409 (in Japanese)
12. N. Srivastava, E. Mansimov, and R. Salakhutdinov, Unsupervised learning of video representations using LSTMs, *Proceedings of the 32nd International Conference on Machine Learning* **37**, 2015, pp. 843-852
13. V. Gross, A. Dittmar, T. Penzel, F. Schüttler and P. V. Wichert, The Relationship between Normal Lung Sounds, *American Journal of Respiratory and Critical Care Medicine* **162**(3), 2000, pp. 905-909