博士論文

Doctoral Dissertation

肺聴診音のコンピュータ支援診断に向けた
深層学習モデルおよび異常検知手法の構築に関する研究

(Study on the Construction of Deep Learning Models and Anomaly Detection Methods for Computer-Aided Diagnosis of Lung Sounds)

2025 年 3 月

March, 2025

福永　亮佑

Ryosuke Fukunaga

山口大学大学院創成科学研究科

Graduate School of Sciences and Technology for Innovation

Yamaguchi University

概要

医療診断において，ある患者の状態を判断するための検査方法は多岐にわたり，検査結果を総合的に判断する必要がある．そこで，医師の診断の補助を目的とし，疾患に関連する症状や病変の解析を行うコンピュータ支援診断（Computer-Aided Diagnosis ; CAD）技術が存在する．しかし，CAD は画像診断に対しての試みが大部分であり，特に音声信号処理の分野においては CAD 技術の十分な検討がなされていない．そのため，人工知能（Artificial Intelligence ; AI）を用いた音声信号処理技術の発展に期待が高まっている．CAD を音声信号に対して適用することができれば，診断にかかる医師の負担を軽減できる．そこで，本論文では，肺疾患の診断方法の 1 つである聴診に対し，AI を用いた音声信号処理を応用することを考える．聴診音は，正常音や異常音について周波数帯域が重なることから統計的な識別が困難であり，さらに，診察環境の様々な要因によって複数種類のノイズが混入し得るため，古典的な音声信号処理では扱いが難しいデータである．近年の深層学習技術の進歩により，このような音声信号に対しても，AI モデルによる高精度な判別が可能になると期待できる．

ここで，肺聴診音の CAD 技術を実現するためには，大きく 3 つの課題がある． 1 つ目の課題は，CAD 技術としてどのようなタスク（分類タスク，異常検知タスクなど）を設定し，AI モデルに解かせるかである．AI 技術を用いる場合，まず解くべきタスクの設定が重要であり，設定したタスクによって得られる精度や実用面の課題が異なる．2 つ目の課題は，学習データの少なさである．聴診音の収集やラベリングは実務上の負担が大きいことに加え，異常音の発生は稀であるため，モデルの学習に十分なデータを用意することが難しい．そのため，少ない学習データに過剰適合しない，汎化性能の高いモデルが求められる．3 つ目の課題は，識別に有効な肺聴診音の特徴量の選択である．学習データの少ない肺聴診音で CAD 向けの AI モデルを構築する場合，学習効率の高さと識別精度の高さを両立した AI モデルが求められる．その実現のためには肺聴診音の前処理と AI アーキテクチャの適切な組

合せによって，各タスクに重要な特徴が抽出できる方式が必要である．

本論文では，これらの課題に対する解決方法の提案及び検証を行った．第 1 章，第 2 章，第 3 章では，研究の背景や目的，本論文で用いる信号処理や AI モデル，使用するデータと評価指標についてそれぞれ説明する．

第 4 章では，学習データの少なさに対処することを目的として，メル周波数ケプストラム係数（Mel Frequency Cepstral Coefficient；MFCC）の拡張と事前学習を提案し，分類タスクの観点から 2 つ目と 3 つ目の課題の解決を図った．実験の結果より，2 つ目の課題に対する事前学習の効果が確認された．また，ノイズ軽減の仕組みを持つアーキテクチャの性能が高い傾向があり，3 つ目の課題について，周波数帯域やその時間変化の情報には識別上有用な特徴量が存在することが認められた．

第 5 章では，深層学習で，肺聴診音の識別において重要な特徴量をより効果的に処理することを目的として，転置 MFCC と独自の分類モデルを提案し，3 つ目の課題についてさらなる解決を図った．ここで，周波数帯域の情報が識別において特に有効であることや，その変化情報についても有用な特徴は含まれるなど，肺聴診音の特徴抽出において考慮すべき点を明らかにした．また， 1 つ目の課題である実用面について，分類モデルの予測結果の解釈性についても検討を行った．

第 6 章では，異常検知タスクに着目し，学習データの少なさに対処した．DAGMM（Deep Autoencoding Gaussian Mixture Model） と Efficient GAN （Efficient Generative Adversarial Network）の 2 種類の異常検知モデルを拡張した手法を提案し，2 つ目の課題の解決を図りつつ，1 つ目の課題である適切なタスク設定の検討を行った．実験では，DAGMM の拡張手法が高い AUC（Area Under the Curve）を達成し，異常検知モデルの性能が優れていることを示した．

第 7 章では，異常検知モデルの解釈性を高めることを目的として，位相的データ解析（Topological Data Analysis；TDA）と Isolation Forest を組み合わせた異常検知手法につ

いて提案し，異常検知における 1 つ目の課題の解決を図った．提案した異常検知手法は，深層学習を用いた従来の異常検知モデルと同等の性能を持ち，高い解釈性を持つことが確認された．

第 8 章では，本論文全体で得られた結果を整理し，結論を示している．本論文の成果は，肺聴診音の CAD 技術の実用化に向けた知見や，今後の精度向上の示唆を与えるものである．今後は，複数の聴診環境を跨いだ汎用的なモデルの構築や，診断支援の実用による医師の評価を得ることで，さらなる性能向上を図っていきたい．

Abstract

In medical diagnosis, there are a variety of examination methods to assess a patient's condition, and the physician must make a comprehensive assessment based on the results. To assist physicians in making diagnoses, Computer-Aided Diagnosis (CAD) technologies have been developed to analyze disease symptoms and abnormalities. However, the majority of CAD applications have focused on imaging diagnostics, and CAD technology has not been sufficiently explored in the field of acoustic signal processing. In recent years, there is growing anticipation for advancements in acoustic signal processing technologies utilizing Artificial Intelligence (AI). If CAD technology for acoustic signals can be realized, it could potentially reduce the workload of physicians during diagnosis. Therefore, in this thesis, we explore the application of AI-based acoustic signal processing to auscultation, one of the diagnostic methods for lung diseases. Lung sounds is difficult to be classified by statistical analysis because normal and abnormal sounds overlap in frequency bands, and they are prone to various types of noise introduced by different diagnostic environments. It is difficult for traditional acoustic signal processing methods to handle such complex data. However, with recent developments in deep learning, it is anticipated that AI models will enable highly accurate classification of these acoustic signals.

In realizing CAD technology for lung sounds, there are three major challenges. The first challenge is defining which tasks, such as classification or anomaly detection, the AI model should solve. When applying AI, it is crucial to first determine the task to be addressed. As the task definition will influence both the accuracy and the problems to be considered for practical use. The second challenge is the small number of training data. In addition to the significant practical burden of physicians to collect and annotate lung sounds, the occurrence of abnormal sounds is rare, making it difficult to obtain enough data for model training. Therefore, there is a need for models with high generalization performance that can avoid overfitting to small datasets. The third challenge is the effective feature extraction from lung sounds for classification or detection. When constructing AI models for CAD with limited auscultation sound data, it is essential to balance high training efficiency with high classification accuracy. To achieve this, it is necessary to develop a method that can extract important features for each task by appropriately combining the preprocessing of lung sounds with AI architectures.

In this paper, we propose and validate solutions to these challenges. Chapters 1, 2, and 3 explain the background and objectives of the study, the signal processing techniques and AI models used in this paper, as well as the data and evaluation metrics

employed.

In Chapter 4, we address the issue of limited training data by proposing the extension of Mel Frequency Cepstral Coefficients (MFCC) and pre-training techniques. These methods aim to solve the second and third challenges from the perspective of the classification task. The experimental results show the effectiveness of pre-training in addressing the second challenge. Additionally, architectures that incorporate noise reduction mechanisms show better performance, and with regard to the third challenge, it was found that features related to frequency bands and their temporal variations contain useful information for classification.

In Chapter 5, to more effectively generate the key features important for lung sound classification through deep learning, we propose the transposed MFCC and a custom classification model, further addressing the third challenge. This chapter clarifies that frequency band information is particularly effective for classification, and that temporal variations also contain valuable features that should be considered in feature extraction for lung sounds. Additionally, regarding the first challenge, which relates to practical application, we examine the interpretability of the predictions obtained by the classification model.

In Chapter 6, we focus on the anomaly detection task to address the issue of limited training data. We propose extended methods for two types of anomaly detection models: DAGMM (Deep Autoencoding Gaussian Mixture Model) and Efficient GAN (Efficient Generative Adversarial Network), aiming to solve the second challenge while also examining the appropriate task setting, which is the first challenge. In the experiments, the extended DAGMM method achieved a high AUC (Area Under the Curve), demonstrating the superior performance of the anomaly detection model.

In Chapter 7, to enhance the interpretability of anomaly detection models, we propose an anomaly detection method that combines Topological Data Analysis (TDA) with Isolation Forest, addressing the first challenge in anomaly detection. The proposed method demonstrated performance comparable to conventional deep learning-based anomaly detection models, while also offering high interpretability.

In Chapter 8, we summarize the results obtained throughout this study and present the conclusions. The findings of this research provide insights for the practical implementation of CAD technology for lung auscultation sounds and offer suggestions for future improvements in accuracy. Moving forward, we aim to further enhance performance by building a generalized model that can operate across multiple auscultation environments and by obtaining feedback from physicians through the practical use of the diagnostic support system.

用語説明

肺聴診音

肺聴診時に聴診器から採取した音を指す．これは，聴診器を介して採取した肺音に加え，心拍音，気道の反響音，聴診器のチューブ内の反響，チェストピースの摩擦音，その他の外部環境音など，聴診において診断に寄与しないノイズも全て含む音である．

肺音

患者の呼吸時に肺の気管から発せられた全ての音を指し，これは正常音と異常音のすべてを含む．本論文で扱う肺聴診音データについて，肺が発する音の特性について述べる場合はこちらを指す．

異常音

疾患を持つ患者に発生する固有の肺音グループを指す．本論文においては，断続性ラ音の一種である Coarse crackle と Fine crackle が該当する．

正常音

病状に関連した肺音を除く，すべての肺音を指す．

異常データ

肺聴診音データのうち疾患を持つ患者から採取された音声データを指す．音声中には正常音と異常音，ノイズが含まれる．

正常データ

肺聴診音データのうち疾患を持たない患者から採取された音声データを指す．音声中には正常音とノイズが含まれる．

目次

概要 1

用語説明 6

第 1 章　はじめに 14

1.1 本論文の背景 14

1.2 本論文の目的 15

1.3 本論文の構成 17

第 2 章　関連研究 20

2.1 音声信号の読み込み 20

2.2 音声信号の前処理 20

2.2.1 フーリエ変換 21

2.2.2 メル周波数ケプストラム係数（MFCC） 22

2.2.3 位相的データ解析（TDA） 24

2.3 深層学習モデルの構成要素 26

2.3.1 ニューロンモデル 27

2.3.2 多層パーセプトロン（MLP） 27

2.3.3 畳み込みニューラルネットワーク（CNN） 28

2.3.4 長・短期記憶（LSTM） 29

2.3.5 畳み込み LSTM（C-LSTM） 31

2.3.6 Transformer 32

2.4 学習モデル 34

2.4.1 自己符号化器（AE） 34

2.4.2 分類モデル 35

2.4.3 異常検知モデル 35

2.4.3.1 Deep Autoencoding Gaussian Mixture Model（DAGMM） 36

2.4.3.2 Efficient Generative Adversarial Network（Efficient GAN） 38

2.4.3.3 Isolation Forest 41
第 3 章　聴診音データと評価指標 42
3.1 使用したデータセット 42
3.2 評価指標 45
3.2.1 分類タスクの評価指標 45
3.2.1.1 正解率（Accuracy） 45
3.2.1.2 再現率（Recall） 45
3.2.1.3 適合率（Precision） 46
3.2.1.3 F 値（F1 score） 46
3.2.2 異常検知タスクの評価指標 46
3.2.2.1 AUC（Area Under the Curve） 47
第 4 章　深層ニューラルネットワークを用いた肺聴診音の識別 49
4.1 背景と目的 49
4.2 方法 50
4.2.1 CNN, LSTM, C-LSTM に対する事前学習と Fine-tuning（提案手法 1） 51
4.2.1.1 CNN の事前学習 51
4.2.1.2 LSTM の事前学習 53
4.2.1.3 C-LSTM の事前学習 54
4.2.2 MFCC 次元数の調整（提案手法 2） 56
4.2.3 動的特徴量（提案手法 3） 57
4.3 結果 58
4.4 まとめ 61
第 5 章　時系列ニューラルネットワークを用いた
肺聴診音の効果的な特徴抽出と識別 63
5.1 背景と目的 63
5.2 方法 65
5.2.1 転置 MFCC（提案手法 1） 65

5.2.2 Cross-encoding Transformer（提案手法 2） 67
5.3 結果 69
5.3.1 識別性能の比較 69
5.3.2 特徴マップ 72
5.3.3 Transformer による解釈性 73
5.4 まとめ 74
第 6 章　深層ニューラルネットワークを用いた肺聴診音の異常検知 76
6.1 背景と目的 76
6.2 方法 77
6.2.1 DAGMM の改良（提案手法 1） 78
6.2.1.1 CAE 79
6.2.1.2 LSTM-AE 79
6.2.1.3 C-LSTM-AE 80
(1) Convolutional Decoder 80
(2) LSTM Decoder 81
6.2.2 Efficient GAN の改良（提案手法 2） 82
6.2.2.1 Efficient GAN with GMM 82
6.2.2.2 Efficient GAN with GMM（C-LSTM） 84
6.3 結果 85
6.4 まとめ 88
第 7 章　解釈性を考慮した肺聴診音の異常検知 **90**
7.1 背景と目的 90
7.2 方法 92
7.2.1 トポロジー特徴量を用いた Isolation Forest による異常検知
（提案手法 1） 92
7.2.2 相関係数と音声 IDF によるスコアリング（提案手法 2） 95
7.2.2.1 相関係数 96

7.2.2.2 音声 IDF（Sound IDF） 96
7.2.2.3 相関係数と音声 IDF による異常スコアの算出 97
7.3 結果 98
7.4 まとめ 101
第 8 章　おわりに 102
8.1 本論文のまとめ 102
8.2 肺聴診音の音声信号処理に関する研究の今後の課題 105
参考文献 107
謝辞 112

図目次

Fig. 1: Frequency spectrum according to the number
of dimensions in MFCC extraction 27
Fig. 2: MFCC with 20 dimensions 24
Fig. 3: $\vec{v}(t)$ generated from lung sound using Takens Embedding Theorem 25
Fig. 4: Persistence diagram from lung sound 26
Fig. 5: Example of convolutional neural network 28
Fig. 6: Structure of LSTM 30
Fig. 7: Structure of Auto Encoder 34
Fig. 8: Structure of DAGMM 36
Fig. 9: Structure of Efficient GAN 38
Fig. 10: Anomaly detection algorithm of Efficient GAN 39
Fig. 11: Example of frequency information, time series information, and
a local feature contained in MFCC 44
Fig. 12: Example of anomaly score and ROC curve 47
Fig. 13: Pre-training of CNN 52
Fig. 14: Fine-tuning of CNN 52
Fig. 15: Pre-training of LSTM 53
Fig. 16: Fine-tuning of LSTM 53
Fig. 17: Pre-training of C-LSTM 55
Fig. 18: Fine-tuning of C-LSTM 55
Fig. 19: Grayscale MFCC images at multiple dimensions 56
Fig. 20: Structure of Cross-encoding Transformer 67
Fig. 21: Visualizing test data features with UMAP
obtained by each architecture 72
Fig. 22: Visualization of attention weights with Cross-encoding Transformer 73

Fig. 23: Structure of LSTM-AE 80
Fig. 24: Structure of C-LSTM-AE（conv） 80
Fig. 25: Structure of C-LSTM-AE（LSTM） 81
Fig. 26: Structure of Efficient GAN with GMM 83
Fig. 27: Structure of Efficient GAN with GMM
（C-LSTM-AE（LSTM Decoder）） 84
Fig. 28: Architecture of anomaly detection using Isolation Forest
with topological features 93
Fig. 29: Enhanced architecture of anomaly detection using Isolation Forest
with topological and time-series features 93
Fig. 30: Visualization of abnormalities with proposed method 1 100
Fig. 31: Visualization of abnormalities with proposed method 2 100

表目次

Table 1: Characteristics of lung sound..... 43

Table 2: Overview of data of each class..... 43

Table 3: Mean accuracy, recall and precision of pre-trained classification models ... 58

Table 4: Mean classification performance by deep learning architecture 70

Table 5: The results of one-sided tests (P-values) in the comparison
of deep learning architecture 70

Table 6: Mean AUC and standard deviation of anomaly detection models
in Chapter 6 by 14-fold cross validation 85

Table 7: The results of one-sided tests (P-values) in the comparison
of anomaly detection models 86

Table 8: Mean AUC and standard deviation of anomaly detection models
in Chapter 7 by 14-fold cross validation 98

第 1 章　はじめに

1.1　本論文の背景

近年，大量のデータを基にした人工知能（Artificial Intelligence ; AI）技術が様々な分野で提案，応用されている．特に，その一種である深層学習[1]は，大量のデータを用いて複雑なパターンや関係性を自動的に学習するアルゴリズムであり，その可用性から，今後は高い正確性が求められる医療分野でも導入がさらに加速していくと考えられる．

医療において AI 技術の研究が進んでいる分野の 1 つとして，デジタル画像処理が挙げられる．特に，畳み込みニューラルネットワーク（Convolutional Neural Network ; CNN）を用いた様々なデジタル画像処理技術が開発されており，医用画像に対して病変領域に特化した特徴抽出や，臓器の 3D モデル化など，様々なアプローチでコンピュータ支援診断（Computer-Aided Diagnosis ; CAD）技術の開発が行われている[2]．CAD 技術は医師の判断を補助することを目的とした，疾患に関連する症状や病変の解析技術である[3]．これらの技術が実用化すると，医師の実務において，より正確な診断と診断の効率化が期待できる．しかし，深層学習を用いた CAD 技術の研究は，画像診断に対しての試みが大部分であり，その他の技術カテゴリ――例えば本論文で対象とする音声信号処理分野への応用については十分な検討がなされていない．

ここで，CAD 技術の開発がデジタル画像処理だけでは充分と言えない理由は，医師の診断方法の多様性にある．医師が患者の疾患を診断する際，その検査方法は多岐にわたる．たとえば肺疾患の診断では，問診や身体診察を行い，ある程度病状を推定したうえで呼吸機能検査，画像検査，心電図検査，歩行試験などを行っている[4]．そのため，CAD においても音声信号処理技術や自然言語処理技術は必要不可欠であり，デジタル画像処理の発展のみでは軽減できない医師の負担が存在する．

CAD 技術のうち，デジタル画像処理が先んじて発展した理由にデジタル化と活用のための基盤整備が先行していたことが挙げられる．X 線 CT 装置の発明により，医用デジタル画像処理の検討が 1970 年代以降活発化していた[5]ことに加え，2000 年頃から画像保存通信システムである PACS（Picture Archiving and Communication System）[6]が普及し，医師の実務環境でデジタル画像を扱うようになったことが大きい．一方で，自然言語処理は電子カルテの普及などにより応用の期待はあったものの，各医療機関の記述表現やデータ構造が標準化されていなかった[7]ことから技術的な検討を行うハードルが高かったと考えられる．そして，音声信号処理に至ってはデジタル聴診器が普及し始めたのが近年ということもあり，聴診に関する音声データがデジタル化されることは一般的ではなかったことから，具体的な診断支援技術の開発が遅れている．

本論文では，特に診断支援技術の検討が遅れている音声信号処理技術に着目した．近年急速に発展している音声信号処理の技術を医療分野に応用することによって，さらなる CAD を実現できる可能性がある．例えば，肺聴診音の診断支援が実現されれば，聴診音から推定される疾患（気管支炎，肺炎，肺結核，間質性肺炎，肺線維症など）の診断について，診断の効率化や遠隔医療の実現が期待できる．

1.2　本論文の目的

本論文では診断支援の観点から，AI を用いた音声信号処理技術を，肺疾患の診断方法の 1 つである聴診に応用することを考える．聴診では，医師が聴診器を介して患者の肺音の識別を行い，肺音中に異常音が存在するか確認する．聴診時に採取される聴診音は，CAD 技術が確立されていないため，識別精度の高い AI モデルや，実用に向けた課題の整理が必要である．そこで，肺疾患患者が含まれるデータを対象として，医師の診断を支援できる AI モデルの開発・検証を行う．

ここで，支援診断への応用にあたって，重要な観点の 1 つは，聴診をどのようなタスクとして AI に解かせるべきかである．ここで述べるタスクには，たとえば分類や異常検知などが含まれる．これらを同一のデータで議論しているものが存在しないため，精度や実用面の課題をそれぞれ確かめる．特に，実用面においては解釈性が重要である．深層学習に代表される複雑な AI モデルは，その予測結果に対してどの変数が影響したかについては示されないことが一般的であり，特にモデルの挙動に悪影響を与える可能性がある多様なノイズが発生し得る聴診において，解釈性を持たない AI モデルでは診断の効率化に寄与できない可能性がある．

また，どのような診断支援の形においても，モデルの精度を高めることが重要となる．このとき，従来の AI 研究で大きな課題として挙げられるのがデータ数の少なさである．肺聴診音の識別に関連した研究として，ヒストグラム統計量を用いた解析手法[8]などが存在し，深層学習による肺聴診音の識別手法[9]なども提案されているが，いずれも分類モデルを構築するためのデータ数が少ないことが課題となっている．音声信号処理は，1 データあたりの特徴量数が多いことから少数データでは学習が難しいが，一方でモデルの学習のために高品質なアノテーションデータを数千件以上用意することは困難である．したがって，前処理，モデル，学習方法などの様々な観点から，少数データでも学習効率の高い AI モデルの構築方法を検討する必要がある．（本論文において，「少数データ」は数十件〜数百件程度の規模のデータを指す．）

これらの課題に対する検証を実施するうえで，共通の前提の課題となるのは AI モデル構築に向けた肺聴診音の特徴量抽出の方法論である．肺聴診音向けの AI モデル構築時の前処理，深層学習アーキテクチャ，AI モデル構造の組合せは一般的な方法論が確立されておらず，汎用性の高い特徴抽出法や，タスクの精度を高めるための特徴抽出について議論がなされていないのが現状である．そのため，本論文では上記の検証を行う際に多様な特徴抽出法の検討・考察を含め，肺聴診音データへの音声信号処理の応用において考慮すべき観点につ

いて示す.

1.3 本論文の構成

本論文は次のように構成される.

第 2 章では関連研究について説明する．AI を用いた音声信号処理を肺聴診音データで実施するにあたり，まずは音声データを解析しやすい形式に変換する必要がある．これらの処理の概観を述べ，音声データ変換手法として本論文が扱うフーリエ変換，メル周波数ケプストラム係数（MFCC），位相的データ解析（TDA）について説明する．次に，本論文の扱う深層学習モデルの主な構成要素となるアーキテクチャのうち多層パーセプトロン（MLP），畳み込みニューラルネットワーク（CNN），長・短期記憶（LSTM），畳み込み LSTM（C-LSTM），Transformer について，それぞれの仕組みと特徴について説明し，これらのアーキテクチャなどから構成される自己符号化器，分類モデル，異常検知モデルについても説明する.

第 3 章では，実験にあたって必要な準備を行う．本論文で用いるデータを説明し，データ形式や読み込み方法について説明する．また，各タスク（分類，異常検知）の評価方法となる正解率，再現率，適合率，F 値，AUC などの指標についても説明する.

第 4 章から第 7 章にかけて，本論文の提案内容の説明と実験・考察を行う．提案内容では，少数の肺聴診音データの CAD 技術（分類モデル，異常検知モデル）の構築に向けた有効な特徴量の探索と組合せを，一貫して追及する.

第 4 章では，肺聴診音の分類モデルを開発する．肺聴診音のデータ数が少ない課題に対して，事前学習や特徴抽出法の変更によるアプローチを提案し，複数の深層学習アーキテクチャによって効果を確認する．これにより，事前学習によってデータ数の課題が改善されることを示し，ノイズに頑健な深層学習アーキテクチャが肺聴診音の特徴抽出に効果的であ

ることを示す.

第 5 章では，肺聴診音の識別に効果的な特徴抽出についてさらなる検討を行う．近年多くの AI モデルに採用されている深層学習アーキテクチャである Transformer を比較対象に加え，深層学習モデルで学習する際に必要な前処理と，学習効率の高いモデル構造について確認する．具体的には，音声信号における前処理として一般的ではない転置処理を加え，さらに，第 4 章で示唆された肺聴診音が内包する音声の周波数・時系列の特徴について，より適切に考慮できるモデル構造を提案し，従来手法と併せて検証・比較を行う．これらによって，音声の周波数情報とその変化情報である時系列情報のそれぞれに肺聴診音を識別可能な特徴が存在することが示される．ここで，得られた実験の結果から肺聴診音に適した深層学習アーキテクチャを整理する．また，CAD の実用面において重要な，「予測結果の解釈性」についても焦点を当て，Transformer の Attention 機構を用いた予測結果の可視化についても試みる．

第 6 章では，肺聴診音の異常検知モデルを開発する．データの少ない状況をさらに効果的に解決する方法として AI モデルによる異常検知を提案し，Deep Autoencoding Gaussian Mixture Model（DAGMM），Efficient GAN などの既存の異常検知モデルについて，肺聴診音に特化した特徴抽出構造に拡張する．さらに，Efficient GAN に Gaussian mixture model（GMM）モジュールを追加し，少数データにおける異常検知タスクでの GMM の効果についても確認する．これらの実験により，肺聴診音への異常検知モデルの適用は効果的であり，実用では異常検知モデルが役立つ可能性があることを示す．また，第 4 章と同様に複数の深層学習アーキテクチャで効果を確認することで，肺聴診音の特徴抽出法の改善は，分類モデルと異常検知モデルいずれにおいても効果が期待できることを示す.

第 7 章では，解釈性を考慮した肺聴診音の異常検知モデルを開発する．AI を用いた異常検知は完璧な予測を保証することが困難であるが，一般的な異常検知モデルにおいては予測の根拠を提示することができず，実用的な診断支援の検討に向けて大きな障壁となって

いる．一方で，医療領域の音声信号に向けた解釈性の研究はほとんどなく，第 4 章で試みた可視化も実用上十分とは言えない．そこで，解釈可能な肺聴診音の異常検知モデルと，モデルの構築に必要な特徴量を検討する．具体的には，肺聴診音の認識において重要な特徴量が，周波数帯域別の音圧レベルの変化であることから，トポロジーを用いた情報圧縮による特徴抽出の応用を提案し，異常検知モデルの一種である Isolation Forest の学習を行う．さらに，時系列の依存性を考慮し，相関係数と独自で定義した音声 IDF を用いて拡張する．これらによって，高い解釈性を持つ異常検知モデルでも，深層学習を用いた異常検知モデルと同等の性能となることを示す．

第 8 章では，本論文の結論を述べる．結論では，本論文で得られた結果をまとめ，肺聴診音の音声信号処理に関する研究の今後の課題についても述べる．

第 2 章　関連研究

本論文の提案手法は，様々な音声信号処理，時系列データ処理，機械学習モデル・深層学習モデルを基盤としており，個々のアーキテクチャの詳細や特徴についての理解が重要であるため，本章で説明する．

2.1　音声信号の読み込み

空気の振動である音声は連続値であるため，音声信号として計算機で扱う場合には離散値に変換する必要がある．そのために行われる処理がアナログ - デジタル（AD）変換（Analogue-to-Digital conversion）である．アナログ - デジタル変換では標本化（サンプリング；Sampling）と呼ばれる処理によって一定間隔でデータを取り出し，そのデータに対して量子化を行うことによって計算機で扱う数値表現を得ることができる．ここで，標本化の時間間隔を決めるにあたっては，Shannon のサンプリング定理[10]が用いられる．次式に示すサンプリング定理では，音声波形$x(t)$が0(Hz)以上，W(Hz)未満の帯域にあるとき，$x(t)$を$\mathrm{T} \leq 1/(2\mathrm{W})(s)$ごとに標本化を行うことで，x(t)を完全に復元できることを示している．

$$x(t) = \sum_{n=-\infty}^{\infty} x(n\mathrm{T}) \frac{\sin\left(\frac{\pi}{\mathrm{T}}(t - n\mathrm{T})\right)}{\frac{\pi}{\mathrm{T}}(t - n\mathrm{T})}. \tag{2.1}$$

2.2　音声信号の前処理

聴診音などの音声信号を機械学習や深層学習などの AI モデルで扱うためには，一般的に前処理が必要となる．たとえば，実世界で保存された音声信号は式（2.1）の方法で読み込まれるが，このとき一般的な記録媒体のサンプリング周波数は 44.1kHz であり，1 秒間あ

たり 44,100 もの離散信号が含まれることになる．肺聴診音のような少数データにおいて，個々のデータがこの規模である場合，AI モデルの学習時に，1）学習データ量に対してモデルのパラメータが過剰になり，AI モデルに学習させることができない（次元の呪い），2）無関係な音声ノイズを学習してしまい局所解に陥る，などの問題が起こり得る．そのため，サンプリング後の音声信号から AI モデルの学習に必要な情報のみを抽出する必要がある．

本節では，前処理として本論文で取り扱う音声データ変換手法である，フーリエ変換，MFCC，TDA について説明する．

2.2.1 フーリエ変換

標本化された離散的な音声波形を周波数成分に分解するために，離散フーリエ変換（Discrete Fourier Transform；DFT）[11]を行う．離散フーリエ変換は以下の式で表される．

$$X_k = \sum_{n=0}^{N-1} x_n \exp(\frac{-\mathrm{j}2\pi nk}{\mathrm{N}}), \qquad k = 0, \dots, N-1, \tag{2.2}$$

ここで，Nは入力信号の長さである．x_nは入力信号のn番目のサンプルを，X_kは周波数領域におけるk番目の成分をそれぞれ表す．音声波形に離散フーリエ変換を行うにあたって，音声に周期性が成立する必要があるため，音声フレームごとに窓関数を適用する（以降，時間窓と呼称する）．本論文では，ハミング窓（hamming window）と呼ばれる窓関数を採用した．ハミング窓は以下の式で表される．

$$W_H(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{\mathrm{N}-1}\right), \quad 0 \le n \le \mathrm{N}-1, \tag{2.3}$$

ここで，N はハミング窓のデータ点の総数を表す．

また実用的には，窓関数をかけた後に離散フーリエ変換を高速に動作させ，計算時間を短縮する高速フーリエ変換（Fast Fourier Transform；FFT）が用いられる．本論文においても，フーリエ変換と呼称するものは高速フーリエ変換の処理を行っている．

本論文で扱う肺聴診音は音の違いに着目すること，つまり，音声信号に含まれる周波数の振幅スペクトルの包絡について，周波数帯域間の関係や時間変化を学習することが必要になる．フーリエ変換を用いることで，音声信号から周波数の振幅スペクトルの変化が抽出され，これは時間窓内における周波数帯域間の関係（周波数情報）を表す．さらに，これらの情報が時間窓ごとに計算されることで，周波数帯域の音圧レベルに対する時間変化を表す情報（時系列情報）となる．

2.2.2　メル周波数ケプストラム係数（MFCC）

メル周波数ケプストラム係数（Mel Frequency Cepstral Coefficient ; MFCC）は人の聴覚に合わせたフィルタ（メルフィルタバンク）を音声信号に適用し，ケプストラムを求める特徴抽出手法である[12].

適用するフィルタバンクは複数のバンドパスフィルタをオーバーラップして重ねることで形成される．人間の聴覚は高周波になるにつれて明瞭に区別することができなくなるため，その性質を反映した音の高さの尺度であるメル尺度を用いてフィルタバンクを作成する．メル尺度にはいくつか種類があるが，本論文では Fant の式を扱う．Fant の式によるメル尺度上の周波数f_{mel}は次式で表される[13].

$$f_{mel} = \frac{1000}{\log_{10} 2} \log_{10}\left(\frac{f_{Hz}}{1000} + 1\right). \tag{2.4}$$

ケプストラムの抽出は，2.2.1 で述べたフーリエ変換後に行われる．フーリエ変換によって振幅スペクトルに変換した後，その対数を取って再びフーリエ変換で時間領域に戻す処理である．MFCC ではフーリエ変換によって得られた振幅スペクトルにメルフィルタバンクをかけ，フィルタ後の振幅を足し合わせて対数をとり，離散コサイン変換（Discrete Cosine Transform ; DCT）を適用してケプストラムを求める．ケプストラムは，一般的には 12～20 次元程度まで抽出される．

ここで，DCT 前の対数振幅スペクトルを Fig. 1 に示す．Fig. 1 は MFCC の抽出次元数

に応じた周波数スペクトルの概形である．Fig. 1 より，MFCC の抽出次元数を増やすにつれて，より詳細な振幅スペクトルの包絡が表現されることがわかる．

上記の処理によって，MFCC はフォルマント成分と呼ばれる音声認識に必要な部分と，ピッチ成分と呼ばれる個人に依存する情報[14]を音声信号から分離することができる．MFCC では，DCT によってフォルマント成分が一定の低周波成分に集中するため，MFCC 変換後の低次成分を抽出することで，個人差に由来するピッチ成分を除去し，情報の次元圧縮を行うことができる．

Fig. 2 は，5 秒間の聴診音から MFCC を 20 次元抽出した例を示す．ここで，Fig. 2 の縦軸は MFCC の次元を表している．これは Fig. 1 で示したスペクトル包絡を DCT で復元した情報であり，各時間窓が内包する周波数帯域の情報を表す．横軸は 5 秒のデータを 20 分割した時間窓を表している．

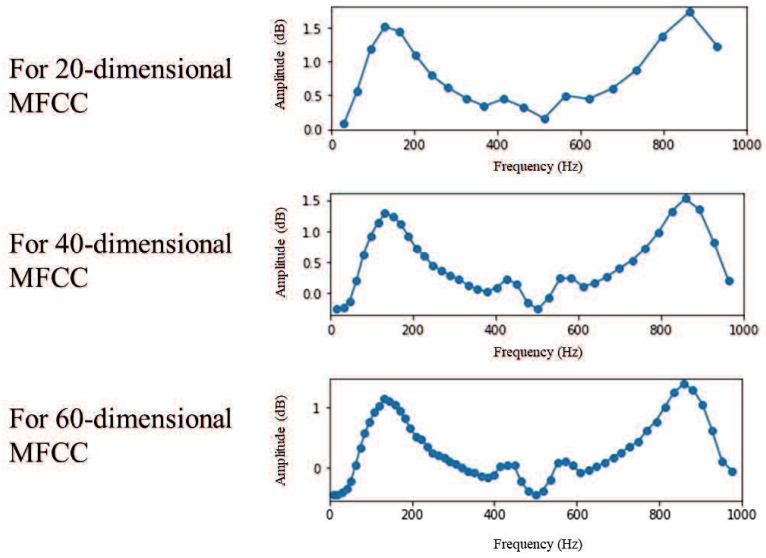


Fig. 1. Frequency spectrum according to the number of dimensions in MFCC extraction

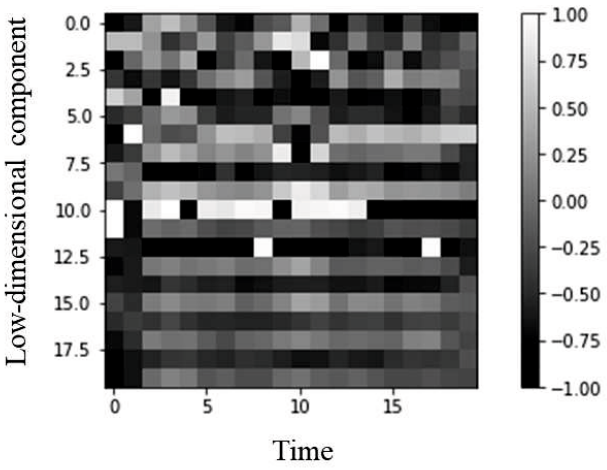


Fig. 2.　MFCC with 20 dimensions

2.2.3　位相的データ解析（TDA）

位相的データ解析（Topological Data Analysis；TDA）[15]は，データの持つ位相幾何学的な構造（形状，連結性など）に関する情報を抽出する手法であり，科学的な構造解析に代表される様々な分野で応用されている[16].

本論文で扱う肺聴診音のデータをはじめとして，音声信号は一次元の離散情報であるため，空間的な構造を持たず，TDA を適用した際に時系列のパターンを正確に捉えることが難しい．このような一変量時系列データに対して，位相的特徴を効果的に抽出する手法として Takens の埋め込み定理（Takens Embedding Theorem）[17]を用いる．Takens の定理は，遅延座標法を用いて時系列データを高次元の軌道空間に埋め込み，疑似アトラクタと呼ばれる TDA に適したデータ構造を得ることができる．ここで，得られる Takens 埋め込み $\vec{v}(t)$は次式で表される．

$$\vec{v}(t) = (y(t), y(t+\Delta t), \dots, y(t+(k-1)\Delta t)\,), \tag{2.5}$$

ここで，$\vec{v}(t)$は時間遅れΔtで作成されたk次元のベクトルであり，Δtとkはハイパーパラメーターである．本論文では，これらを最適化によって導出するアルゴリズム[18]を肺聴診音データに適用し，Δt=3，k=2 の値を設定した．上記の設定で肺聴診音の時間窓から得られた

Takens 埋め込み$\vec{v}(t)$の例を Fig. 3 に示す.

音声信号を$\vec{v}(t)$として再構築した後，トポロジカル特徴量を抽出する．本論文では，アトラクタから抽出できるトポロジカル特徴として，パーシステンス図（Persistence Diagram）[19]とパーシステンスエントロピー（Persistence Entropy）[20]を扱う．パーシステンス図は TDA で用いられる手法で，データ内に存在する空間的な穴や連結成分などの位相的構造が，スケールに応じてどのように変化し，消失するかを表した特徴量である．連結成分を表す 0 次ホモロジーH_0，穴（ホール）を表す 1 次ホモロジーH_1，空洞を表す 2 次ホモロジーH_2などのホモロジー群H_kを計算し，これらの位相的特徴が生成されるタイミング（birth）と消滅するタイミング（death）を記録する．Fig. 4 に，Fig. 3 の Takens 埋め込み$\vec{v}(t)$から得られたパーシステンス図を示す.

パーシステンスエントロピーは，パーシステンス図で得られた各次元の点について平均情報量（シャノンエントロピー）を計算したものであり，これによってパーシステンス図の情報を要約する．TDA では，ここまでに述べた Takens 埋め込み，パーシステンス図，パーシステンスエントロピーの一連の処理によって，音声信号に対しても位相的特徴の要約

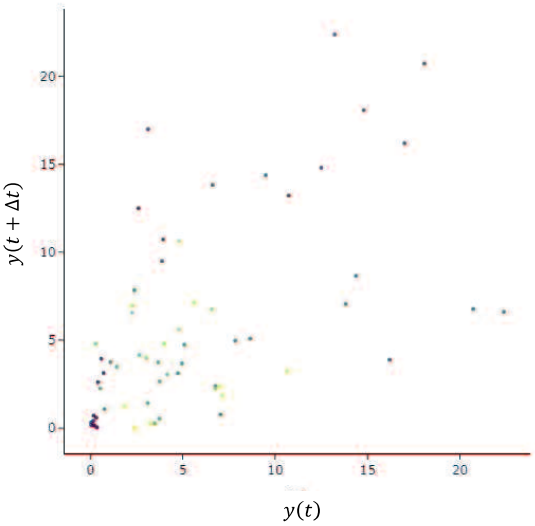


Fig. 3. $\vec{v}(t)$ generated from lung sound using Takens Embedding Theorem

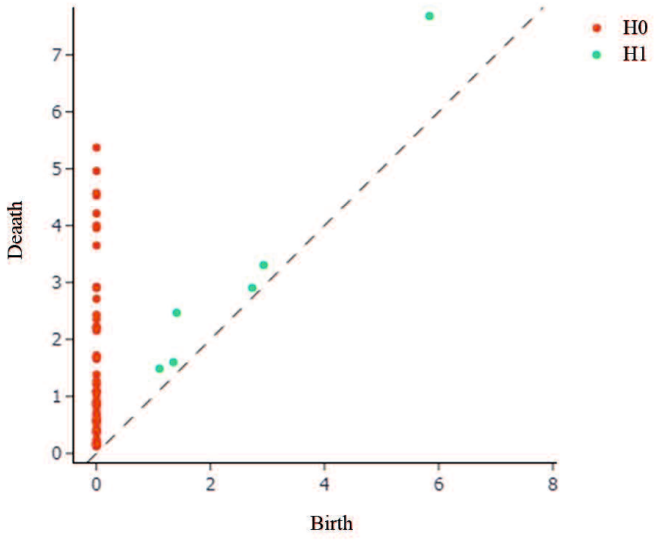


Fig. 4.　Persistence diagram from lung sound

表現を獲得することができる.

2.3　深層学習モデルの構成要素

2.2 で前処理された音声信号に対して，機械学習・深層学習などの AI モデルを用いて学習を行う. 本論文で扱う深層学習モデルは複数のアーキテクチャから構成されており，個々のアーキテクチャのアルゴリズムと処理特性は大きく異なる.

本論文が対象とする肺聴診音のような音声信号は，本質的には時系列の特徴パターンが重要であり，歴史的に動的計画法（DP）マッチング[21]や隠れマルコフモデル（HMM）[22]など，時系列データをモデル化できる手法が利用されてきた. 近年では深層学習の発展に伴い，RNN，LSTM，Transformer など，時系列情報の扱いに長けたニューラルネットワークの採用へとシフトしているのが現状である. 本節では，本論文で扱う深層学習モデルの主な構成要素となるアーキテクチャのうち MLP，CNN，LSTM，C-LSTM，Transformer について説明する.

2.3.1　ニューロンモデル

脳は情報の処理や伝達を行う神経細胞の集合体であり，1 つ 1 つの神経細胞（ニューロン）は，他のニューロンからの信号を結合しているシナプスを通して受け取り，入力信号の集積値が閾値を超えた際にニューロンが発火し，シナプスを通してさらに次のニューロンへと伝達される．神経細胞によるこの一連の信号伝達の様子をモデル化したものが W. McCulloch と W. Pitt によって提案されたニューロンモデル[23]である．ニューロンモデルは次式で表現される．

$$Z = f\left(\sum_{i=1}^{N} w_i\, x_i - \theta\right), \tag{3.1}$$

ここで，入力x_iは結合しているニューロンからの入力信号，w_iはニューロン同士を結合しているシナプスの結合荷重（重み），θは発火の閾値を表しており，N は信号を受け取るニューロンの数に対応している．また，$f(\cdot)$は活性化関数を指し，この関数の出力値がニューロンから出力される信号の値となる．本論文中で用いられる活性化関数$f(\cdot)$はシグモイド（sigmoid）関数，ReLU（Rectified linear unit）関数[24]，tanh 関数[25]，ソフトマックス（softmax）関数[26]などがある．

2.3.2　多層パーセプトロン（MLP）

複数のニューロンモデルを組み合わせ，非線形分類問題に対応したのが D. E. Rumelhart らによって提案された多層パーセプトロン（Multi-Layer Perceptron ; MLP）である[27]. MLP は入力層，中間層，出力層の三層からなっており，各層は前節のニューロンモデルから構成される. ニューロンが前の層のニューロンと完全に接続されること（全結合）により，ネットワーク全体で非線形である複雑なパターンについても学習する能力を持つ.

2.3.3　畳み込みニューラルネットワーク（CNN）

畳み込みニューラルネットワーク（Convolutional Neural Network；CNN）とは，主に画像認識に応用されるニューラルネットワークである[28]．画像認識においては，入力画像内で学習対象となる物体の位置がずれたり，多少変形していたりしても認識結果が変わらないことが望ましいことが多い．CNN は，内部に固有の畳み込み窓を保有し，そうした位置のずれや多少の変形に不変な内部表現を学習することができる．

CNN は一般に入力層（input layer），畳み込み層（convolution layer），プーリング層（pooling layer），ニューロンモデルで構成される全結合層（fully connected layer），出力層（output layer）によって構成される．CNN の構成例を Fig. 5 に示す．

ここで，畳み込み層では，多チャネルの画像に複数のフィルタを畳み込む計算を並行して行う．本論文で MFCC に対して CNN を適用するとき，初期のチャネル数はK=1 である．次元が$H \times W \times C$（Hは高さ，Wは幅，Cはチャネル数を表す）である入力データXについて，プーリング窓のサイズが$K \times K$，ストライド（移動幅）がs，パディングがpであるとき，畳み込み演算の数式は次式で表される．

$$Y_{i,j,c'} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{c=1}^{C} X_{i \cdot s+k-1, j \cdot s+l-1, c} \cdot W_{k,l,c,c'} + b_{c'}, \quad (3.2)$$

ここで，$Y_{i,j,c'}$は出力の位置(i,j)におけるチャネルc'を表し，$W_{k,l,c,c'}$は対応する重みを，$b_{c'}$はバイアス項を表す．

プーリング層では，畳み込みで抽出された画像のどの位置でフィルタの応答が強かった

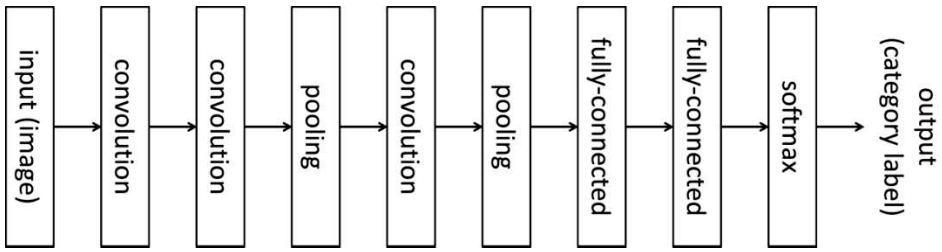


Fig. 5.　Example of convolutional neural network

かという情報を要約し，画像の特徴の微小な位置変化に対する応答の不変性を実現する．本論文で扱うプーリング手法には最大プーリング（Max pooling）[29]やグローバル平均プーリング（Global average pooling）[30]がある．Max pooling はプーリング領域中の最大値を，Global average pooling はチャネルごとの平均値を，それぞれ新たな特徴の値とするプーリング手法である．2 次元の入力データXについて，プーリング窓のサイズを$K \times K$，ストライドをsとしたとき，プーリング後の位置(p, q)における Max pooling は次式で表される．

$$\mathrm{MaxPool}(X)_{p,q} = \max_{i \in [p \cdot s, (p \cdot s)+K-1]} \max_{j \in [p \cdot s, (p \cdot s)+K-1]} X_{i,j} . \tag{3.3}$$

また，データXについて，次元が$H \times W \times C$であるとき，データXの各チャネルcにおける Global average pooling は次式で表される．

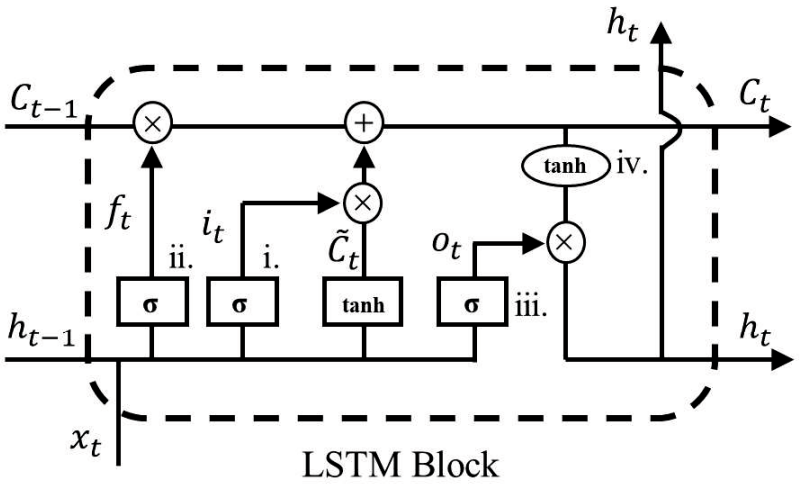
$$\mathrm{GAP}(X)_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,c} . \tag{3.4}$$

CNN では，ここまでに述べた畳み込みやプーリングなどの操作によって局所的な情報を捉えることができるほか，畳み込み層における重みの共有や一部の結合の固定によって，層数が同じ場合の MLP と比較して，学習すべき結合重み数（自由度）が大幅に減っていることから，学習データが少ない場合にも過学習を抑制し，より頑健な推論が期待できる．

2.3.4 長・短期記憶（LSTM）

長期的な依存関係を学習できる時系列ニューラルネットワークとして 1997 年に S. Hochreiter と J. Schmidhuber によって提案されたのが長・短期記憶（Long Short Term Memory ; LSTM）である[31]．LSTM の構造を Fig. 6 に示す．

ここで，σはシグモイド関数を表す．LSTM はセル状態の情報を削除・追加する機能を持ち，それらはゲートと呼ばれる構造によって制御される．LSTM は，入力ゲート（Fig. 6 の i.），忘却ゲート（Fig. 6 の ii.），出力ゲート（Fig. 6 の iii.）の 3 つのゲートと過去の情報を保持するセルを有する．これらの仕組みによって，LSTM では再帰的に時系列中の重要な情報を学習しており，深層学習を用いた信号処理分野で応用されている．



σ: Sigmoid function

C: Cell memory

h: Cell output

f: Forget gate

i: Input gate

o: Output gate

Fig. 6. Structure of LSTM

以下，LSTM のセル状態のパラメータの更新方法について説明する．最初に，パラメータの更新にあたってセルから捨てる情報の判定を，忘却ゲートと呼ばれるシグモイド層で行う．これは一時刻前のセル状態C_{t-1}の中の各数値をどれだけ忘却するかを決定する．忘却ゲートは次式で表され，h_{t-1}とx_tをもとに 0 から 1 の間の数値を出力する．

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)\,. \tag{3.5}$$

次に，セルで保存する新しい情報の判定を行う．これは，入力ゲートと呼ばれるシグモイド層で行う．入力ゲートは次式で表される．

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)\,. \tag{3.6}$$

入力ゲートによってどの値を更新するかの判定を行った後，tanh 層でセルに加える候補値のベクトル$\tilde{C}_t$を作成する．情報を更新するために，入力ゲートと tanh 層の 2 つをかけて用いる．なお，tanh 層は次式で表される．

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)\,. \tag{3.7}$$

次に，一時刻前のセル状態C_{t-1}にf_tをかけることによって忘却ゲートで判定したものを忘れさせた後，$i_t \cdot \tilde{C}_t$を加えてセル状態を更新する．この更新式は次式で示される．

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t\,. \tag{3.8}$$

最後に出力の判定を行う．セル状態の中から出力する部分の判定を出力ゲートと呼ばれるシグモイド層で行う．判定された部分のみを出力するために，セル状態に tanh を適用してセル中の値を-1 と 1 の間に圧縮する．これにシグモイド層をかけたものが実際の出力となる．出力ゲート値o_tと最終的な出力h_tは次式でそれぞれ計算される．

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)\,, \tag{3.9}$$

$$h_t = o_t \cdot \tanh(C_t)\,. \tag{3.10}$$

LSTM では上記の構造により，系列の情報を再帰的に入力し，学習上重要な情報を記憶することによって，時系列方向の特徴抽出に特化した学習を行うことができる．

本論文では，データの各時刻について双方向に学習を行う双方向 LSTM（Bidirectional Long Short Term Memory；BiLSTM）[32]についても扱う．BiLSTM は時間軸の順方向に隠れ層が結合する通常の LSTM と，時間軸の逆方向に隠れ層が結合する LSTM を結合したものである．順方向と逆方向の LSTM はどちらも独立であるため，BiLSTM の学習は通常の LSTM の学習と変わらず，一般的に識別精度が向上するとされる．

2.3.5　畳み込み LSTM（C-LSTM）

畳み込み LSTM（Convolutional LSTM；C-LSTM）は，X. Shi らによって提案された，CNN と LSTM を組み合わせたニューラルネットワークである[33]．C-LSTM の構造の概観は Fig. 6 の LSTM と同一であり，LSTM と異なる点は一部の計算方法である．具体的には，LSTM 内の 3 つのゲートと tanh の計算（式（3.5），式（3.6），式（3.7），式（3.10））における重みの乗算が畳み込み演算となっている．C-LSTM における忘却ゲート，入力ゲ

ート，出力ゲートの計算式を以下にそれぞれ示す.

$$f_t = \sigma\left(W_f * [h_{t-1}, x_t] + b_f\right), \tag{3.11}$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i), \tag{3.12}$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o), \tag{3.13}$$

ここで*は畳み込み演算を表している．C-LSTM における tanh 層の計算式を以下に示す.

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C). \tag{3.14}$$

通常の LSTM では，本論文で扱う MFCC などの 2 次元以上の情報を入力とした場合に，各行の情報を 1 時刻ごとに独立して入力するため，空間的な情報が破壊されてしまう．上記の式のように畳み込み演算に変更することにより，画像の状態を維持したまま LSTM への入力とすることができる．この構造によって，畳み込み演算によって得られる周辺情報と，それらの時系列方向への依存関係を LSTM で学習することができるようになる.

2.3.6 Transformer

Transformer は，2017 年に A. Vaswani らによって提案された深層学習モデルであり，自然言語処理のみならず画像処理など他の分野でも幅広く応用されている[34]．元の論文においては Encoder と Decoder によって構成される自己符号化器であったが，本論文において Encoder のみを使用した特徴抽出を行う．Transformer の Encoder は，主に Multi-Head self-Attention，Feed Forward Network によって構成される．また，Encoder への入力時に Positional Encoding が適用される.

まず，Positional Encoding の役割を説明する．Positional Encoding の役割は位置情報の付与である．Transformer の Encoder は，LSTM とは異なり再帰的な構造を持たないため，入力データの順番を考慮した学習ができない．そのため，次式で表される位置ベクトル$\vec{v}_i^{\,pos}$を入力ベクトルの時刻iに対応する要素に加算する.

$$\vec{v}_i^{\,pos} = \left(\sin\left(\frac{i}{T_1}\right), \cos\left(\frac{i}{T_1}\right), \ldots, \sin\left(\frac{i}{T_{d_{model}/2}}\right), \cos\left(\frac{i}{T_{d_{model}/2}}\right),\right). \tag{3.15}$$

次に, 位置情報を付与されたベクトルは Multi-Head self-Attention によって処理される. Multi-Head Attention は入力をQ, K, Vとして次式で表される.

$$MultiHead(Q,K,V) = concat(head_1, \dots, head_h)W^O\,, \tag{3.16}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right). \tag{3.17}$$

式（3.17）の$Attention$関数は Scaled Dot-Product Attention と呼ばれ，次式で表される.

$$Attention(Q,K,V) = sofmax\left(\frac{QK^T}{\sqrt{d}}\right)V\,, \tag{3.18}$$

ここで, Multi-Head self-Attention においてQ, K, Vは同じ値を入力とする. QK^Tによって内積を計算し，後述の正規化によってデータ内の各時刻の注目度（Attention）を計算している. dはデータの次元数d_{model}を指し，次元数による内積の発散を防いでいる. Multi-Head Attention は入力Q, K, Vに対して複数の行列の組（W_i^Q, W_i^K, W_i^V）をそれぞれ適用した線形変換を施すことで，学習を通して様々な角度から注目度の算出を行う.

後続の Feed Forward Network では，2 層の全結合モデル（間に ReLU 関数を挟む）が適用される. また，Feed Forward Network の前後においては，学習の安定化・高速化のため, ResNet[35]などで用いられる residual connection によって元の入力が出力に加算され, ベクトルの正規化処理（Layer Normalization）が行われる.

Transformer は上記の仕組みから，特に離れた要素同士の依存関係も頑健に学習できることが強みであり，近年発展している大規模言語モデル（Large Language Model；LLM）で採用されるのみならず, 大規模な音声認識モデル，もしくはマルチモーダル大規模言語モデル（Multilingual Large Language Model；MLLM）などにも採用されている[36].

2.4　学習モデル

本節では，本論文で用いる自己符号化器，分類モデル，異常検知モデルについて説明する．それぞれのモデルのうち，深層学習を用いたモデルについては 2.3 のアーキテクチャを用いて構成される．

2.4.1　自己符号化器（AE）

自己符号化器（Auto Encoder, AE）とは，教師なし学習モデルの一種であり，入力データを符号化するために特化した深層学習モデルである[27]．2.3 で説明したアーキテクチャを用いて，入力データを圧縮した特徴を獲得（符号化）し，その特徴から入力データを復元（復号化）するよう学習を行う．入力データを圧縮することによって，入力データを表現する普遍的な特徴を獲得できるため，クラス分類などの性能を向上させるための事前学習などにも応用される．自己符号化器の構造を Fig. 7 に示す．

自己符号化器の学習において，入力層から中間層への変換$Z_j = g(u_j)$を符号化，中間層から出力層への変換$y_i = h(u_i)$を復号化と呼び，Z_jをデータの特徴という[28]．なお，符号化器を Encoder，復号化器を Decoder と呼称する．

自己符号化器の学習は，通常の MLP とは異なり，入力データX^kと出力データY^kを比較した平均二乗誤差（Mean Squared Error ; MSE）などを用いて学習時の損失を計算する．

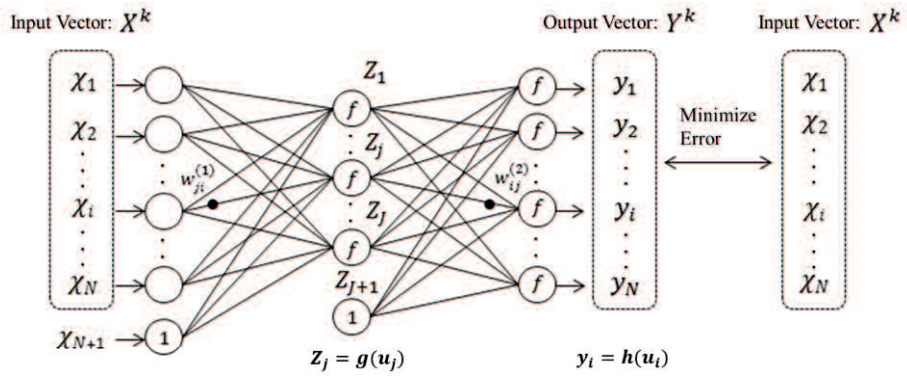


Fig. 7.　Structure of Auto Encoder

自己符号化器の目的は，入力ベクトルのパターンを出力で復元するネットワークの重みとバイアスを学習することによって，中間の層で入力データの特徴を抽出することである．

2.4.2　分類モデル

分類とは，データ固有のクラスラベル（正常，異常など）を各データに対して割り当てる手法の総称である．ここで，本論文で扱う分類モデルは教師あり学習を前提とする．教師あり学習の識別では，事前に各データが所属するクラスラベルを定義し，そのクラスラベルに応じた学習と評価を行う．

本論文では基本的な構成として，2.3 で説明したアーキテクチャに対して全結合層（fully connected layer）を追加することで分類モデルを構築した．出力層の全結合層に使用されるニューロンの数はクラスラベルの数と等しく，最終的な分類モデルの出力は softmax 関数によって各クラスラベルに属する確率値に変換される．

2.4.3　異常検知モデル

異常検知とは，正常データのみを学習して正常データの分布を作成し，分布に属さないデータを異常データ（外れ値）とみなす手法の総称である．異常検知モデルは，異常として検出することが望ましいデータの多様性が学習データで確保できない場合に取り得る，一般的なアプローチである．本論文で扱う肺聴診音の「異常」の性質は既知であるが，データ数の少なさが課題であることから，異常グループを正確に検知するための十分な学習ができない可能性がある．そのため，正常ではないものを異常として検出するアプローチは，本論文で扱う異常検知においても有効であると考えられる．ここでは，本論文で扱う異常値検出手法として，DAGMM，Efficient GAN，Isolation Forest を説明する．

2.4.3.1 Deep Autoencoding Gaussian Mixture Model（DAGMM）

Deep Autoencoding Gaussian Mixture Model（DAGMM）は，特徴抽出とクラスタリングを同時に行うことを特徴とする外れ値検出手法である[37]．従来の一般的な教師なしの外れ値検出手法では，特徴抽出と分布生成を独立して行っていた．例えば，入力データから特徴量を抽出するために自己符号化器を用い，次に k-means やガウス混合モデル（GMM）などのクラスタリング手法を用いて分布を作成する方法がある．このような学習構造の問題点の 1 つは，特徴抽出とクラスタリングの学習が別々に実行されることである．具体的には，自己符号化器から抽出された特徴量が「入力情報を復元するための潜在表現の獲得」にのみ特化しているため，クラスタリングに適した特徴となっているとは限らず，クラスタリング時に正常データと異常データの分布が重なってしまい，良い識別性能が得られない場合があることが挙げられる．

ここで，DAGMM の構造について説明する．DAGMM の構造の概観を Fig. 8 に示す．特徴抽出は自己符号化器を用いた圧縮ネットワーク（Compression network）で実現し，クラスタリングは推定ネットワーク（Estimation network）と GMM を組み合わせて実現する．

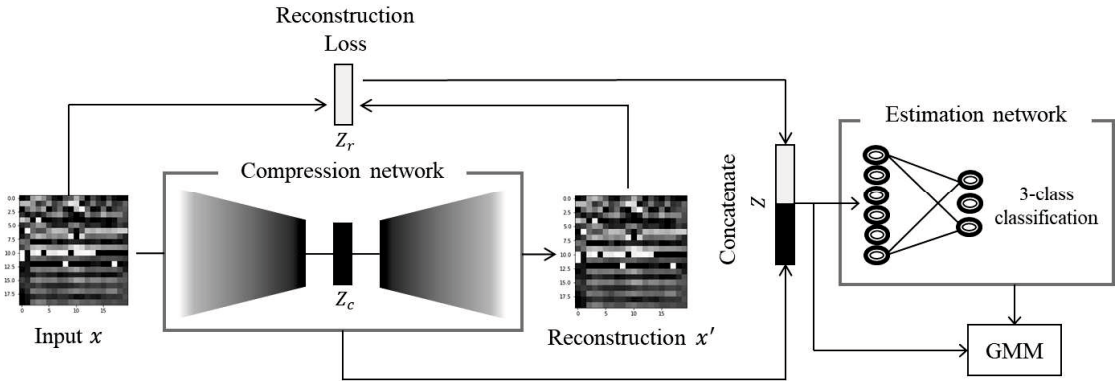


Fig. 8.　Structure of DAGMM

DAGMM の学習フローについて説明する．まず，入力xは最初に圧縮ネットワークに通され，特徴量Z_cに符号化される．さらに，Z_cをデコードして再構成画像x'を得る．入力xと再構成画像x'をもとに，次式で再構成誤差Z_rを算出する．

$$Z_r = (d_1, d_2) = \left(\frac{\|x - x'\|_2}{\|x\|_2}, \frac{x \cdot x'}{\|x\|_2 \|x'\|_2} \right). \tag{3.19}$$

そして，圧縮ネットワークで生成された特徴量Z_cと誤差Z_rを連結したものが新たな特徴量Zとなる．これを推定ネットワークへの入力とし，推定ネットワークはZがどのクラスタに所属するかの確率$\hat{\gamma}$を出力する．そして，特徴量Zと所属確率$\hat{\gamma}$から各クラスタについて混合比φ，平均行列μ，共分散行列σの 3 つを計算し，混合ガウス分布を生成する．クラスタkにおけるφ_k，μ_k，σ_kの計算式はそれぞれ以下で与えられる．

$$\varphi_k = \sum_{i=1}^{N} \frac{\hat{\gamma}_{ik}}{N}, \tag{3.20}$$

$$\mu_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik} Z_i}{\sum_{i=1}^{N} \hat{\gamma}_{ik}}, \tag{3.21}$$

$$\sigma_k = \frac{\sum_{i=1}^{N} \hat{\gamma}_{ik} (Z_i - \mu_k)(Z_i - \mu_k)^T}{\sum_{i=1}^{N} \hat{\gamma}_{ik}}, \tag{3.22}$$

ここで，kはクラスタの番号，Nはデータ数，$\hat{\gamma}_{ik}$はi番目のデータがクラスタkに所属する確率，Z_iはi番目のデータの特徴量である．

DAGMM は，上式の計算後に特徴量Zのエネルギーを計算する．データのZが，GMM の分布の中心に位置する場合にはエネルギーは小さくなり，分布から外れるほどエネルギーは大きくなる．エネルギー関数は次式で表される．

$$E_{(Z)} = -\log \left(\sum_{k=1}^{K} \varphi_k \frac{\exp\left(-\frac{1}{2} (z - \mu_k)^T \sum_k^{-1} (z - \mu_k) \right)}{\sqrt{|2\pi\sigma_k|}} \right). \tag{3.23}$$

また，DAGMM は目的関数が小さくなるように学習を行う．ここで，DAGMM の目的関数は，次式のように表される．

$$V = \frac{1}{N} \sum_{i=1}^{N} \|x_i, x_i'\|_2^2 + \frac{\lambda_1}{N} \sum_{i=1}^{N} E_{(Z)} + \frac{\lambda_2}{N} \sum_{k=1}^{K} \sum_{j=1}^{d} \frac{1}{(\sigma_k)_{j,j}}, \tag{3.24}$$

右辺第 1 項は入力データxと再構成画像x'の L2 ノルム，第 2 項は式 (3.23) のエネルギー，第 3 項は分布の対角要素を 0 としないための正則化項である．また，Kはクラスタ数，dは特徴量Zの次元数を表し，本論文では係数λ_1＝0.1，λ_2＝0.0001 に設定している．

異常検知時には正常データの分布を基に肺聴診音データのエネルギーを計算する．DAGMM は正常データで学習されるため，異常データを入力した場合に特徴量や再構成誤差，予測確率が正常データと乖離し，エネルギーが正常データよりも大きくなることが期待される．

2.4.3.2 Efficient Generative Adversarial Network（Efficient GAN）

Efficient Generative Adversarial Network（Efficient GAN）は，H. Zenati らによって 2018 年に提案された外れ値検出手法である[38]．Efficient GAN の構造を Fig. 9 に示す．DAGMM は正常データを可能な限りコンパクトな分布に落とし込む一方，肺聴診音のような微細な音の変化について，正常との違いを見出せずに異常と判定できない可能性がある．そこで，疑似的な異常データを生成し，それを見分ける学習を繰り返す異常検知モデルである Efficient GAN についても検証を行った．

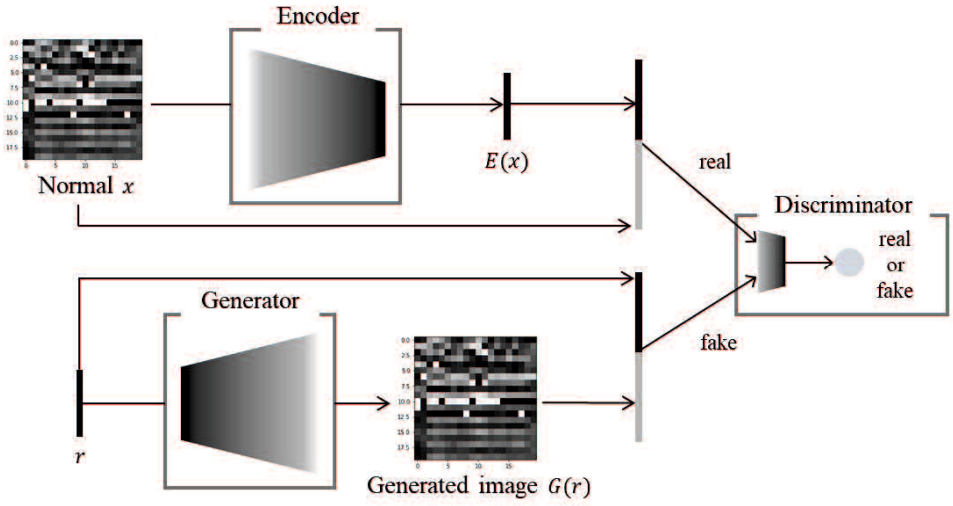


Fig. 9.　Structure of Efficient GAN

Efficient GAN は生成モデル BiGAN（Bidirectional Generative Adversarial Network）[39]の構造をベースとした学習モデルであり，正常データに似たデータを生成することを学習の目的としている．Fig. 9 に示すように，Efficient GAN は，Encoder，Generator，Discriminator の 3 つのネットワークから構成される．ここで，Efficient GAN の学習方法について説明する．まず入力は，正常データxとランダム（一様乱数に従う）ノイズrの 2 つである．Fig. 9 上段より， xは Encoder によって特徴量$E(x)$と結合される．この結合された特徴量$[x, E(x)]$は，本物のデータとして Discriminator で識別される．一方 Fig. 9 下段より，ランダムノイズrは，r が Generator によって変換された$G(r)$と結合される．結合された特徴量$[r, G(r)]$は，偽物（フェイク）のデータとして Discriminator で識別される．このとき，Discriminator は本物とフェイクについて正しく識別できるよう学習を行う．一方で，Encoder と Generator は Discriminator を騙すような学習を行う．これらの学習の結果，rと$E(x)$はそれぞれ類似した特徴量となり，$G(r)$は正常データxに類似したデータとなっていく．この学習で訓練された Encoder，Generator，Discriminator は，正常データについてのみ変換が行えるようになるため，これらのネットワークを用いて異常検知を実現する．

異常検知時には，Efficient GAN は異常スコア（Anomaly Score）の算出を行う．Efficient GAN の推論アルゴリズムの概観を Fig. 10 に示す．Fig. 10 より，まず Encoder を用いて，テスト用の入力データxを符号化した特徴量$E(x)$を得る．その後，xと$E(x)$のペアを

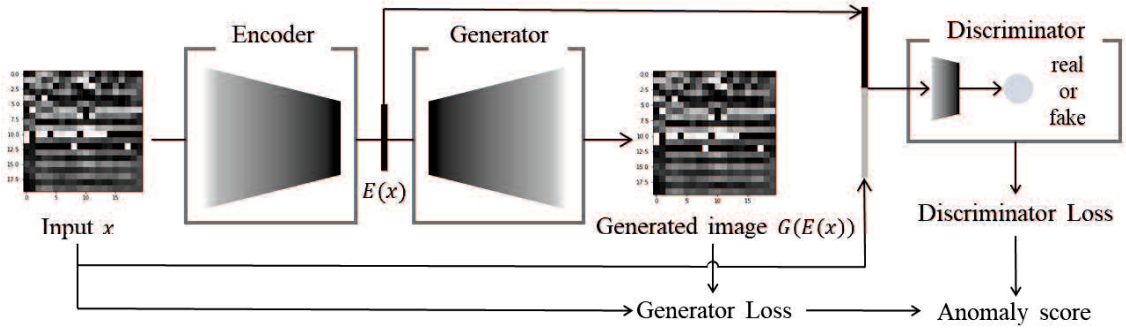


Fig. 10.　Anomaly detection algorithm of Efficient GAN

Discriminator で識別させ，Discriminator Loss を得る．一方，$E(x)$を Generator に入力し，再構成した復元データ$G(E(x))$を得る．そして，xと$G(E(x))$から Generator Loss を算出し，Discriminator Loss と Generator Loss から Anomaly Score を得る．Anomaly Score は Efficient GAN における最終的な出力である．以下，Anomaly Score $A(x)$，Generator Loss $L_G(x)$，Discriminator Loss $L_D(x)$の計算式をそれぞれ示す．

$$A(x) = \alpha L_G(x) + (1-\alpha)L_D(x)\,, \tag{3.25}$$

$$L_G(x) = \|x - G(E(x))\|_1\,, \tag{3.26}$$

$$L_D(x) = -\log\Big(D\big(x, E(x)\big)\Big)\,, \tag{3.27}$$

ここで，$A(x)$ の値は大きいほど，入力データxが異常であることを表している．αは任意の係数であり，本論文では$\alpha = 0.5$とした．また，式（3.27）では入力データxと$E(x)$を結合して Discriminator で識別し，正常クラスとみなした確率について交差エントロピーを計算している．これにより，Discriminator が正常データとみなすほど$L_D(x)$は小さい値を取り，異常データでは推論時に$A(x)$が大きくなることが期待される．

ここで，$A(x)$によって異常スコア（Anomaly Score）を得ることが可能となる理由について述べる．再構成損失$L_G(x)$は入力データxと，それを Generator で再構成した$G(E(x))$の L1 ノルムである．異常データを入力とした場合，Encoder による符号化が学習したデータと異なるものとなるため，Generator でデコードした際の復元の精度が低くなる．そのため，異常データにおける再構成損失$L_G(x)$は大きくなることが期待される．一方，識別損失$L_D(x)$は，入力データxとそれを符号化した$E(x)$のペアに対して Discriminator で識別した際の確率値と，本物と判定する場合のクラスとの交差エントロピー損失σである．異常データを入力とした場合，Discriminator は正常データの識別を学習しているため，Discriminator は偽物であると識別する可能性が高い．このため，異常データにおける識別損失$L_D(x)$についても大きくなることが期待できる．これら 2 種類の損失はどちらも異常なデータに対して高い

値を返すため，これらに係数をかけて足し合わせた異常スコア$A(x)$によって外れ値検出が可能となる．

2.4.3.3 Isolation Forest

Isolation Forest は，F. T. Liu と Z. Zhou によって提案された，異常検知のための二分決定木を用いたアンサンブル学習アルゴリズムである[40].

Isolation Forest は,「異常データは少数かつ他の特徴量と分離されている」という考えに基づき，ランダムに選択された特徴量に対して，データを二分決定木で複数回に分けて領域分割することにより，異常を検出する．異常データは通常のデータよりも少ない分割回数で隔離されることが多いため，各決定木でのパス長が基準値よりも短い場合，異常と判定される．

本論文では，1）計算資源が限られた環境や少数の特徴量・データにおいても効率的に動作すること，2）二分決定木に基づくため異常点検出プロセスや決定ルールの解釈が比較的容易であること，などの理由から Isolation Forest を採用した．ただし，Isolation Forest では前述の仕組みから時間的な依存関係を考慮しないため，本論文で使用する MFCC などの時系列の情報が表現される特徴量に対しては，Isolation Forest を用いることは一般的でない．

第 3 章　聴診音データと評価指標

3.1　使用したデータセット

本論文では，山口大学医学部附属病院から提供された肺聴診音データを使用した．本論文で呼称する「肺聴診音データ」とは，被験者を座位にして，できるだけ周囲の雑音が入らない個室で肺聴診を行った際に録音された肺聴診音のうち，医師が正常な肺音（Normal），水泡音（Coarse crackle ; Coarse），捻髪音（Fine crackle ; Fine）と判断した部分を切り取った各 5 秒のデータを指す．Coarse crackle と Fine crackle は，いずれも肺音の異常音の一種である断続性ラ音に属する．断続性ラ音の特徴として，突発的に出現すること，高い音圧レベルを持っていること，持続時間が非常に短いことなどが挙げられる[4].

ここで, 肺聴診を行う部位は, 左右両肺をそれぞれ上下に二分し, さらに, それぞれ前面, 側面，後面で取得したため合計 12 部位である．各部位とも 3 呼吸相（吸気＋呼気）以上，約 15 秒以上を録音している．このとき，チャネル数は 1，サンプリングレートは 11kHz として，デジタル聴診器（パワー聴診器，スターキージャパン製）をボイスレコーダー（ICD-MS1，ソニー製）に接続し，16bit の WAV ファイル形式で保存した．

データの取得時に混入するノイズは，呼吸の指示や聴診器への力の加え方により異なり，たとえば聴診に熟練していない医師の場合，聴診中に聴診器がずれてしまうことで患者との接触部であるチェストピースと肌とのわずかなこすれが発生し，雑音を多く拾ってしまうことがある．一方，本データの取得は本問題を熟知する医師によって行われた．

Table 1 は各肺音の特徴を表している[4]．肺聴診音のデータ数の少なさに加えて，正常な肺音と異常な肺音は周波数帯域の重複や個人差があるため，周波数解析で分析を行うことは難しい．さらに，本論文で扱う肺聴診音データは，前述のチェストピースの摩擦音の他にも，心拍音，気道の反響音，聴診器のチューブ内の反響，その他外部環境音（足音，会話）

などのノイズが微量に含まれており，これらの音声が分離不可能な形で保存されている．これらのノイズの存在も，周波数解析が難しい要因の 1 つである．

本論文では，異常検知モデルの学習と評価においては，Coarse crackle と Fine crackle のデータをまとめて異常データ（Abnormal）としてグループ化している．各クラスのデータ数と，データを取得した被験者数患者数，性別，疾患名を Table 2 に示す．

Table 1.　Characteristics of lung sound

Class	Abnormal		Normal
	Coarse crackle	Fine crackle	
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
	Coarse crackle	Fine crackle	
Number of data	36	43	140
Number of patients (male, female)	14 (9, 3)	10 (7, 3)	12 (12, 0)
Diseases	Bronchitis, Chronic bronchitis, Emphysema, Lung cancer, Organized pneumonia, Pulmonary tuberculosis	Interstitial pneumonia, Non specific interstitial pneumonia, Usual interstitial pneumonia	–

* Coarse crackle includes two of unknown gender.

肺聴診音データは，各実験への入力とする前に，データのサンプリング（サンプリング周波数 2000Hz），前処理による次元圧縮（2.2 参照）を行い，前処理が TDA の場合を除き正規化（平均 0, 分散 1）を行っている．なお, Table 1 より, 本論文における肺聴診音は 1000Hz 以下の周波数帯域を持つことから，2.1 で述べた式（2.1）に基づき，2000Hz でサンプリングを行うこととした．

本論文で作成された前処理済み音声データは，時間窓ごとに周波数帯域の情報が集約された形で表現される．本章以降，各時間窓の周波数帯域を表す特徴を「周波数情報」，周波数情報の時間変化を表す情報を「時系列情報」，周波数情報と時系列情報の局所的な特徴パターンを「周辺情報」と呼称する．Fig. 11 は，MFCC に対する周波数情報・時系列情報・周辺情報の例を示す．Fig. 11 にて囲まれた箇所のうち，赤枠が周波数情報を，青枠が時系列情報を，緑枠が周辺情報をそれぞれ示す．MFCC において，周波数情報は，時間窓中の周波数スペクトルについて DCT を適用することによって得られた系列情報である．また，時系列情報は周波数情報の変化を表しており，時間窓ごとに突発的に変化が発生したかについて確認することができる．さらに，周辺情報はこれらの情報に対する任意の大きさのパターンである．

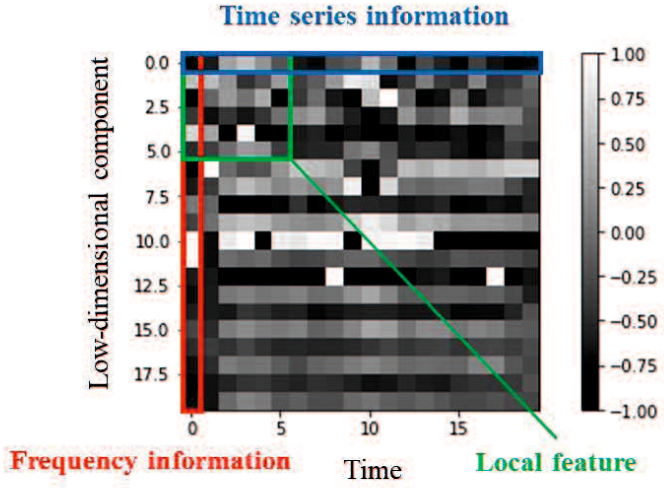


Fig. 11. Example of frequency information, time series information, and a local feature contained in MFCC

3.2　評価指標

本論文で扱う分類タスク・異常検知タスクでは，訓練データをもとにモデルを構築し，テストデータに対して性能評価を行う．ここで，性能評価のために用いられるのが評価指標である．本節では，それぞれのタスクの評価指標の定義について説明する．

なお，本論文において有意差の検定は，有意水準を 5%($p = 0.05$)に設定した片側検定によって行われる．

3.2.1　分類タスクの評価指標

本節では，分類モデルの性能を評価するための指標として，正解率（Accuracy），再現率（Recall），適合率（Precision），F 値（F1 score）の 4 つを取り上げ，それぞれの定義と計算方法について説明する．以下において，TPは True Positive（真陽性），TNは True Negative（真陰性），FPは False Positive（偽陽性），FNは False Negative（偽陰性）を表す．

また，各章においてこれらの分類タスクの評価指標（Accuracy，Recall，Precision，F1 score）はすべて%形式で記載している．

3.2.1.1　正解率（Accuracy）

正解率（Accuracy）は，分類モデルの全体的な性能を示す最も一般的な指標であり，分類モデルが正しく識別したデータの割合を示す．正解率は次式で表される．

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.28}$$

3.2.1.2　再現率（Recall）

再現率（Recall）は，陽性に属するデータのうち，分類モデルが正しく陽性であると予測したデータの割合を示す指標である．再現率は次式で表される．

$$Recall = \frac{TP}{TP + FN} . \tag{3.29}$$

本論文においては，陽性は異常データであり，再現率が高いことは異常データの見逃しが少ないことを意味する．

3.2.1.3　適合率（Precision）

適合率（Precision）は，分類モデルが陽性であると予測したデータのうち，実際に陽性に属していたデータの割合を示す指標である．適合率は次式で表される．

$$Precision = \frac{TP}{TP + FP} . \tag{3.30}$$

適合率では，上記の式によって分類モデルの予測結果の信頼性を示すことができる．

3.2.1.4　F 値（F1 score）

F 値（F1 score）は再現率と適合率の調和平均であり，分類モデルの性能をより包括的に評価する指標である．F 値は次式で表される．

$$F1score = 2 \times \frac{Recall \times Precision}{Recall + Precision} . \tag{3.31}$$

F 値は再現率と適合率のバランスを考慮し，両者を同等に重視して評価するために用いられる．F 値が高いほど，分類モデルが陽性に対して誤検出が少なく，かつ見逃しが少ない一貫した予測ができていることを示す．

3.2.2　異常検知タスクの評価指標

本節では，異常検知モデルの性能を評価する指標として使用される AUC（Area Under the Curve）について説明する．

3.2.2.1 AUC（Area Under the Curve）

AUC は, 異常検知モデルの性能を評価するために用いられる代表的な指標の一つであり, モデルが陽性（異常）と陰性（正常）を明瞭に区別できるかを数値化する．AUC は，受信者操作特性曲線（Receiver Operating Characteristic curve ; ROC 曲線）の下側の面積を表し，0 から 1 の範囲の値を取り，1 に近いほどモデルの性能が高いことを示す．たとえば，AUC = 1 の場合は，異常と正常を完璧に区別できる異常検知モデルであることを，AUC = 0.5 の場合はランダムな識別が行われていることを示す．

ここで，ROC 曲線は，異なる閾値での真陽性率（True Positive Rate ; TPR）と偽陽性率（False Positive Rate ; FPR）をプロットしたものである．TPR，FPRは次式で表される．

$$TPR = \frac{TP}{TP + FN}, \tag{3.32}$$

$$FPR = \frac{FP}{FP + TN}. \tag{3.33}$$

FPR が低いほど誤検知が少なく，TPR が高いほど異常を見逃さず検知できていることを表す．本論文における異常検知モデルの結果と ROC 曲線の例を Fig. 12 に示す．

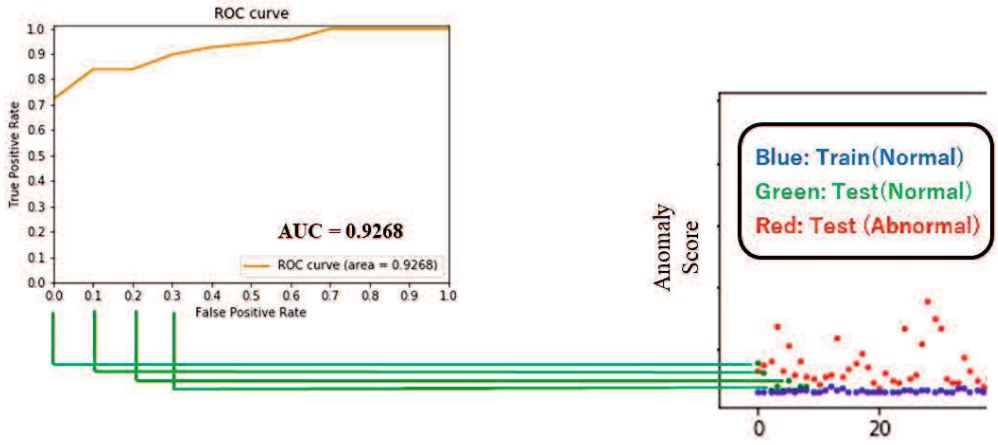


Fig. 12. Example of anomaly score and ROC curve

Fig. 12 に示すように，本論文の異常検知タスクでは，テストデータにおける正常データの件数を 10 件に固定し，各正常テストデータの異常スコア（Anomaly Score）を降順にソートすることでFPR，TPRの閾値とした.

第 4 章　深層ニューラルネットワークを用いた肺聴診音の識別

聴診の際に医師が行っているのは，本質的には肺聴診音の分類である．本章では，分類タスクに注目し，分類モデルの開発・検証を行う．精度の高いモデルを構築するため，肺聴診音に対する特徴エンジニアリングの検討や，既知の課題である「データの少なさ」についても対応する．

4.1　背景と目的

医師は聴診において，聴診機器を介して得られる音声信号の各時刻に注目し，異常音の手がかりの有無を基に肺聴診音の識別を行うことで診断を行っている．診断支援に向けた取組みとして，本論文で対象とする肺聴診音についても，はじめに分類モデルを構築することを検討する．

3.1 で述べた通り，肺聴診音を統計的に識別することは，音声信号の複雑性や内的・外的要因によるノイズの影響を受けるため困難である．そのため，従来の統計ベースの手法では限界があり，より精度の高いモデルを求める場合には，深層学習などのより高次の表現力を持つモデルを用いることが必要となる．

このとき，深層学習で肺聴診音の学習を行うには，通常大量の音声データが必要である．しかし，深層学習モデルの学習に必要なアノテーション（クラスラベル付け）が行われている音声データを各医療施設が十分な量取得するには長い時間がかかる．これは，1.1 で述べたように，聴診音をデジタル化する取り組みが遅れていること，そして，デジタル化したデータを医師がトリミングし，ラベル付けする労力が大きいためである．深層学習を用いて少

ないデータで学習を行う場合，個人差が大きくかつ識別に寄与しない様々なノイズが混在する聴診音では，診断に重要で本質的な特徴を捉えきれず，汎化性能が得られない可能性がある．

そこで，本章では，深層学習を用いて汎化性能の高い聴診音識別器を構築することに焦点をあて，深層学習においてデータ数の少なさに対応できる手法を提案する．これには，ラベルなしデータによる事前学習が効果的である．さらに，肺聴診音を効率的に学習する機構について検証するために，肺聴診音に適した深層学習アーキテクチャの検討や，音声特徴抽出法の拡張などについても上記と組み合わせて検証する． 具体的には，3 種類のネットワーク： 1) CNN，2) LSTM，3) C-LSTM をベースとする自己符号化器を構成し，ラベルなしデータであらかじめ聴診音の特徴を学習させる．その後，少数のラベルありデータで訓練させることで，効果的な学習を行う．3 種類のネットワークのうち，CNN は入力情報の全体を俯瞰した学習を行うことが可能である．LSTM は，時間軸の依存関係を考慮した学習が可能であり，時系列の性質を持つ聴診音への効果を検証する．C-LSTM は CNN と LSTM のどちらの特徴も有しているため，データを俯瞰した局所的特徴と時系列の両方を考慮した学習が可能である．これらの複数のアーキテクチャを採用・比較することにより，データが少ない場合でも肺聴診音を深層学習モデルで効果的に学習するための特徴抽出方式を明らかにする．

4.2 方法

ここで，一般的な深層学習を用いた分類モデルにおいて，少数の訓練データを用いた学習に効果的な特徴抽出法の検討を行う．1 つ目に挙げられるのは，事前学習によるドメイン知識の獲得である．事前学習では，（訓練データとは異なる）音声データを事前に AI モデルに学習させ，その後訓練データによる学習を行う．このように多段の学習を行うことによっ

て，AI モデルが音声に関する特徴表現を事前に把握でき，事前学習がない場合と比較して学習効率が向上する．ここでは， CNN，LSTM，C-LSTM に対して事前学習法の設計を行った．具体的には，CNN，LSTM，C-LSTM を 3.3.1 で述べたような自己符号化器として構成し，ラベルなしデータを用いて聴診音の特徴を事前学習（Pre-training）する．その後，事前学習済みニューラルネットワークをベースとした識別器を構成し Fine-tuning[41]を行う．2 つ目に挙げられるのが，ニューラルネットワークへの入力の前処理の拡張である．ここでは，入力として一般的に行われる前処理である MFCC について拡張を行う．具体的には，MFCC の最適な次元数を検討するとともに，MFCC の動的特徴量を併用する．次節以降，各提案手法について述べる．

ここで，対象とするタスクは，Coarse，Fine，Normal の 3 クラスの聴診音に対するクラス分類である．訓練データは Table 2 のうち各クラスから 27 個ずつ，テストデータは各クラスから 9 個ずつランダムに抽出し，これを 100 回繰り返して正解率，再現率，適合率の平均を求めるためのデータセットとした．事前学習に使用するデータは，上述のデータセットに含まれず，ラベルも付与されていない約 30 秒の肺聴診音であり，5 秒のフレームを 0.25 秒のストライドで 107 個に分割したものである．

4.2.1　CNN, LSTM, C-LSTM に対する事前学習と Fine-tuning（提案手法 1）

CNN，LSTM，C-LSTM で自己符号化器の構成方法が異なるため，本節ではその詳細について説明する．

4.2.1.1　CNN の事前学習

CNN を用いた事前学習と Fine-tuning の概観を Fig. 13 ,Fig. 14 にそれぞれ示す．ここでは，3.1 で説明した 5 秒の肺聴診音データを識別するにあたり，まず畳み込み自己符号化器（Convolutional Autoencoder ; CAE）で事前学習させた後，分類タスクに特化した Fine-

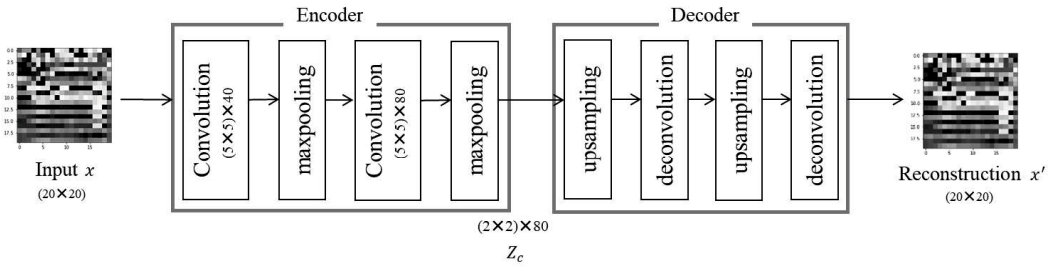


Fig. 13.　Pre-training of CNN

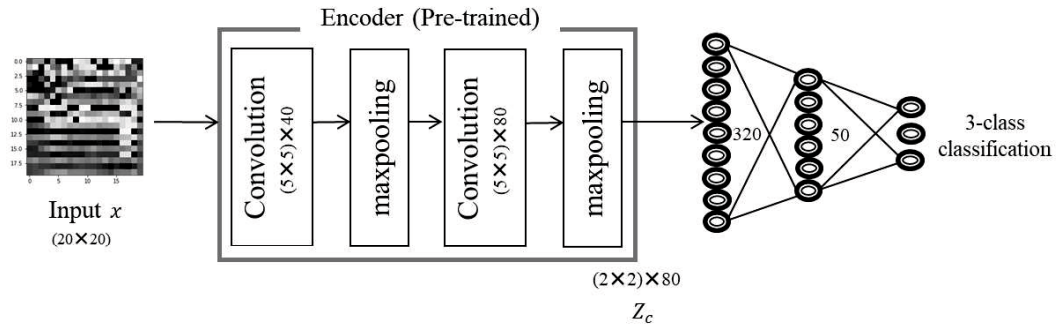


Fig. 14.　Fine-tuning of CNN

tuning を行う．このとき，入力データxは，元の音声データに MFCC を適用し Fig. 2 のような 2 次元データに変換したものである．CAE の学習は 2 段階にわたって行われる．1 段階目はラベルなしデータの学習によるドメイン知識の獲得である（Fig. 13）．Fig. 13 の入力データxには, 訓練/テストとは異なるラベルなしのデータを 5 秒ごとに分割したものを使用する．1 段階目では，2 個の畳み込み層と 2 個のプーリング層で構成された Encoder によって入力xを符号化し，その後，2 個の逆畳み込み層と 2 個のアップサンプリング層で構成された Decoder によって復元する．Fig. 13 において，たとえば 1 つ目の Convolution では，5×5 のサイズの畳み込み窓を 40 個保持していることを示しており，符号化後の中間出力（Encoder の出力）となる特徴量Z_Cは 2×2 の特徴マップが 80 個生成されることを示している．CAE では，学習において入力と出力から計算される MSE を損失関数として最小化する．1 段階目の学習後，2 段階目の学習である分類モデルとしての学習を行う（Fig. 14）．

Fig. 14 の入力データxには, クラスラベル付きの肺聴診音データを用いる． Fig. 13 の CAE の Decoder を切り離し，Encoder の重みを固定したまま，全結合層を 3 層付加した CNN を構成し, ラベル付きの肺聴診音データで損失関数を Cross-entropy とした Fine-tuning を行う.

4.2.1.2 LSTM の事前学習

LSTM を用いた事前学習と Fine-tuning の概観を Fig. 15，Fig. 16 にそれぞれ示す．ここで，Fig. 15 が LSTM による自己符号化器である[42]．これは，入力を特徴量に変換する LSTM である「Encoder LSTM」と，特徴量から復元を行う LSTM を「Decoder LSTM」によって構成される．LSTM を用いた事前学習として，CAE の場合と同様，1 段階目では 5 秒ごとに分割したラベルなしのデータxを入力し, Fig. 15 の自己符号化器（Encoder LSTM, Decoder LSTM）で学習させる．たとえば 20 次元の MFCC を学習するとき， ネットワー

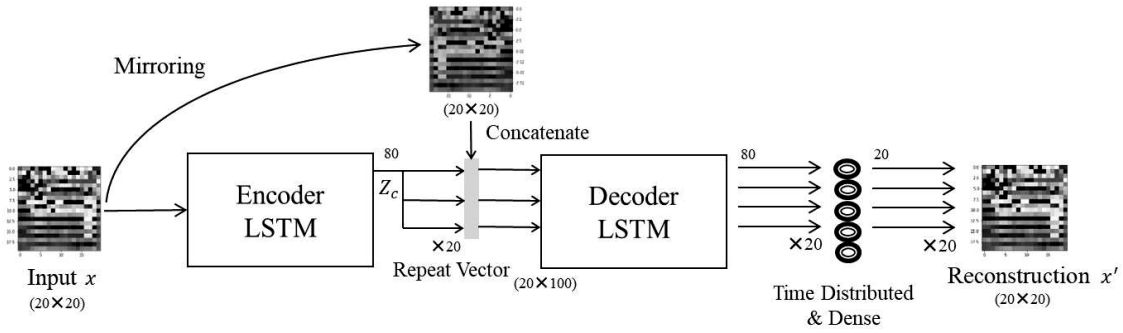


Fig. 15.　Pre-training of LSTM

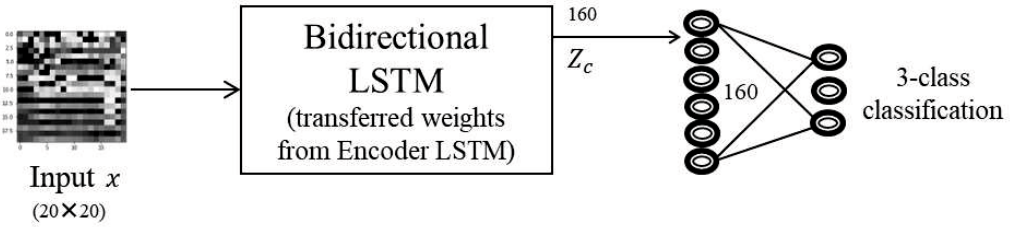


Fig. 16.　Fine-tuning of LSTM

ク構造の詳細は以下のとおりである．まず，Encoder LSTM で，LSTM の隠れ層から 80 個の特徴量が出力される．次に，得られた特徴量に対して，Encoder LSTM に入力したデータ（20×20）を鏡映したものを加える．この処理は，Fig. 15 のように，一次元配列である LSTM の特徴量（80）を入力データの行数（20）だけ複製して 2 次元配列（20×80）とし，鏡映データ（20×20）を結合させることで行う．これにより，新たな特徴量（20×100）を得る．ここで，鏡映データは，時間軸の情報が失われている Encoder LSTM の特徴量（80）に対し，特徴量から入力データを復元するためのヒントとしての役割があり，また逆時間軸の情報を与える効果もある．以上のように生成された特徴量（20×100）を Decoder LSTM に時刻順（1×100）に入力する．Decoder LSTM では，時刻順に入力されるデータ（1×100）に対して毎時刻出力（80）を行う．出力された一次元配列（80）を，全結合層に入力することで，元の入力データの横軸と同等のサイズ（20）に変換する．ここで，復元された情報（20）を全時刻（20）で順番に結合することで，元の入力データと同じサイズのデータ（20×20）が生成される．LSTM を用いた自己符号化器においても CAE と同様にこれを復元データとみなし，MSE を損失関数として入力と同じになるように学習を行う．

自己符号化器の学習終了後は，Fig. 16 に示す識別器を構築し，2 段階目の学習を行う．ここで分類モデルは BiLSTM とした．BiLSTM における順時間軸の LSTM の重みと逆時間軸の LSTM の重みはどちらも Fig. 15 の Encoder LSTM の学習済み重みを用いている．Fig. 16 では，BiLSTM に全結合層を 2 層追加し，ラベル付き入力データxを用いて Fine-tuning を行う．

4.2.1.3　C-LSTM の事前学習

畳み込み LSTM（C-LSTM）は，2.3.5 で述べた通り，LSTM 内部の計算において直積の計算を畳み込みに置き換えた LSTM である．この構造によって C-LSTM は CNN と LSTM の両方の特徴を持っているため，周辺情報と時系列情報の両方の特徴を考慮できる．C-

LSTM の出力形状は（畳み込み後の特徴量マップの大きさ）×（特徴量マップの数）であり，CNN と同様である． C-LSTM を用いた事前学習と Fine-tuning の概観を Fig. 17，Fig. 18 にそれぞれ示す． C-LSTM の事前学習でも，CNN や LSTM の場合と同様に，5 秒ごとに分割したラベルなしのデータを自己符号化器で学習した後，Fine-tuning を行う．Fig. 17 に示すように，1 段階目の学習では，自己符号化器は 2 個の C-LSTM で構成された Encoder によって入力データxを符号化し，その後，2 個の逆畳み込み層で構成された Decoder によって復元する．LSTM と異なり，Decoder を逆畳み込み層としたのは，Encoder の出力が CNN と同様に 2 次元の特徴マップとなるためである．

また，C-LSTM を用いた自己符号化器については，使用される Decoder に確立された手法が存在しない．そこで，時系列情報が失われる可能性を考慮し，本節における CAE の Encoder では Max pooling による特徴マップの縮小を用いないこととした．したがって，

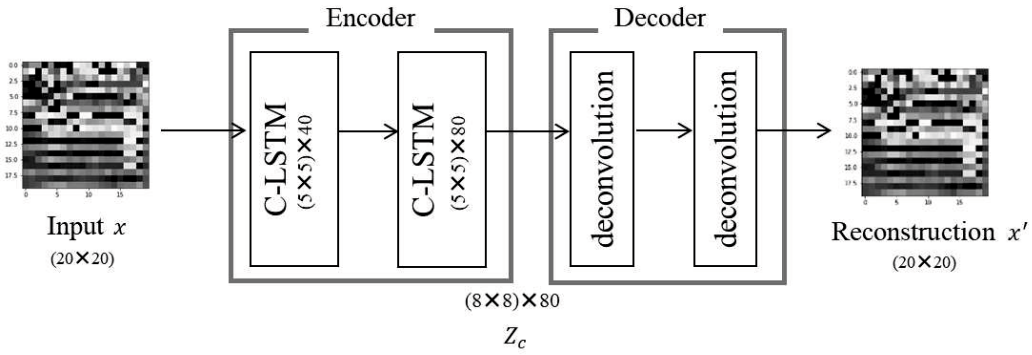


Fig. 17.　Pre-training of C-LSTM

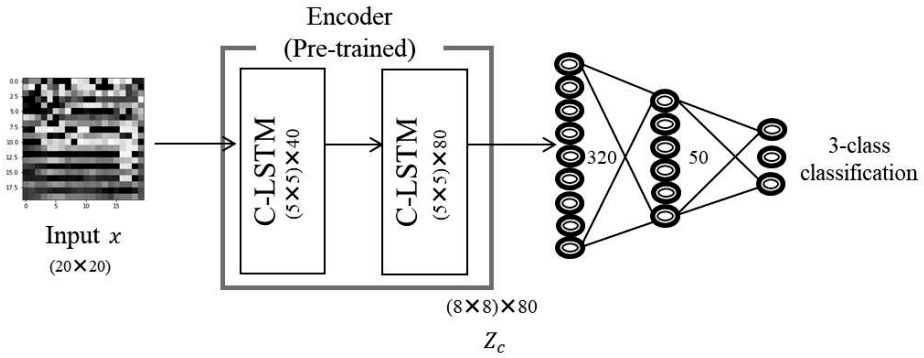


Fig. 18.　Fine-tuning of C-LSTM

これに対応し Decoder においても Up-sampling による特徴マップの拡大を行っていない. Fig. 18 に示すように，1 段階目の学習後は，Fig. 17 の Encoder の重みを固定したまま全結合層を 2 層付加した C-LSTM による分類モデルを構成し，ラベル付き入力データxを用いて Fine-tuning を行う.

4.2.2　MFCC 次元数の調整（提案手法 2）

5 秒間の聴診音データについて，複数の次元数で MFCC を抽出した場合のグレースケール画像を Fig. 19 に示す．MFCC の高次元に離散することによって得られるピッチ成分は一般に分析が難しいため，通常はフォルマント成分が凝集している 12～20 次元までを使用する．しかし，Fig. 19 における 20 次元以上の高次成分にも特徴の変化が表れている．これは微細なスペクトル包絡を表しており，より細かい波形を表現する情報が含まれているが，ノイズとなるピッチ成分が混入し得るため一般的な MFCC では用いられない. 肺聴診音は，3.1 で述べた類似した周波数帯域・ノイズなどの性質から，モデルの精度を高めるために必要な特徴が一般的な音声認識とは異なる可能性がある．そのため，MFCC の高次成分も一定の次元数までは有用である可能性があり，MFCC の抽出次元数を増加させることによっ

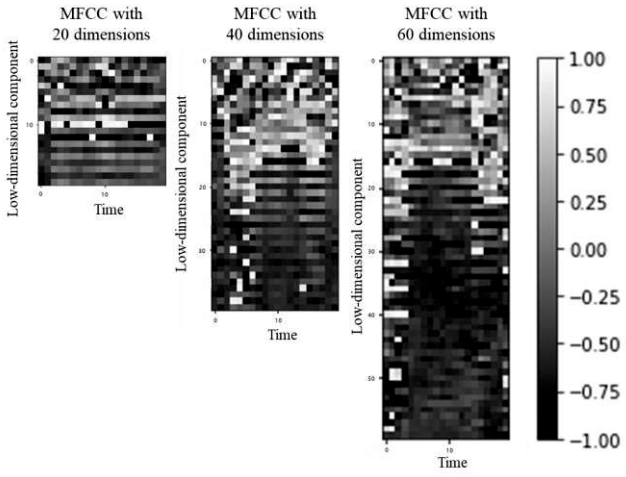


Fig. 19.　Grayscale MFCC images at multiple dimensions

て識別精度が向上するかを検証する．また，個人差の大きいピッチ成分の情報を加えすぎると性能の低下が予想されるため，通常用いられる MFCC の次元数の 2 倍，3 倍までの情報量が妥当であると判断し，拡張を行った．

4.2.3　動的特徴量（提案手法 3）

MFCC に関連した特徴量であるΔメルケプストラム（ΔMFCC），ΔΔメルケプストラム（ΔΔMFCC）[13]を組み合せた特徴量を使用する．ΔMFCC は MFCC の隣接スペクトルの微分値，ΔΔMFCC はΔMFCC において隣接する値の微分値であり，動的特徴量として用いられている一般的な手法である．本論文の CNN では，MFCC で得られた特徴量を 1 チャネルの入力画像として扱っているが，ΔMFCC とΔΔMFCC を同時に用いることの有効性が示されている[43]ことから，MFCC，ΔMFCC，ΔΔMFCC を合成し 3 チャネル画像として学習を行うことを検討する．これにより，CNN では明示的に捉えられない，MFCC の周波数情報の動的な時間変化を CNN でも捉えることが期待される．

動的特徴量の使用については，時系列ニューラルネットワークの LSTM では，一般に 3 チャネル以上の情報を扱うことができず，また，時系列情報を元々考慮できる構造であるため，実験では CNN のみに動的特徴量を適用した．このとき，前処理後の入力データのサイズは，MFCC において 20 次元まで抽出する場合では 20×20 であり，40 次元では 40×20，60 次元では 60×20 である．

4.3 結果

提案手法を組み合せて得られた正解率，再現率，適合率の平均を Table 3 に示す．Table 3 より，MFCC を用いた肺聴診音分類モデルの正解率は CNN に提案手法の 1 と 2 を組み合せたものが最もよい結果（78.07%）となった．比較のため，提案手法を組み込まない従来法（MLP，CNN，LSTM，C-LSTM）についても実験を行ったところ，最も性能が高い

Table 3. Mean accuracy, recall and precision of pre-trained classification models

	Method	Accuracy (%)	Recall (%)		Precision (%)
			Fine	Coarse	
	MLP	69.59	68.39	63.07	77.21
	CNN	70.37	69.00	64.86	76.58
	LSTM	69.52	65.44	65.11	78.00
	C-LSTM	73.81	75.56	70.56	75.33
P1	CNN with pre-training	73.48	72.33	67.89	80.22
P1	LSTM with pre-training	73.33	73.11	71.22	75.67
P1	C-LSTM with pre-training	75.85	77.44	71.78	78.33
P2	CNN with pre-training and 40D MFCC	76.56	76.44	74.11	79.11
P2	CNN with pre-training and 60D MFCC	**78.07**	80.44	72.44	81.33
P2	LSTM with pre-training and 40D MFCC	73.22	74.78	67.78	77.11
P2	LSTM with pre-training and 60D MFCC	72.07	71.44	65.89	78.89
P2	C-LSTM with pre-training and 40D MFCC	73.59	72.56	70.11	78.11
P2	C-LSTM with pre-training and 60D MFCC	72.33	69.11	69.33	78.56
P3	CNN with pre-training and dynamic features	76.15	77.00	69.78	81.67
P4	CNN with pre-training, 40D MFCC and dynamic features	74.30	72.22	70.33	80.33
P4	CNN with pre-training, 60D MFCC and dynamic features	76.37	76.56	69.67	82.89

P1: Proposed method 1, P2: Proposed method 1+2, P3: Proposed method 1+3, P4: Proposed method 1+2+3

のは C-LSTM の正解率 73.81%であり，続いて CNN の正解率が高かった．平均的に識別性能の高いこれらの深層学習アーキテクチャはどちらも畳み込み処理を内部で行っていることから，肺聴診音の識別においては，畳み込み処理が重要な役割を果たしていると言える．

ここで，各提案手法の有用性について考察する．提案手法 1（Table 3 の P1 行）では，事前学習を行うことによる識別性能の向上が確認できた．いずれのアーキテクチャにおいても従来手法と比べて性能が向上していたことから，事前学習は小規模な深層学習においてもモデルの性能向上に効果的であり，その効果はアーキテクチャに依らないことが示された．ただし，本論文で用いた事前学習用の肺聴診音データは，4.2 で述べた通り，類似する診察環境（これは，聴診機器，収音環境，被験者の属性などが含まれる）であることから，事前学習に用いるデータがどのようなデータでも機能することを保証するものではない．

次に，提案手法 2 における MFCC の次元数調整では，CNN と LSTM 及び C-LSTM で結果が異なった．まず事前学習付き CNN に対して MFCC の次元数を 20 から 40，60 に増やしたとき（Table 3 の P2 行），60 次元が最も高い正解率であり，次元数を増やすにつれて高い正解率を示した．したがって，MFCC では高次元にも識別に有用な特徴を持っていると言える．一方，LSTM 及び C-LSTM については，MFCC の次元数を増やすほど性能は低くなった．これは，MFCC の高次元に有用な特徴成分が存在しつつも，ノイズに対する頑健性が低い時系列ニューラルネットワークがノイズであるピッチ成分を学習することで，識別に悪影響が出ていると考えられる．

提案手法 3 においては，事前学習付き CNN に動的特徴量（MFCC + ΔMFCC + ΔΔMFCC の 3 チャネル画像）を与えて学習を行ったところ，正解率の改善がみられた（Table 3 の P3 行）．したがって，CNN 単体で計算する畳み込み窓の内部の情報のみで捉えられない，識別に有用な時系列の情報を補完できることが確認できた．

しかし，提案手法全ての組合せ（提案手法 1 + 2 + 3）は効果が得られなかった（Table 3 の P4 行）．各手法で得られた新しい特徴量は識別に有用であったものの，組み合わせるこ

とによって一部の特徴量部分が学習できないノイズとなったことが考えられる．提案手法 2 と提案手法 3 は，CNN のみ性能向上の効果が確認されていることから，いずれも特徴量に含まれるノイズが増加する手法である．提案手法 2 で発生したピッチ成分情報のノイズに対して，ΔMFCC・ΔΔMFCC などの差分情報を計算することにより，本来はフォルマント成分である箇所までノイズが広がり，識別上意味のないノイズとして学習を阻害したことにより，CNN においても性能が低下してしまったと考えられる．ただし，この効果の前提条件となるのは，(1) データが少数であること，(2) 熟達した医師が取得した質の高いデータであること，である．本論文で扱うデータは 100 件に満たない少数データであるが，データの規模が拡大することによって提案手法の全ての組合せが効果を発揮する可能性がある．

ここで，肺聴診音とアーキテクチャの相性に関して洞察する．まず，本章で扱った，少数のデータかつ音声信号の分類という局所解に陥りやすいタスクにおいては，畳み込み処理を含むアーキテクチャ（C-LSTM，CNN）による特徴抽出が有効であることがわかる．これは，第一に，畳み込み処理によるノイズの軽減の効果であるといえる．Table 3 より，LSTM の性能が最も低かったことから，本論文で扱う肺聴診音データについても，初期値などの諸条件によっては時系列ニューラルネットワークで補足することが難しい複雑な周期性，もしくはクラスラベルと関連しない突発的なノイズが含まれる可能性が高い．時系列ニューラルネットワークは学習によって内部状態を更新するため，これらの情報が累積することによる影響を受けやすく，特に少ないデータではモデルが周期性の一部，もしくはクラス分類とは関連しない周期性に着目する可能性がある．一方，CNN による特徴抽出では畳み込みを用いることで，提案手法 2・提案手法 3 の考察で示したように，入力データのノイズによる過学習が緩和されることが確認できた．

第二に，肺音の周波数帯域の情報と，その変化情報が識別に有用である可能性がある．畳み込みの処理では，周辺情報を特徴量として学習するため，周波数情報，時系列情報，もしくはこれらの組合せが識別において有効であると言える．Table 3 において C-LSTM の性

能が CNN を上回っていたことや，異常データの識別というタスクの特性からも，時系列情報は肺聴診音の識別上有用である．

また，識別上有用である時系列情報をどのように与えるべきかについても考察する．時系列ニューラルネットワークは本来時系列の学習に特化したニューラルネットワークであるが，単体では肺聴診音の時系列を適切に学習できていないことが Table 3 で確認された．肺聴診音において，時系列情報を考慮した学習を行う際，LSTM 単体では MLP よりも性能が低かったのに対して CNN では動的特徴量（差分情報による時系列情報）を付与したことで正解率が高まっている．このことから，ノイズに頑健なアーキテクチャで時系列を扱うことが適切であると考えられる．しかし，ノイズに頑健な畳み込みを LSTM に追加した C-LSTM においても，Table 3 において一貫して識別性能が高かったわけではないことから，時系列情報に対する特徴抽出アーキテクチャはさらに検討する余地がある．他にも，MFCC の時系列情報と時系列ニューラルネットワークの相性の悪さも考えられるため，前処理においても改良できる可能性がある．

4.4 まとめ

本章では，データ数の少ない肺聴診音識別に対して有用な深層ニューラルネットワークの構成を明らかにするため，事前学習の適用，および特徴量の改良（MFCC の抽出次元数の調整と動的特徴量の使用）を行った分類モデルを提案し，正解率の評価を行った．

実験結果より，事前学習はすべての手法において，少ないデータに対する学習の効果が顕著であり，特徴量の改良（MFCC の次元数調整，動的特徴量）は CNN に対して有効であった．従来法の C-LSTM および事前学習のみを付加した C-LSTM の性能は同条件の他手法と比較して優れていたため，周辺情報と時系列情報はどちらも聴診音の識別において有効と言える．

今回比較したアーキテクチャのうち，C-LSTM，もしくは CNN（MFCC の次元数増加を含む）が肺聴診音の特徴抽出に向いていると言える．CNN は特徴量にノイズが混入する前処理データ変換（提案手法 2，提案手法 3）についても頑健に識別率を向上させることができた．また，C-LSTM はベースの性能が高く，アルゴリズムの特性上，CNN と比較するとノイズに対する頑健性は低下するものの，時系列情報を加えることによる有用性も確認できた．

次章では，本章で扱った特徴量やアーキテクチャをさらに深く検討し，1）肺聴診音の特徴抽出において有効な情報とそれぞれの情報が持つ性質，2）診断支援の実用上の課題である解釈性について述べる．

第 5 章　時系列ニューラルネットワークを用いた肺聴診音の効果的な特徴抽出と識別

第 4 章では，一般的な分類モデルのアプローチについて，データが少ない事象に対する対策と音声特徴量を拡張することの効果について検証した．本章では，肺聴診音と深層学習モデルの特性をより考慮し，深層学習モデルに適した肺聴診音特徴の抽出法について検討する．特に，4 章において十分な性能が得られなかった時系列ニューラルネットワークに着目し，肺聴診音に効果的に適用する方法を検討する．

5.1　背景と目的

本論文で扱う肺聴診音は，第 1 章で述べたようにデジタル化が遅れており，医用診断への応用が進んでいない．そのため，データが不足しており，AI モデル，特に深層学習向けにどのような特徴量エンジニアリングが適しているのかについて述べている文献はほとんど存在しない．

そのため，第 4 章では，複数の深層学習アーキテクチャによる特徴抽出の検討や，一般的な前処理の拡張を行った．これにより，1) 特に入力のノイズに頑健な深層学習アーキテクチャの性能が高い傾向があり，入力のノイズに弱い時系列ニューラルネットワークの性能が劣っていたこと，2) データの時系列情報に識別上有用な情報が存在する，などの示唆が得られた．しかし，1)の時系列ニューラルネットワークの性能劣化は，必ずしも肺聴診音の特徴抽出の全てに当てはまるとは限らない．第 4 章にて行った検証は，一般的な処理を肺聴診音に対して応用・拡張したものであり，前処理の方法によっては時系列ニューラルネットワークにおいても精度を発揮できる可能性がある．

さらに，2)についてもさらなる検討が必要である．前処理後の MFCC に含まれる周波数情報や時系列情報について，アーキテクチャも含めてどのような特徴抽出が識別上重要な貢献を果たしているのかについては明らかにできていない．特に，周波数情報単体の有用性について，Table 1 より周波数帯の違いから有用であることは考えられるものの，明確な実験結果が得られていない．

そこで，本章では，肺聴診音の特徴量エンジニアリングについてさらに洞察を深めるための検証を行う．第 4 章において効果を発揮できなかった時系列ニューラルネットワークと，周波数情報の有用性の評価に焦点をあて，前処理やアーキテクチャを含む特徴抽出法の変更によってこれらが識別性能の向上に寄与するか検証する．

また，肺聴診音の CAD 技術の確立にあたり，重要な観点として挙げられるのが「結果の解釈性」である．本論文における解釈性は「モデルの出力結果の直接的な根拠が，医師を補助可能な形で提示できること」として定義される．機械学習分野で類似する単語には，説明可能性（Explainability）もあるが，こちらは「モデルの予測が一貫してホワイトボックスであること」を指しており，異なる概念である．

複雑な 1 次元データを扱う音声信号処理分野において，CAD 技術に求められる高い性能を発揮するためには，複雑な計算を表現するためにブラックボックスな推論が必要となる場合がある．そのような場合にも，医師がモデルの結果を確認する機構があることが望ましいため，本論文では解釈性の検討を行った．

本章では，第 4 章で扱った深層学習アーキテクチャに加え，時系列ニューラルネットワークの一種である Transformer を追加した．Transformer は Attention と呼ばれる機構により，解釈性に関する出力が可能であるため，肺聴診音に対する解釈性の検討を併せて行った．第 4 章と同様に複数のアーキテクチャで分類モデルを構築し，より詳細な比較を通じて，肺聴診音の識別に必要な特徴量を整理する．

5.2　方法

本章では，肺聴診音に対する深層学習における，新たな特徴抽出手法を提案する．提案手法では, 時系列ニューラルネットワークの処理特性を考慮し, MFCC の転置行列を用いて, 音声に含まれる周波数情報に対する抽出能力を向上させる．同時に，さらなる識別性能の向上を目指し，周波数情報と時系列情報を組み合わせたアーキテクチャを提案する．次節以降, 各提案手法について述べる．

ここで，対象とするタスクは，第 4 章と同様，Coarse，Fine，Normal の 3 クラスの聴診音に対するクラス分類である．訓練データは Table 2 のうち各クラスから 27 個ずつ，テストデータは各クラスから 9 個ずつランダムに抽出し，これを 100 回繰り返して正解率, 再現率，適合率，F1 スコアの平均を求めるためのデータセットとした．

また，検証する深層学習アーキテクチャについて，時系列ニューラルネットワークの一種である Transformer を追加する．Transformer は近年では事前学習を前提とした大規模なモデルで使用されることが一般的ではあるが，本章では比較のため，小規模なデータセットに合わせて層数の少ない Transformer Encoder を構成した．本論文における Transformer Encoder は 2.3.6 に示した Transformer1 層と Global average pooling，全結合層によって構成される．Transformer において，式（3.16）のh を 3 とし，各時刻の埋め込みサイズd_{model}を 20, Feed Forward Network の隠れ層のサイズを 4 とした. このとき, Transformer と Positional Encoding は，BERT（Bidirectional Encoder Representations from Transformers）[44]に基づく手法を採用した．本論文で使用する Positional Encoding は正弦波（sinusoidal）に基づいており，入力データの各位置に対するエンコーディングは事前に固定値として決定される．

5.2.1　転置 MFCC（提案手法 1）

3.2 で述べたそれぞれの深層学習アーキテクチャは，特徴抽出に関して独自の着眼点を備

えている．例えば，CNN は畳み込み処理によって周辺情報（MFCC 中に発生する，周波数情報と時系列情報の組合せによる局所的なパターン）を考慮することに優れている．そのほか，LSTM と Transformer などの時系列ニューラルネットワークは，時系列となる各時刻の情報の変化を学習することに優れている．

4.3 の実験結果より，肺聴診音に関する特徴抽出について考えられる仮説として，（特に本論文のようなデータ数においては）MFCC の時系列情報をメインとした学習は十分な効果を発揮できないことが挙げられる．一方で，Table 1 のデータ特性や，4.3 において CNN・C-LSTM などの局所的特徴を抽出するアーキテクチャの性能が比較的高かったことから，周波数情報には識別上有用な特徴が存在する可能性がある．

ここで，周波数に特化した学習を行うことを検討する．そのために，時系列ニューラルネットワークが活用できる可能性がある．具体的には，時系列ニューラルネットワークへ MFCC を入力する直前に，前処理として転置を実施する．このとき新たに入力とする$MFCC'$は次式で表される．

$$MFCC' = MFCC^T. \tag{5.1}$$

MFCC が持つ周波数情報は，2.2.2 で述べた通り，時間窓ごとに離散コサイン変換された周波数スペクトルである．これは，DCT によって低周波のコサイン波ごとに集積された，系列としても意味を持つ値であるため，時系列ニューラルネットワークへの入力を式（5.1）とすることで，各時間窓の周波数表現に着目した学習ができる．なお，時系列ニューラルネットワークで MFCC を直接入力した場合には，周波数帯域間の相互依存性は明示的に学習されない．

臨床研究において，転置 MFCC を採用する，MFCC の 2 つの時系列（時間，周波数）を考慮することは標準的ではないが，肺聴診音のような小規模かつ特定の周波数帯域に集中した音声信号の識別が求められる場合は周波数帯域の時間的変動を捉えることによる効果が期待される．ただし，CNN など非時系列ニューラルネットワークについては原理上効果

がないため実験は実施していない.

5.2.2 Cross-encoding Transformer（提案手法 2）

5.2.1 で述べた通り，前処理に変更を加えた時系列ニューラルネットワークにより，周波数帯の系列パターンに対する学習効率の向上が期待される．しかし，このとき，持続時間などの時系列パターンの学習効果については低下する可能性がある．第 4 章において時間情報についても有用である示唆が得られているため，周波数と時間の 2 つの時系列情報を適切に学習させることが必要である．

本節では，MFCC の持つ 2 つの時系列（時間，周波数）情報をより深く考慮するアーキテクチャを提案する．本アーキテクチャは，転置前と転置後の MFCC を組み合わせて特徴抽出を行うもので，ここでは Cross-encoding Transformer と呼称する．Fig. 20 に Cross-encoding Transformer の構造を示す.

Cross-encoding Transformer は入力として，転置前の MFCC（Fig. 20 中のInput x）と転置後の MFCC（Fig. 20 中のTransposed x）の 2 つの入力を用いる．これらの入力データは Positional Encoding[44]によって位置情報が付与され，それぞれ独立した Transformer Block に入力される．Transformer Block は 5.2 の Transformer Encoder で使用しているものと同様の 1 層の Transformer を表す．このとき，Positional Encoding は行方向を時系列とみなして位置情報を付与するため，転置前の MFCC が時間方向の時系列情報を，転置

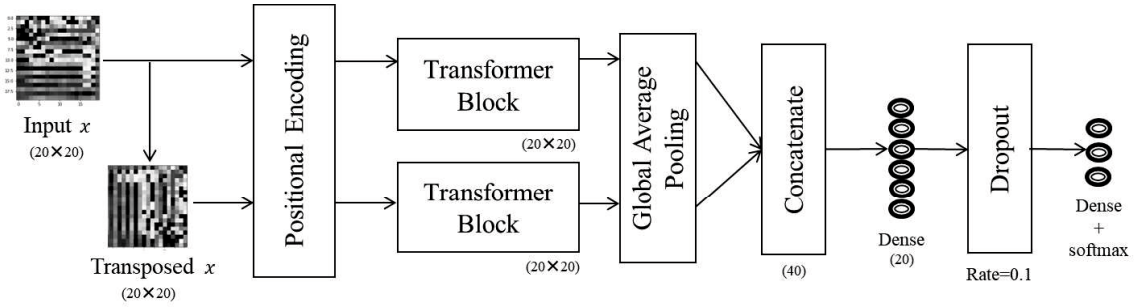


Fig. 20.　Structure of Cross-encoding Transformer

後の MFCC が周波数に関する時系列情報を，それぞれ学習することを期待する．Transformer によって学習されたそれぞれの時系列情報は，Global average pooling によってマージされ，さらに全結合層を追加することによって分類モデルとしてクラスを識別可能な特徴の学習を行う．上記のアプローチにより，肺聴診音の持つ時間情報と周波数情報について，従来の手法と比較してより効果的な情報統合を実現し，その結果として信号処理タスクの性能改善が期待できる．

さらに，Transformer には解釈可能性が期待される．医療かつ信号処理分野における AI の解釈可能性に関する論文は少ないが，コンピュータ支援診断を実現するための重要な要素の一つである．通常，MFCC のような 2 次元で表現されるデータに深層学習を適用する場合には，SHAP（SHapley Additive exPlanation）[45]や LIME（Local Interpretable Model-agnostic Explainations）[46]などの 2 次元の可視化手法が用いられる．しかし，音声信号処理で一般的な特徴量である MFCC は不可逆変換であり，ヒートマップ形式で表現されるこれらの手法は出力の根拠として解釈することは困難である．

ここで，Transformer を用いた解釈性について検討する．本論文で提案する Cross-encoding Transformer では，時間方向（Fig. 20 上部の Transformer Block）の Attention の重みに基づき，肺聴診音データのどの時間が識別根拠となったかを計算した（以降，Attention スコアと呼称する）．サンプリングされた肺聴診音データsに対する各時間系列への Attention スコアを次式に示す．

$$Score_s = repeat\Big(\big(\max\big(0, \min\big(1, GAP\big(F(MFCC_s)\big)_i\big)\big)\big)_{i=1}^{len(MFCC_s)}, \frac{len(s)}{len(MFCC_s)} \Big), \tag{5.2}$$

ここで，$F(x)$は Positional encoding を含む時間方向の Transformer モジュールを表す．肺聴診音データsは MFCC に変換（$MFCC_s$）され，内部の時間方向の Transformer モジュー

ルで処理される（$F(x)$）．ここで，$F(x)$はグローバル平均プーリング（式（5.2） の$GAP(x)$）によって変換され，$MFCC_s$の各時刻に対して 1 つの値を取る．このときの各時刻の値は$\max(0, min(1, x))$によって 0 から 1 の間の数値にクリッピングされ，$repeat(x, k)$によって，配列xの各要素を指定した回数kだけ繰り返して音声データsの長さに戻すように展開する．ここで得られる Attention の重みは，予測におけるラベルを決定する際に Cross-encoding Transformer が注目した度合いを表している．Attention スコアが高い時刻ほど，予測において重要な決定点となった時刻であると解釈できる．

5.3 結果

本節では，各手法の性能と可視化された特徴マップなどをもとに，得られた結果を述べ考察する．さらに，Transformer によって得られる識別結果の解釈性についても議論する．

5.3.1 識別性能の比較

提案手法を含む各深層学習アーキテクチャについて，正解率，再現率，適合率，F1 スコアの試行平均を Table 4 に示す．このとき，再現率，適合率，F1 スコアはマクロ平均によって導出を行っている．医療分野において病気を見逃すことは重大なリスクにつながることから，以降では，再現率（Recall）に基づいてアーキテクチャ間の性能について述べる．また，各手法間の片側 T 検定の結果を Table 5 に示す．

Table 4.　Mean classification performance by deep learning architecture

		Accuracy(%)	Recall(%)	Precision(%)	F1 Score(%)
Methods	MLP	75.74	75.74	77.30	75.41
	CNN	76.48	76.48	78.44	76.17
	LSTM	71.41	71.41	74.39	70.74
	LSTM (transposed)	76.96	76.96	79.03	76.55
	C-LSTM	75.81	75.81	78.36	75.45
	C-LSTM (transposed)	76.26	76.26	78.46	75.87
	Transformer	60.93	60.93	62.13	58.07
	Transformer (transposed)	76.48	76.48	78.18	76.21
	Cross-encoding Transformer	**77.85**	**77.85**	**79.82**	**77.38**

Table 5. The results of one-sided tests (P-values) in the comparison of deep learning architecture

		Methods							
		MLP	CNN	LSTM	LSTM (T)	C-LSTM	C-LSTM (T)	Transfor-mer	Transfor-mer (T)
Methods	CNN	$1.4e^{-1}$	-	-	-	-	-	-	-
	LSTM	$3.0e^{-9}$	$1.5e^{-9}$	-	-	-	-	-	-
	LSTM (T)	$5.2e^{-2}$	$2.8e^{-1}$	$\mathbf{2.5e^{-11}}$	-	-	-	-	-
	C-LSTM	$4.6e^{-1}$	$1.9e^{-1}$	$1.6e^{-6}$	$7.1e^{-2}$	-	-	-	-
	C-LSTM (T)	$2.6e^{-1}$	$3.8e^{-1}$	$6.9e^{-8}$	$1.8e^{-1}$	$\mathbf{2.6}e^{-1}$	-	-	-
	Transformer	$5.4e^{-27}$	$1.1e^{-26}$	$5.3e^{-17}$	$1.4e^{-28}$	$6.7e^{-25}$	$9.7e^{-24}$	-	-
	Transformer (T)	$1.6e^{-1}$	$5.0e^{-1}$	$1.8e^{-9}$	$2.7e^{-1}$	$1.9e^{-1}$	$3.9e^{-1}$	$\mathbf{8.7e^{-26}}$	-
	Cross-encoding Transformer	$\mathbf{3.0}e^{-3}$	$\mathbf{4.4}e^{-2}$	$\mathbf{7.3e^{-17}}$	$1.2e^{-1}$	$\mathbf{4.0}e^{-3}$	$\mathbf{2.2}e^{-2}$	$\mathbf{2.1e^{-30}}$	$\mathbf{1.3}e^{-2}$

(T): with transposed MFCC (proposed method 1)

NOTICE: e represents the base of the natural logarithm.

Table 4 より，MFCC を転置しない従来手法を比較すると，（他の手法より有意に高いわけではないものの）CNN が最も高い性能を示し，LSTM や Transformer などの時系列情報を考慮したモデルの性能が目立って劣る結果となった．これは，時系列情報にノイズが存在し，学習の妨げになっているため，あるいは時系列情報に過剰に適合したこと（過学習）が原因であると考えられる．

ここで，提案手法について考察する．MFCC を転置した場合（提案手法 1，Table 4 の(transposed)付き），LSTM，C-LSTM，Transformer の識別性能はすべて向上し，Table 5 より有意な結果となった．特筆すべきは，通常の MFCC では時系列に特化した学習を行う LSTM と Transformer の性能が大幅に向上したことである．いずれの時系列ニューラルネットワークも同じ傾向であったことから，肺聴診音の特徴抽出において時系列ニューラルネットワークを扱う場合には，周波数方向の系列に関する情報を学習させることで高い性能を引き出すことができた．第 4 章のように MFCC をノイズに頑健なアーキテクチャで扱う他にも，周波数に特化させた AI モデルの学習によって識別性能の向上が可能であると言える．ただし，本章の実験においては，音声信号は MFCC に変換された時点でピッチ成分などのノイズが除去されており，周波数に特化した場合においてもノイズの除去は必要であると考えられる．

また，従来の時系列ニューラルネットワークはいずれも識別性能が低かったのに対し，転置によって性能向上の効果が見られたことから，少なくとも MFCC においては時間情報よりも周波数情報の方が深層学習アーキテクチャにとって解析しやすく，有用な特徴が含まれていると言える．

次に，Cross-encoding Transformer（提案手法 2）の性能について述べる．Table 4，Table 5 によると，Cross-encoding Transformer は比較した深層学習アーキテクチャのうち最も高い性能を示し，転置 MFCC を入力とした LSTM（提案手法 1）を除くすべての手法に対して有意差が観察された．時間情報の取り込みを行っているアーキテクチャについては，時間情報の重みづけを柔軟にモデルで行っている Cross-encoding Transformer（77.85%），時間情報をパターンとして頑健に取り込んでいる CNN（76.48%），パターンの動的変化を学習している C-LSTM（75.81%），時系列の学習を中心に行っている LSTM・Transformer という順番に識別性能が並んでいることから，時系列情報は過学習のリスクはあるものの，識別上有用な情報も含んでいるということが改めて示唆された．

5.3.2　特徴マップ

肺聴診音の特徴量への理解をさらに深めるため，学習後の各深層学習アーキテクチャについて，出力層（識別）直前の特徴量を抽出した結果を Fig. 21 に示す．このとき，特徴量の圧縮には，UMAP（Uniform Manifold Approximation and Projection）[47]を採用した．Fig. 21 について，x 軸は圧縮された特徴量の第 1 成分を，y 軸は第 2 成分を表しており，散布図の色は各クラスラベルを表す（青 : normal，橙 : coarse crackle，緑 : fine crackle）．たとえば，Fig. 21 で MLP に注目すると，単純な MLP であっても，ラベルごとにクラスタが存在することが確認でき，識別可能な特徴がある程度抽出されていることが読み取れる．

UMAP を用いて，MFCC の転置あり・転置なしの結果を比較すると，転置がない場合に時系列ニューラルネットワーク（LSTM，Transformer）では，2 種類の異常データ（coarse crackle，fine crackle）間の特徴が混在していることが観察された．このことから，周波数系列の情報に注目した学習を行わない場合，異常データ間の識別が比較的困難になることが示唆される. 時系列ニューラルネットワークの機構を一部保有する C-LSTM については，MFCC の転置による特徴量の変化は観察されなかった．

また，時系列と周波数系列の両方の情報を柔軟に取り入れたモデル（Cross-encoding Transformer，C-LSTM）ほど，正常データと異常データ間でマージンを取った分類境界を決定できていることが読み取れる．しかし一方で，明瞭に分類境界が設定されているように

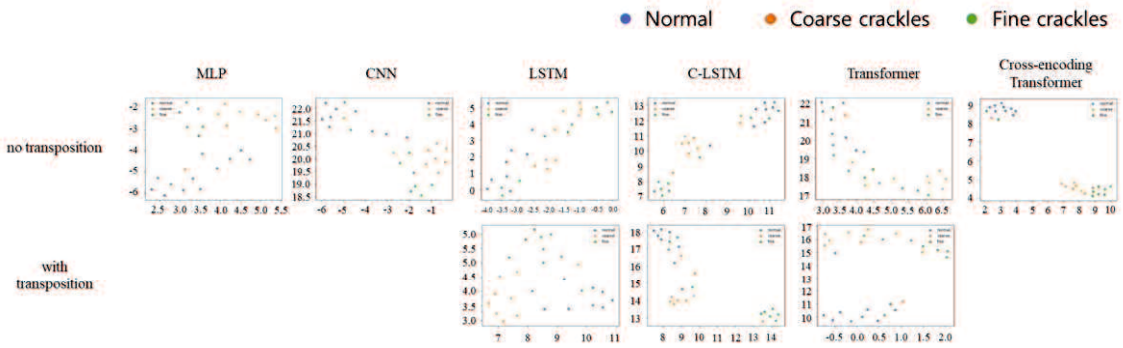


Fig. 21.　Visualizing test data features with UMAP obtained by each architecture

観察された場合でもクラスラベルの混入が確認された．これは，1）正常データ内のノイズを学習した結果，異常データ内のノイズを正常と判断してしまう，2）学習で使用した正常データに含まれる以上のノイズがテストの正常データで見られ，異常のクラスに分類してしまう，などが原因として挙げられる．本論文で扱った分類モデルは訓練データ内の分類境界を学習する識別関数であるため，訓練データ内の各クラスのデータの品質や，テストデータでのドメインの変化に大きく影響を受ける．これは，少数データのみによる分類モデルの大きな課題である．

5.3.3　Transformer による解釈性

提案手法 2 の Cross-encoding Transformer については，式（5.2）によってクラスラベルの予測結果について Transformer の Attention 重みを用いた解釈が可能である．ここで，式（5.2）で述べた Attention スコアをテストデータに対して算出し，音声信号に重ねて可視化したものを Fig. 22 に示す．

Fig. 22 において，緑色の箇所が Attention スコアの高い時刻である．音響強度のピークに影響されることなく，モデルが判断根拠とした時間を明示的に強調することに成功していることが確認できる．一方で，coarse crackle（Fig. 22 中央）と fine crackle（Fig. 22 右側）の注目箇所時刻に大きな重なりが見られ（たとえば 1.5 秒～2.0 秒付近）データに関係なく特定の時間領域に注目してしまう問題があるなど，正確な解釈が難しいことが確認された．

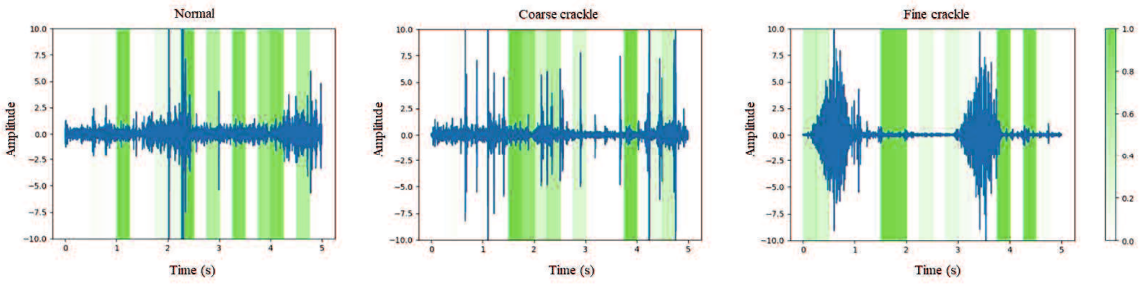


Fig. 22.　Visualization of attention weights with Cross-encoding Transformer

Attention スコアは Transformer が予測時に注目した箇所を示しており，Attention が高い箇所にその音声の重要な特徴が必ずしも含まれることを保証しない．さらに，1) Attention 後に全結合層が用いられているため，直接的な推論根拠ではないということ，2) Transformer は複数の Attention head で構成されており，それぞれの Attention head で注目している箇所が異なること，などから，本章で提案した Attention スコアのみでは分類モデル全体の挙動を説明するには至っていない．そのため，より信頼性の高い解釈性を求める場合は，単純なアーキテクチャを用いることによって説明可能なモデル構成にすることや，分類モデルを時間系列ごとに独立させて学習する，などの他のアプローチが求められ，さらなる工夫が必要である．

5.4 まとめ

本章では，時系列ニューラルネットワークの性能を引き出すための特徴抽出法の検討と，データ数の少ない肺聴診音識別に対して有用な特徴抽出法の整理を行った．

発展的な前処理法である転置 MFCC や，より時間情報と周波数情報を考慮できる深層学習アーキテクチャである Cross-encoding Transformer を提案し，いずれも有意に効果があることを確認した．また，実験において，周波数情報が肺聴診音の識別において重要であること，時間情報は過学習のおそれが存在すること，分類モデルの予測結果の解釈性については課題が残ること，などについて述べた．肺聴診音の特徴抽出に関するより深い洞察を提供するとともに，近年あらゆる深層学習分野で基盤モデルとして広く用いられている Transformer[34]についても肺聴診音の効果的な特徴抽出法を提案できたことは，将来的な診断支援技術の発展に向けて価値がある．

また，本章においては取り扱わなかったが，周波数情報の重要性は他の音声信号向け前処理手法に対しても一般化されるのか，一般化される場合，どのような音声信号向け前処理手

法が肺聴診音の AI モデルの学習に適した特徴を備えているのかは検討の余地がある．

分類モデルについては，特徴量エンジニアリングを継続しても識別性能の頭打ちが見込まれるとともに，モデルの解釈性の課題を解決することが難しい．次章では，これまでに行った肺聴診音の特徴抽出に関する議論を踏まえつつ，異常検知タスクについて検討を深める．

第 6 章　深層ニューラルネットワークを用いた肺聴診音の異常検知

第 4 章および第 5 章では，肺聴診音の分類モデルについて検証し，前処理やアーキテクチャ，学習方法の改良によって少ないデータでも効果的な学習ができることを示した．本章では，異常検知タスクに注目し，異常検知モデルの開発・検証を行う．異常検知は正常データのみをもとに学習するため，異常データの収集が比較的困難な医療現場においても，データの少ないタスクをさらに効果的に解決する方法である．複数の深層学習異常検知モデルについて性能を比較し，肺聴診音に対しての効果を検証する．

6.1　背景と目的

肺聴診音に関する AI モデルを支援診断に応用するにあたって，重要な観点の 1 つは，聴診をどのようなタスクとして AI に解かせるべきかである．本論文で扱う分類・異常検知をはじめとする様々なタスクについて，複数のアーキテクチャや学習方法を組み合わせたタスク特化のモデル構造が提案されている．実用に向けてどのような AI モデル（タスク）が適しているのかについては，得られるアウトプットの形式や精度などを加味する必要がある．

第 4 章と第 5 章では，肺聴診音について分類タスクでの検証を行った．少数データを考慮した特徴抽出は，分類性能の向上に効果があることが示されたものの，5.3.2 で述べたように，少数データを用いた分類モデルは訓練データの影響を特に受けやすく，雑音の多い肺聴診音においてよい精度を得ることが難しいケースがある．

ここで，少数のデータからより効果的な特徴抽出が見込めるタスクとして，異常検知が挙

げられる．異常検知タスクでは，正常データの特徴分布から外れたものを異常とみなすため，少数のデータに対して複数のラベルを学習する分類モデルと比較して，予測バイアスが発生する可能性の軽減が期待できる．

本章では，異常検知の枠組みにおいて，深層学習を用いた特徴抽出と，その特徴を用いた異常検知手法の開発・検証を行う．すなわち，正常と異常の学習データを用いた分類問題の枠組みではなく，正常データのみで学習が可能な異常検知方式の枠組みに焦点をあて，より精度の高いモデルの構築を目指す．

本章で扱う異常検知手法では，正常な肺聴診音データから特徴を抽出し，抽出された特徴をもとに異常検知を行う．異常検知システムの予測精度に影響を与える要素は大きく 2 つあり，1）異常検知モデル内部の特徴抽出，2）異常検知モデルの目的関数に起因する正常分布の学習メカニズムである．異常検知モデルとして，2.4.3.1 で説明した DAGMM（Deep Autoencoding Gaussian Mixture Model）と 2.4.3.2 で説明した Efficient GAN を肺聴診音データに適用し，これらの異常検知モデルを先述の観点から改良する．提案手法 1 では DAGMM の特徴抽出（圧縮ネットワーク）の構造を変更し，異常検知において重要な特徴の抽出が可能な圧縮ネットワークの構造を提案する．提案手法 2 では Efficient GAN に GMM を組み合わせ，目的関数の改良を施し，DAGMM や Efficient GAN らと比較することによって異常検知性能の高いモデル構造を提案する．これら 2 つの提案手法について，第 4 章・第 5 章と同様に MFCC を用いた肺聴診音データを入力とし，従来の異常検知手法と比較する．

6.2 方法

異常検知では，比較的豊富に取得可能な正常データを最大限活用するため，正常データが持つ特徴の分布を正確に捉える特徴抽出及びモデル構造が必要となる．これには，特徴抽出

を行うアーキテクチャが 1）正常データからまとまった共通点を見出し，2）その共通点が異常データに見られないことが必要となる．このため，1）については DAGMM 内で正常分布を生成する GMM の構造に着目，2）については Efficient GAN のような疑似的な異常データを生成して真の正常データと差異を学習する構造に着目し，肺聴診音の異常検知に適したモデル構造を検討する．

さらに，2.4.3 で述べたこれらの異常検知手法について，より肺聴診音に特化した特徴抽出を行う．提案手法は 2 つに大別され，1 つ目は DAGMM の改良である．クラスタリングの目的に適合した特徴抽出が可能な DAGMM について圧縮ネットワークの改良を行う．2 つ目は Efficient GAN に DAGMM のアルゴリズムを組み合せた新たなモデル構造を提案する．

ここで，対象とするタスクは，正常データ（Table 1 のうち Normal），異常データ（Table 1 のうち Coarse crackle，Fine crackle を含む Abnormal）に対する異常検知である．実験では，Table 2 のうち 140 個の正常データをシャッフル後に 14 分割して交差検証を行い，1 分割あたりの訓練用正常データ数は 130，テスト用正常データ数は 10，異常データ数は 79 とした．これにより，テストデータは各 fold 間で重複しておらず，データセット全体での評価が可能になる. 評価指標は, ROC 曲線下の面積である AUC（Area under the Curve）スコアとした．

6.2.1 DAGMM の改良（提案手法 1）

異常検知モデルのうち, 正常データからまとまった共通点を見出す機構について, 改良する. これを行うのは DAGMM のうち, 圧縮ネットワーク（Fig. 8 中の Compression network）と呼ばれる箇所である．2.4.3.1 で説明したように，この機構は自己符号化器によって成り立っており，第 4 章・第 5 章で述べたような内部のアーキテクチャの検討が必要になる．

本節では．DAGMM の圧縮ネットワークを 3 種類のネットワーク：1）CAE，2）LSTM-

AE，3) C-LSTM-AE に置き換え，肺聴診音の異常検知性能向上に効果的な特徴抽出アルゴリズムを構築する．これら 3 種類のネットワークに用いられる 1) CNN，2) LSTM，3) C-LSTM は，いずれも抽出できる特徴が大きく異なるとともに，第 4 章や第 5 章において識別に有用であることが示されており，異常検知においても従来の DAGMM で用いられている MLP より優れた性能を示すことが期待できる．CAE，LSTM-AE，C-LSTM-AE の 3 種類の自己符号化器は，それぞれ構造や特徴が異なる．本節では，それぞれの詳細について説明する．

6.2.1.1 CAE

DAGMM の圧縮ネットワークに適用する CAE は，4.2.1.1 で説明した CAE（Fig. 13）と同一の構造とした．圧縮ネットワークへの入力データxは，5 秒間の教師ラベルなし音声データ（本章においては訓練データとなる正常データ）に MFCC を適用したものである

このとき GMM の入力に用いる特徴量Z_Cは，Encoder で符号化された 3 次元の特徴量（2×2×80）を 1 次元に変換（flatten）したものである．CNN の畳み込みとプーリングを用いた CAE の構造により，音声信号にノイズが混入した場合でも特徴量への変換に及ぼす影響を軽減できる．

6.2.1.2 LSTM-AE

DAGMM の圧縮ネットワークに適用する LSTM-AE の構造を Fig. 23 に示す. LSTM-AE は LSTM で構成された自己符号化器であり，本論文で扱う LSTM-AE については Fig. 15（4.2.1.2）で扱った構造をベースにしている．CAE の場合と同様に，自己符号化器の学習には教師ラベルなしの音声データに MFCC を適用したものを入力としている．ここで，GMM の入力に用いる特徴量は Encoder から出力される 1 次元特徴量Z_C（200）である．

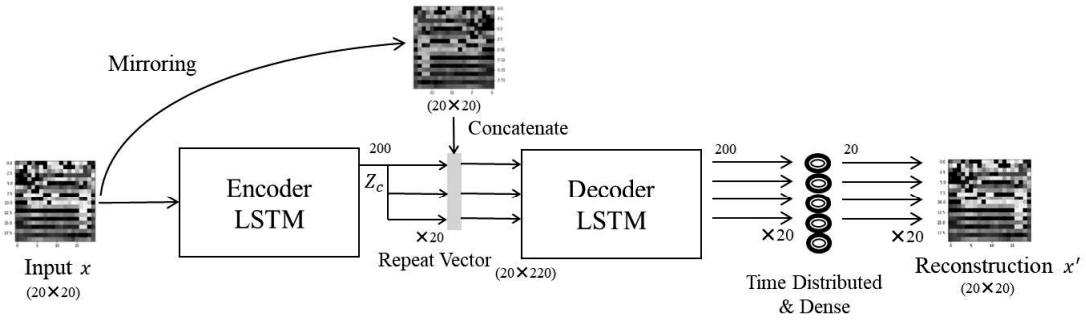


Fig. 23.　Structure of LSTM-AE

6.2.1.3　C-LSTM-AE

C-LSTM-AE は C-LSTM をベースとした自己符号化器であり，4.2.1.3 で述べた通り，使用される Decoder には確立された手法が存在せず，第 4 章・第 5 章においても十分な検討はされていなかった．そこで，本節で扱う C-LSTM-AE では 2 種類の Decoder：1）Convolutional Decoder，2）LSTM Decoder を用いた構造を提案する．それぞれの構造の違いについて，次節で述べる．

(1) Convolutional Decoder

DAGMM の圧縮ネットワークに Convolutional Decoder を用いた場合の構造（以降，C-LSTM-AE （conv）と呼称する）を Fig. 24 に示す．Decoder は 2 つの逆畳み込み層と 2 つのアップサンプリング層で構成されており，4.2 で提案した C-LSTM の自己符号化器（Fig. 16）よりも CAE で使用した Decoder に近い構成になっている．Fig. 16 とは異なり，Max

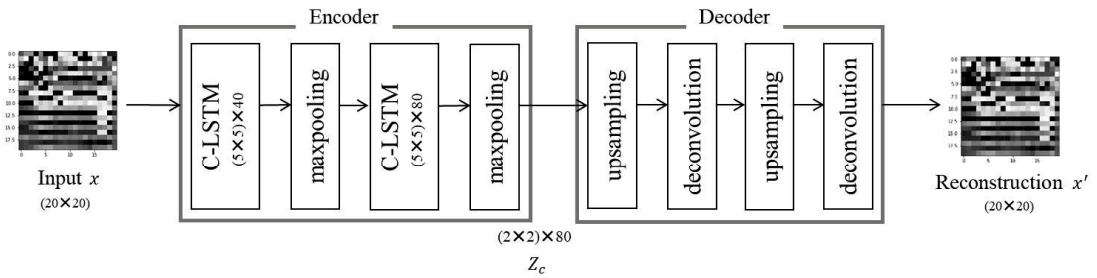


Fig. 24.　Structure of C-LSTM-AE（conv）

pooling が採用されたのは，第 4 章・第 5 章において C-LSTM が音声信号の時系列の影響を強く受けており，学習 Loss が安定しないケースが観察されたためである．Max pooling によって畳み込み後の情報量を削減することにより，時系列変化による影響が軽減され，学習の安定が期待できる[48].

ここで，Encoder における C-LSTM の出力として得られる特徴マップは，MFCC 内の周辺情報を含んでおり，かつ CNN と同様の特徴サイズであるため，逆畳み込み層を用いることによって CAE での復元と同様に，局所的な特徴に着目した復元を行える．そのため，時系列情報と周辺情報を考慮した特徴表現を得ることができる．なお，GMM の入力に用いる特徴量は，Encoder で符号化された 3 次元特徴量（2×2×80）Z_cを 1 次元に変換（flatten）したものである．

(2) LSTM Decoder

DAGMM の圧縮ネットワークに LSTM Decoder を用いた場合の構造（以降， C-LSTM-AE（LSTM）と呼称する）を Fig. 25 に示す．本構造では，Fig. 24 と同様の Encoder（C-LSTM とプーリング層）によって符号化した後，LSTM Decoder によって時系列情報を重視した復元を行う．ここで，本ネットワークの Decoder は LSTM-AE（Fig. 15）で使用した Decoder と同様の構成になっている．

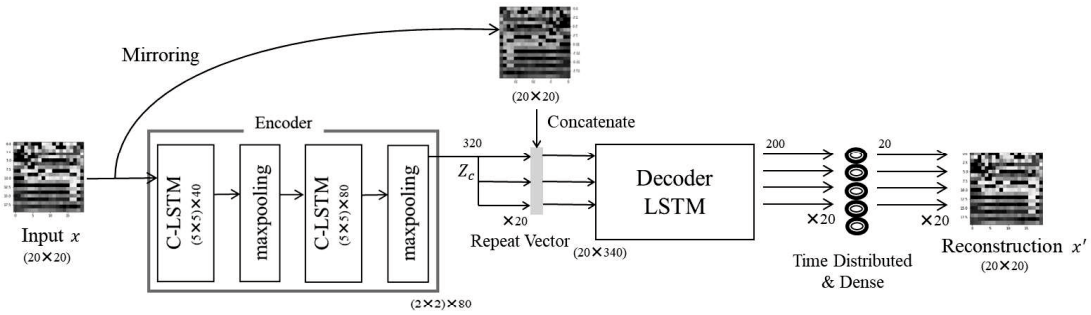


Fig. 25.　Structure of C-LSTM-AE（LSTM）

Decoder を LSTM とする利点は以下のとおりである．異常検知タスクにおいて，本論文で扱う正常データは無症状患者の肺聴診音データであり，少量ながらも外部環境によるノイズが存在する．肺聴診音データを入力としたとき，C-LSTM ベースの Encoder では，ノイズは一定まで軽減された状態で，時系列の情報を含めた特徴量の抽出が行われる．ここで LSTM を Decoder とすることにより，より時系列に着目した特徴量が抽出されることが期待できる．本章の C-LSTM の Encoder では max pooling によるさらなるノイズ軽減を行っているため，6.2.1.2 で述べた LSTM-AE と比較して入力データ$\boldsymbol{x}$に含まれるノイズに対して頑健な特徴抽出が行われる．

なお，GMM の入力に用いる特徴量は，Encoder で符号化された 3 次元特徴量（2×2×80）Z_Cを 1 次元に変換（flatten）したものである．

6.2.2 Efficient GAN の改良（提案手法 2）

異常検知モデルにおいて，異常データにはない共通の特徴量を抽出する手法として，Efficient GAN を検証する．GAN では正常データに類似するデータを内部で生成し，それらを見分けるための学習を行う．

本節では Efficient GAN を改良し，肺聴診音の異常検知性能向上に効果的なモデルを構築する．具体的には，DAGMM で用いた GMM を Efficient GAN の学習に追加することで性能向上を図る．さらに GMM を追加した Efficient GAN について，提案手法 1 の「C-LSTM と LSTM を用いた改良」を特徴抽出に施し，改良後の Efficient GAN についても提案手法 1 の導入が有効であることを示す．以下，それぞれの詳細な構造を説明する．

6.2.2.1 Efficient GAN with GMM

Efficient GAN with GMM の構造を Fig. 26 に示す．Efficient GAN with GMM は，Efficient GAN の構造をベースに，DAGMM の構造を取り入れた異常検知モデルである．

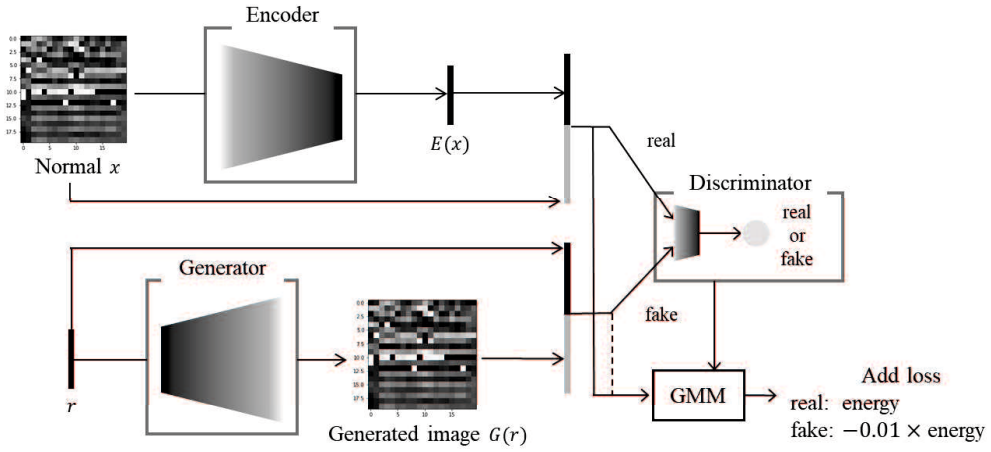


Fig. 26.　Structure of Efficient GAN with GMM

Fig. 26 右下部に示すように，入力データxとxを符号化した特徴量$E(x)$を結合した新たな特徴量を生成し，それに対する Discriminator の識別結果から GMM でガウス分布を作成している．また，DAGMM と同様に，GMM で計算されたエネルギー（式（3.24）の右辺第 2 項）と正則化項（式（3.24）の右辺第 3 項）を Efficient GAN の目的関数に加えている．このときエネルギーに関する項について，本物のデータ（入力xとxの符号化特徴量$E(x)$のペアを結合した特徴量）ではエネルギー値をそのまま目的関数に代入した．一方，フェイクデータ（ランダムノイズrとrから Generator で生成したデータ$G(r)$のペアを結合した特徴量）はエネルギーに係数a（本手法では-0.01）を乗算したものを代入した．モデルの学習では勾配降下法によって目的関数の最小化を目指すため，フェイクデータのエネルギーにマイナスを乗算することで，本物のデータ（正常データの特徴量）についてはエネルギー（異常スコア）が小さくなるよう，フェイクデータについてはエネルギーが高くなるように学習を行っている．aの絶対値を小さい値に設定している理由は，フェイクデータのエネルギーを高くする制約が強くなることにより，Loss が収束しなくなる現象を防ぐためである．このモデル構造によって，本物のデータはよりまとまった分布となり，Efficient GAN 内部で生成

されるフェイクデータは正常データの分布から外れるようになる．異常検知の評価時には，オリジナルの Efficient GAN と同様に，Anomaly Score（式（3.25））を用いる．

6.2.2.2 Efficient GAN with GMM （C-LSTM）

Efficient GAN with GMM に C-LSTM-AE（LSTM Decoder）を組み合わせた構造を Fig. 27 に示す．以降，本モデル構造を Efficient GAN with GMM（C-LSTM）と呼称する．Fig. 27 上段に示すように，Fig. 26 との違いは特徴量$E(x)$を抽出後に Decoder が追加されている点にある．C-LSTM-AE（LSTM Decoder）では，入力xと再構成したデータ$D(E(x))$についての再構成誤差（L2 ノルム）を目的関数に追加している．ここで，Encoder は C-LSTM-AE（LSTM）（Fig. 25）の Encoder をベースとし，Decoder と Generator は C-LSTM-AE（LSTM）（Fig. 25）の Decoder 部分をベースとしたネットワークである．このとき，Fig. 25 と異なり新たに Decoder を追加している理由は，Discriminator に入力する本物画像とフェイク画像の質を同等のものとし，構造上の問題によって両者の画質に優劣が生じないようにするためである．Generator によるフェイクデータ生成では，LSTM-AE で使用していた鏡映データ [42] を用いることができず生成画像が粗くなってしまうため，

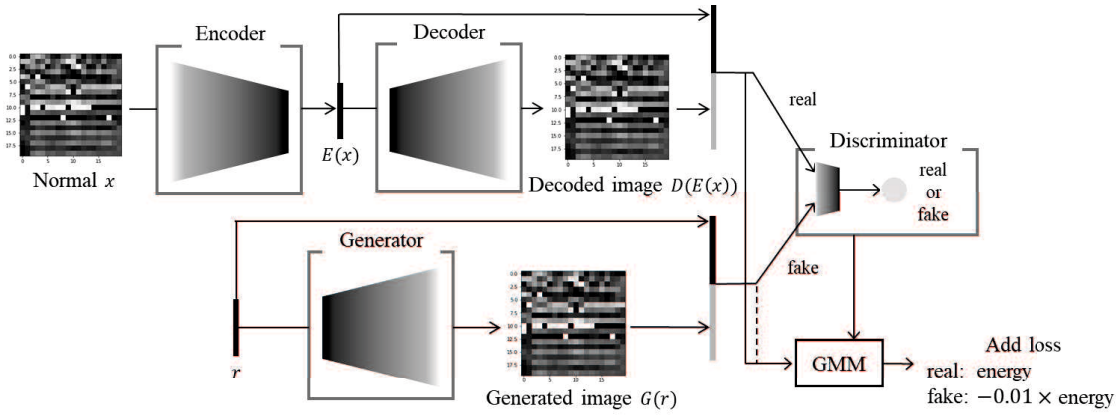


Fig. 27. Structure of Efficient GAN with GMM
（C-LSTM-AE（LSTM Decoder））

Discriminator による本物とフェイクの識別が容易になってしまう．したがって本物のデータについても，Decoder で再構成したデータ$D(E(x))$を特徴量として扱う構成とした．

C-LSTM-AE を適用した Efficient GAN では，Encoder に C-LSTM，Decoder と Generator に LSTM を用いることによって，DAGMM における特徴抽出の改良に近い効果が期待できる．なお，異常スコアの算出は Anomaly Score（式（3.25））で行われ，この計算時には Decoder が不要であるため用いない．

6.3 結果

従来手法及び提案手法によって得られた平均 AUC と標準偏差を Table 6 に示す．また，全手法の片側 T 検定（P 値）の結果を Table 7 に示す．Table 6 より，提案手法 1 の DAGMM with C-LSTM-AE（LSTM）（6.2.1.3（2））で得られた AUC（0.9439）が全ての手法のうち

Table 6. Mean AUC and standard deviation of anomaly detection models in Chapter 6 by 14-fold cross validation

	Methods	Mean AUC	Standard deviation
Conventional method	DAGMM	0.8730	0.056
	Efficient GAN	0.8480	0.044
Proposed method 1	DAGMM with CAE	0.9193	0.051
	DAGMM with LSTM-AE	0.8757	0.084
	DAGMM with C-LSTM-AE (conv)	0.9390	0.045
	DAGMM with C-LSTM-AE (LSTM)	**0.9439**	0.036
Proposed method 2	Efficient GAN with GMM	0.8765	**0.035**
	Efficient GAN with GMM (C-LSTM)	0.9166	0.045

(conv): Convolutional Decoder, (LSTM): LSTM Decoder, (C-LSTM):C-LSTM-AE with LSTM Decoder

Table 7. The results of one-sided t-tests (P-values) in the comparison of anomaly detection models

		Methods						
		DAGMM	Efficient GAN	DAGMM with CAE	DAGMM with LSTM-AE	DAGMM with C-LSTM-AE (conv)	DAGMM with C-LSTM-AE (LSTM)	Efficient GAN with GMM
Methods	Efficient GAN	$6.87e^{-2}$	-	-	-	-	-	-
	DAGMM with CAE	$1.06e^{-3}$	$3.29e^{-4}$	-	-	-	-	-
	DAGMM with LSTM-AE	$4.48e^{-1}$	$1.46e^{-1}$	$1.43e^{-2}$	-	-	-	-
	DAGMM with C-LSTM-AE (conv)	$1.58e^{-4}$	$1.56e^{-5}$	$2.29e^{-4}$	$1.76e^{-3}$	-	-	-
	DAGMM with C-LSTM-AE (LSTM)	$1.43e^{-5}$	$3.22e^{-6}$	$3.39e^{-4}$	$1.41e^{-3}$	$1.20e^{-1}$	-	-
	Efficient GAN with GMM	$4.13e^{-1}$	$2.70e^{-3}$	$5.86e^{-3}$	$4.88e^{-1}$	$1.83e^{-4}$	$4.31e^{-5}$	-
	Efficient GAN with GMM (C-LSTM)	$1.99e^{-3}$	$2.58e^{-4}$	$3.68e^{-1}$	$9.33e^{-3}$	$3.90e^{-3}$	$1.88e^{-3}$	$4.92e^{-3}$

(conv): Convolutional Decoder, (LSTM): LSTM Decoder, (C-LSTM):C-LSTM-AE with LSTM Decoder

NOTICE: e represents the base of the natural logarithm.

最も優れている．また，C-LSTM を用いた特徴抽出は提案手法 2 で示した Efficient GAN の改良（6.2.2.2）においても有効だった（Table 6 において AUC が 0.8765 から 0. 9166 に向上，かつ Table 7 より P 値 4.92×10^{-3}で有意差あり）．このことから，肺聴診音の異常検知タスクにおいて，提案した C-LSTM の特徴抽出が有効であり，モデルを問わず周辺情報と時系列情報のどちらも考慮した特徴抽出の有用性を示した．周辺情報と時系列情報を扱う C-LSTM-AE のうち LSTM Decoder が優れていた理由は，時系列情報に特化した復元構造を持つためである．第 4 章・第 5 章で示した通り，LSTM はノイズに弱いため，異常データは LSTM で復元が難しくなり，異常データと判定しやすくなる．また，時系列情報が有用であるにも関わらず，LSTM-AE を用いた DAGMM の性能が大きく向上しなかった理

由として，LSTM のみの Encoder ではノイズに脆弱であることが挙げられる．時系列情報には，Table1 の特徴では判別できない肺音の重要な特徴（肺音の発生パターンの情報）が含まれている．しかし，LSTM は時刻順に入力を一つずつ処理する仕組みであるためノイズの影響を受けやすく，抽出された特徴量は分散が大きくなると考えられる．そのため，第 4 章や第 5 章の結果からも，肺聴診音において時系列情報をそのままアーキテクチャで扱う場合には，畳み込みとプーリングを併用する必要があると言える．

提案手法 2 の Efficient GAN の改良では，Table 6 より，GMM を Efficient GAN の学習構造に取り入れた Efficient GAN with GMM のスコア（0.8765）が，従来の Efficient GAN（0.8480）より高い値を示した．このことから，GMM を用いることで異常検知性能が向上し，かつ，DAGMM のモデル構造は GAN ベースの異常検知モデルにも転用可能であることがわかった．ここで，GMM の適用によって性能が向上した理由について考察する．Efficient GAN では学習の過程で，訓練データ（正常データ）に類似したデータが生成された場合でもフェイクデータ（異常データ）と識別するよう学習しているため，Efficient GAN は未知のデータに対して感度が高くなると考えられる．このとき，正常な肺音の特徴を Efficient GAN で学習できていない場合，正常データも異常と判定する可能性が高くなる問題が発生する．そこで，正常データについてまとまった特徴分布を形成，つまり GMM を用いた正則化を行うことでこの問題を解決している．これにより，モデルの学習を経て正常な肺音間で揺らぎの少ない特徴を捉えることが可能になり，検証時に正常データを異常データと誤判定する割合を軽減したことで，AUC の向上につながった．

特徴抽出の改良について，C-LSTM の適用による性能の向上は Efficient GAN より DAGMM が顕著であった．ここで，C-LSTM による特徴抽出がモデル（DAGMM と Efficient GAN with GMM）に与えた効果の違いについて考察する．C-LSTM を用いた特徴抽出は，本論文で扱った 2 つのモデルの異常検知性能を向上させたが，同程度の性能向上にはならず，標準偏差に与える効果についても違いが見られた．特に，標準偏差の違いは大きく，C-

LSTM の適用によって，DAGMM では標準偏差が下がり（Table 6 より 0.056 から 0.036 に低下），Efficient GAN with GMM においては標準偏差が上がる結果となった（Table 6 より 0.035 から 0.045 に向上）．この理由として，pooling の処理を用いたことで GAN の性能が安定していないことが挙げられる[49]．そのため，本論文で提案した C-LSTM による特徴抽出は DAGMM に適した構造となっており，Efficient GAN with GMM の特徴抽出については改良の余地がある．たとえば，GAN の学習の妨げにならない特徴抽出を実現できればさらなる異常検知性能の向上が期待できる．

本章で行った提案の範囲において，DAGMM と Efficient GAN を比較すると，DAGMM のような「正常データの特徴量を可能な限り圧縮する」ことの有効性が顕著に示される結果となった．一方で，Efficient GAN が勝る fold も確認されており，GMM の学習構造が Efficient GAN による異常検知モデルにも転用可能であることから，他の GAN ベースの異常検知モデルについても今後の改良や応用が期待できる．

6.4 まとめ

本章では，DAGMM の特徴抽出の改良，及び Efficient GAN の改良を行い，肺聴診音データを用いて評価を行った．実験の結果より，提案手法はいずれも従来手法の性能を上回っており，特に提案手法 1 の DAGMM with C-LSTM-AE（LSTM）はどの異常検知モデルよりも性能が優れていた．このことから，肺聴診音の異常検知タスクにおける特徴抽出アーキテクチャの重要性を示し，同時に，畳み込みと時系列情報を同時に扱うことの有用性を示した．提案手法 2 では Efficient GAN に，GMM を追加すること及び C-LSTM を特徴抽出に用いることの有用性を示し，GMM を用いた特徴量の正則化と，C-LSTM を用いた特徴抽出がモデルに依らず有効であることを示した．

本章で提案した異常検知モデルの提案手法は高い AUC が計測されており，分類モデルの正解率や，各クラスラベルの学習データを確保する難しさを考慮すると実用では異常検知モデルが役立つ可能性がある．一方で，本論文で扱った DAGMM，Efficient GAN は第 4 章で扱ったモデルと同様にブラックボックス構造であり，「なぜ異常とみなされたのか」については異常スコアの数値以上の解釈を行うことができない．さらに，今回性能が高かった機構は特徴量を圧縮するものであり，聴診環境の変化や機器の変更に耐えられない可能性があり，異常検知においても解釈性を持つモデルが重要である．

次章では，解釈性を考慮した肺聴診音の異常検知モデルについて検討を行う．

第 7 章　解釈性を考慮した肺聴診音の異常検知

第 6 章において，肺聴診音の異常検知モデルは高い性能が確認された．一方，ここまでに述べた異常検知モデルが出力する結果は異常スコアのみであり，異常検知タスクにおいても結果の解釈性について課題が残っている．本章では，解釈性を持つ異常検知モデルについて開発・検証を行う．解釈性を担保することに加えて，精度を保つための特徴抽出や異常検知モデルの構築について検討する．

7.1　背景と目的

本論文が目的とする肺聴診音の CAD 技術において，「モデルの結果の解釈性」は重要なテーマである．診断の補助として深層学習モデルを用いる場合は，収音環境の変化や機器の劣化など，様々な外的要因によってモデルの予測が変動する可能性がある．そのため，解釈性が不足しているモデルを採用すると，医師の診断と乖離があった場合にモデルの予測の妥当性を検証することが困難であり，かえって診断の妨げとなるリスクがある．さらに，本論文で扱う肺聴診音については，実際の聴診環境で収音されたデータが少ない場合が多く，異常検知モデルが，実際の異常音とは無関係な予測バイアス（たとえば，データの 1 秒目にノイズがあると異常とみなされやすい，など）を持ちうるリスクがあるため，予測結果に解釈性を持つ異常検知モデルは診断支援の実現に向けて大きな価値がある．

本論文がここまで扱った深層学習モデルは，大半がその内部構造の複雑性からブラックボックスであり「なぜ異常と判断されたのか」を示すことができない．第 5 章においては Transformer の Attention 機構による可視化を行ったが，解釈が難しく，Attention の出力値を解釈可能な値として代替するには不十分であることを示した．5.2.2 で述べた通り，一般的に MFCC のような 2 次元データに対して，深層学習で解釈性を担保するアプローチと

して挙げられるのは，SHAP[45]や LIME[46]などの深層学習アーキテクチャ内部の重みを可視化するアプローチである．しかし，SHAP 値では，2 次元情報である MFCC に対して 2 次元のマップとして可視化を行うため，肺聴診音の異常検知において，どの時間が予測に寄与したかを可視化することは困難である[50]．解釈可能な異常検知モデルとして，異常音発生期間のセグメンテーションを学習するモデルも存在するが，収集が難しい肺聴診音データについて，追加でセグメンテーション用のラベル付きデータを用意することは医師の負担が大きく，かつ音声信号中に多種多様な正常音・異常音・ノイズが混在することから，肺聴診音の CAD には適さない．

そこで，本章では，正常な肺音データのみを用いた異常検知モデルについて，精度を維持しつつ，解釈可能な異常検知手法を構築する．具体的には，肺聴診音データを時間窓に分割し，それぞれの時間窓について異常検知の推論を行った後，その結果を集約することで，異常信号の発生期間を判定できる手法を提案する．このとき，入力データを分割して学習するため，ここまで第 4 章・第 5 章・第 6 章で行ったものとは異なる特徴抽出アプローチが求められる．そこで，本章では TDA によるトポロジー特徴を採用した．さらに，時間窓に分割した入力データは次元数が小さくなるため，異常検知モデルについても機械学習ベースの異常検知モデルである Isolation Forest（2.4.3.3 を参照）を用いる．さらに，第 6 章までで時系列情報の有効性を確認できていることから，上記に加えて， 相関係数と音声 IDF（Inverse Document Frequency）によって時系列情報を加味させ異常検知精度の向上を図る．本章では，これらの提案手法と従来の異常検知手法について精度を比較するとともに，提案手法の解釈性について評価する．

7.2 方法

本章では，解釈性を考慮したモデルの検討を行うが，肺聴診音の異常検知について最低限必要となる情報は「どの時刻の情報を異常とみなした根拠とするか」である．これを実現するためには，各時刻について「異常である」という判断が示される必要があるが，第5章で述べたように，これはTransformerなどのAttentionの値などで代用することができない．

そこで，本節では解釈性を考慮した異常検知モデルとして，2つの手法を提案する．1つ目は，フーリエ変換のトポロジー特徴とIsolation Forestを組み合わせた異常検知手法であり，時間窓ごとに異常検知を行うことによって予測結果の直接的な解釈が可能になる．2つ目は，1つ目の手法をベースとした，相関係数と音声IDFを組み合わせた異常スコアの改良である．

ここで，対象とするタスクは，第6章と同様，正常データ（Table 1のうちNormal），異常データ（Table 1のうちAbnormal．Coarse crackle，Fine crackleを含む）に対する異常検知である．実験では，Table 2のうち140個の正常データをシャッフル後に14分割して交差検証を行い，1分割あたりの訓練用正常データ数は130，テスト用正常データ数は10，異常データ数は79とした．評価指標は，ROC曲線下の面積であるAUC（Area under the Curve）とした．

7.2.1 トポロジー特徴量を用いたIsolation Forestによる異常検知（提案手法1）

第6章までに導入した深層学習アーキテクチャでは，前処理後の音声信号であるMFCCに対して，周辺情報や時系列などの様々な内部パターンに着目した特徴抽出の変換を行っていた．このとき，MFCCから特徴量への変換，そして特徴量から予測結果の算出は複雑な変換が行われており，入力のどの時刻が予測に貢献したのか解釈することは難しい．そこで，本節では，時刻（時間窓）ごとの独立した学習を行う異常検知モデルとする．

本節で扱うトポロジー特徴量と Isolation Forest を用いた異常検知モデル（提案手法 1）の構造を Fig. 28 に示す．Fig. 28 に示すように，提案手法 1 では，元の音声信号から分割された各時間窓に対して異常検知を行い，集約されたスコアを異常の度合いとみなす．

まず，異常検知モデルの学習方法について説明する．最初に，訓練データ（正常な肺聴診音データ）について時間窓を作成し，フーリエ変換を適用する．本論文では，サンプリング周波数を 2000Hz，ウィンドウサイズを 32 に設定しているため，フーリエ変換時の周波数ビンは 62.5Hz である．次に，フーリエ変換によって得られたそれぞれの時間窓がもつ各周波数ビンのパワースペクトルから，トポロジカル特徴を抽出する．具体的には，Takens の埋め込み定理を用いて，音声信号のパワースペクトルから疑似アトラクタを構成することで TDA を行いやすいデータ形式に変換し，その後，パーシステンス図とパーシステンスエントロピーについて計算する（2.2.3 を参照）．この一連の処理によって求められたトポロジー特徴量を，Isolation Forest の学習データとして使用する．このとき，TDA を実施する前にフーリエ変換を用いる理由は，2.2 で述べたように，元の音声信号は信号量が多く，学習データに過剰に適合する可能性があるためである．

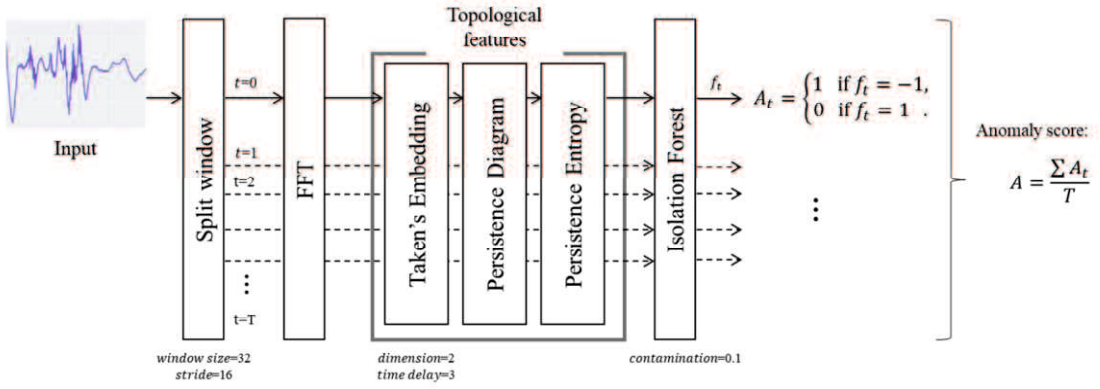


Fig. 28. Architecture of anomaly detection using Isolation Forest with topological features

ここで，トポロジー特徴を用いる理由について説明する．第 5 章において，周波数情報に見られる特徴パターンが特に肺聴診音の識別において重要であることが示され，このとき，MFCC を転置した周波数情報を LSTM・Transformer で学習する機構が有効であった．そのため，時間窓に含まれる周波数帯域の包絡を，時系列とみなして特徴パターンを学習することは，肺聴診音の識別において効果があると言える．そこで，時間窓ごとに異常検知を行うアプローチをとるにあたって同様の枠組みを検討した．ここで，フーリエ変換後の各時間窓に対して TDA を行うことにより，フーリエ変換で表現される振幅スペクトラムの位相的な特徴量を抽出する．これは Isolation Forest のような時系列を考慮しないアルゴリズムでも学習可能な「周波数帯域別の音圧レベルを系列とみなした特徴量」であり，Table 1 で示した正常データの特徴量の分布の概形（たとえば，正常データでは 150Hz～600Hz に振幅スペクトラムの山が存在する，など）を学習できる．

予測時には，前述の処理をテストデータに適用し，前述の学習を行った Isolation Forest を用いて予測を行う．時刻tにおける Isolation Forest の予測値をf_t（正常とみなしたデータは 1，異常とみなしたデータは-1が出力される)とすると，時刻tにおける異常スコアA_tは次式で与えられる．

$$A_t = \begin{cases} 1 & if\ f_t = -1 \\ 0 & if\ f_t = 1 \end{cases}. \tag{7.1}$$

次に，各時間窓の異常スコアを表すA_tを基に，音声信号全体の異常スコアを決定するAについて計算する．ウィンドウ総数をTとすると，Aは次式で与えられる．

$$A = \frac{\sum A_t}{T}. \tag{7.2}$$

また，本章ではトポロジー特徴の有効性について検証するため，上記の異常検知手法に加えて，Isolation Forest への入力をフーリエ変換後の特徴量とした場合についても実験を行った．

7.2.2 相関係数と音声 IDF によるスコアリング（提案手法 2）

7.2.1 で説明した提案手法 1 では，時間窓ごとに閉じたデータに対する独立した推論を行っており，周波数情報のみによって異常検知を行っている．そのため，肺聴診音の持つ時系列情報については考慮されていない．しかし，第 4 章・第 5 章において「時系列情報が音声の識別において役立ちうる情報であること」，第 6 章において「異常検知タスクにおいても時系列情報を取り込んだモデルが有効であること」が示されている．そのため，7.2.1 についても，時系列の情報を適切に考慮させることによって異常検知性能の向上が期待できる．ただし，密に時系列情報と周波数情報を結合させた推論を行うと，第 5 章と同様に予測アルゴリズムが過度に複雑になり，予測結果の解釈性が失われかねないため，周波数情報を主とした異常検知のまま，スコアリング方法を変更することによって時系列情報を付与することとした．

7.2.1 の提案手法 1 について，相関係数と音声 IDF によるスコアリングの変更を加えた異常検知モデル（提案手法 2）の構造を Fig. 29 に示す．提案手法 2 では，提案手法 1 に加えて，音が持つ時系列情報のうち，安定性と希少性を異常スコアに取り入れている．Isolation Forest を用いて各時間窓に対して異常検知を行う場合，前後の時間窓との依存関係を考慮できないため，十分な性能を得られない可能性がある．たとえば，音声信号の異常データに

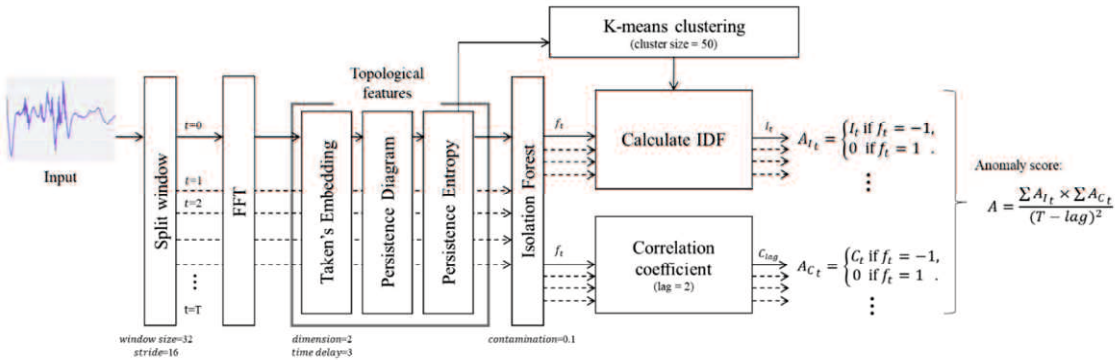


Fig. 29. Enhanced architecture of anomaly detection using Isolation Forest with topological and time-series features

は，一般的に， 1) 不安定性（異常な信号が発生したタイミングで過去との周波数特性が異なる），2) 希少性（異常音の発生は稀で，正常データには見られない特徴を持つ）のような特徴がある．本節では，提案手法 1 について，上記の観点から異常スコアの算出方法を改良することで，解釈性を残したまま異常検知性能の向上を目指す．提案手法 2 では， これらの特性を考慮できるようにするために，正常データの特徴量から，相関係数を用いて 1)を，本章で独自に設計した音声 IDF（Sound IDF）を用いて 2)をそれぞれ算出した．

7.2.2.1 相関係数

相関係数では，時系列情報が持つ不安定性について評価する．時刻 t の音が異常検知における非定常的なパターンであるかどうかを判断するために，時刻$t-lag$（lagは任意の自然数をとる）の音との間で相関係数C_tを計算し，異常スコアに組み込む．時刻 tにおける異常検知向けの相関係数C_tは次式で表される．

$$C_t = \begin{cases} -1 \times \left(\dfrac{v_{t-lag} \cdot v_t}{\|v_{t-lag}\| \|v_t\|} - 1 \right) & if\ t \geq lag \\ 0 & if\ t < lag \end{cases}, \tag{7.3}$$

ここで，v_tは時刻 tの特徴量を表す．ウィンドウサイズが 32 であることとストライドが 16 であること（7.2.1 を参照）を考慮し，本論文では，時刻tについて 2 時刻前の特徴量を用いて相関係数C_tを計算することとした（$lag = 2$）．Isolation forest の学習では時系列による依存関係が存在しないため，予測時の異常スコアの算出にのみ相関係数を使用する．

7.2.2.2 音声 IDF（Sound IDF）

IDF（Inverse Document Frequency）は，主に自然言語処理の分野で使用される，文書中の単語の希少性を定量化するための手法である[51]．本論文では，音声信号に対して，自然言語と同様に IDF を算出する仕組みを検討した．

信号処理で IDF を算出するためには，音声信号を単語と同様に扱えるようにする必要が

ある．そこで，音声信号からトポロジー特徴を計算し，クラスタリングを適用することによって音声 IDF を算出する．本節では，クラスタリング手法として k-means クラスタリング[52]を用いた．クラスタリングによって，時間窓単位に分割された正常データはクラスタ数の数だけグループ分けが行われる．このとき，予測時には異常な肺聴診音が稀少なクラスタに所属するような挙動を想定し，クラスタ数は 50 に設定している．クラスタkの IDF を意味するIDF_kを次式に示す．

$$IDF_k = \log\left(\frac{N+1}{n_k+1}\right) + 1 . \tag{7.4}$$

式（7.3）において，n_kはクラスタkに所属する時間窓を持っている訓練データ数を，Nが訓練データの総数を表す．したがって，時刻tの時間窓に割り当てられたクラスタ番号をk_tと定義すると，時刻tにおける IDF を意味するI_tは次式で表せる．

$$I_t = IDF_{k_t} . \tag{7.5}$$

ここで，正常なデータであるが音声窓が属することがほとんどないクラスタについては IDF が高く算出され，異常とみなされやすくなる．

7.2.2.3 相関係数と音声 IDF による異常スコアの算出

相関係数（7.2.2.1）と音声 IDF（7.2.2.2）を用いて，提案手法 1 で述べた異常スコアの算出を改良する．時刻tにおける Isolation forest の予測値をf_t （正常とみなしたデータは 1， 異常とみなしたデータは-1が出力される)としたとき，時刻tにおける相関係数A_{I_t}と音声 IDFA_{C_t}の異常スコアはそれぞれ次式で表すことができる．

$$A_{I_t} = \begin{cases} I_t \; if \; f_t = -1 \\ 0 \;\; if \; f_t = 1 \end{cases}, \tag{7.6}$$

$$A_{C_t} = \begin{cases} C_t \; if \; f_t = -1 \\ 0 \;\; if \; f_t = 1 \end{cases}. \tag{7.7}$$

さらに，A_{I_t}とA_{C_t}を用いて，肺聴診音データの最終的な異常スコアを計算する．予測対象の音声信号が持つ時間窓の総数をTとすると，異常スコアAは次式で表すことができる．

$$A = \frac{\sum A_{I_t} \times \sum A_{C_t}}{(T - lag)^2} . \tag{7.8}$$

上記のプロセスは，2 段階の異常検知，すなわち，第 1 段階の Isolation forest による各時間窓における独立した異常検知と，相関係数と音声 IDF を用いて第 1 段階の結果をブースティングさせた異常検知手法とみなすことができる．異常音に含まれる重要な特性に絞った情報を異常スコアに組み込むことで提案手法 1 の解釈性は失われず，時系列情報の依存関係によるブースティングによって，異常検知の精度向上が期待できる．

7.3 結果

従来手法と提案手法で得られた AUC の結果を Table 8 に示す．Table 8 より，Efficient GAN の AUC は全ての手法の中で優れていたものの，提案手法 2 が同等であったことが確認された．両者の P 値は 2.43×10^{-1}であり，有意差は見られなかった．また，提案手法 2

Table 8. Mean AUC and standard deviation of anomaly detection models in Chapter 7 by 14-fold cross validation

Folds	Methods				
	DAGMM	Efficient GAN	Proposed method 1	Proposed method 2	Proposed method 1 Without topological features
1	0.6475	0.8241	0.8110	**0.8589**	0.8292
2	0.7816	**0.8899**	0.8627	0.8867	0.8258
3	0.7595	**0.9215**	0.8097	0.8690	0.7671
4	0.6905	0.8146	0.8101	**0.8475**	0.5880
5	0.7361	0.8025	0.7722	**0.8032**	0.5278
6	0.7272	0.8285	0.8375	**0.8443**	0.6617
7	0.6772	**0.7987**	0.7013	0.7525	0.8022
8	0.6873	**0.9525**	0.9194	0.9146	0.8143
9	0.8418	**0.9424**	0.9364	0.9329	0.7445
10	0.6829	0.8481	0.9221	**0.9354**	0.7308
11	0.7203	**0.9247**	0.8554	0.8722	0.8081
12	0.6563	**0.9101**	0.9063	0.8987	0.5180
13	0.6266	**0.9032**	0.7810	0.8006	0.7896
14	0.7354	0.8418	**0.8620**	0.8589	0.7627
Means	0.7122	**0.8716**	0.8419	0.8625	0.7264

の AUC が DAGMM よりも高かったことからも，提案手法 2 は既存の深層学習異常検知モデルに匹敵する異常検知性能を持っていることがわかる．

一方，提案手法 1 は DAGMM と比較して高い AUC を示したものの，Efficient GAN より AUC が有意に低い結果(P 値は 3.10×10^{-2})となった．提案手法 1 は，ほとんどの fold で提案手法 2 を下回っており，提案手法 1 と 2 の P 値は 2.10×10^{-2}と有意差が認められたことからも，提案手法 2 の有効性が示されたと言える．これにより，時間窓の依存関係（時系列情報）を考慮することが，肺聴診音データにおいて有効であることが改めて示された．

また，7.2.1 で述べたトポロジー特徴の有効性についても検証する．Table 8 より，トポロジー特徴の計算を行わず，フーリエ変換後の特徴量で学習を行った場合の提案手法 1 の平均 AUC は 0.7264 であり，提案手法 1 の平均 AUC よりはるかに低い結果となった．提案手法 1 にトポロジー特徴を導入することによって，過学習によって精度が大幅に低下する fold の数が減っており，独立した時間窓での異常検知におけるトポロジー特徴の有用性が示された．しかし，Isolation Forest 単体では時系列依存を捉える学習を行うことができないため，単純なフーリエ変換で得られる特徴では，時間窓における周波数スペクトラム包絡中の変化情報を安定して学習できていない．

最後に，本章の提案手法の解釈性について述べる．それぞれ Fig. 30 は提案手法 1，Fig. 31 は提案手法 2 によって得られた，各クラスラベルの異常検出の例である．Fig. 30 では，オレンジ色の領域が，Isolation forest によって異常であると予測された時間窓を示す．Fig. 31 では，提案手法 2 によって得られた異常スコアをカラーマップを用いて示しており，時刻tの色の値は相関係数A_{I_t}と音声 IDFA_{C_t}を乗算することによって算出されている．式 (7.1)（7.2.1 を参照）より，Fig. 30 の提案手法 1 では，各時間窓に対して 0 または 1 の値を取るため，Fig. 31 のようなヒートマップによる可視化は行うことができない．

Fig. 30 より，正常なデータでは突然のノイズが異常として検出されることもあるが，概ね正しい異常領域が検出されることがわかる．提案手法 2 は提案手法 1 をベースとしてい

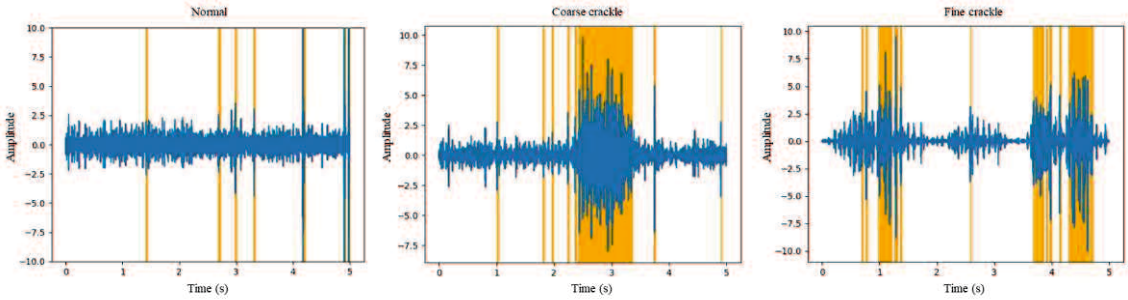


Fig. 30. Visualization of abnormalities with proposed method 1

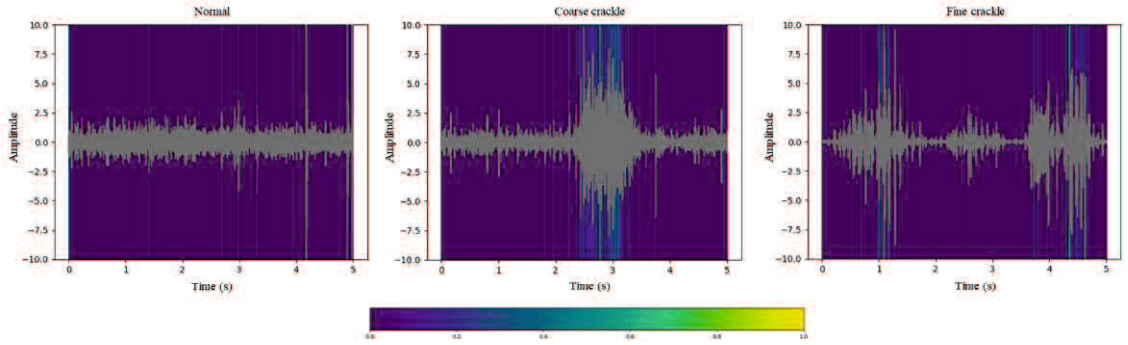


Fig. 31. Visualization of abnormalities with proposed method 2

るため，Fig. 31 においても同じ領域は検出されるが，重みづけを行っているため，（前時刻の時間窓との関係性が強い，もしくは正常データで頻発する）無視することが可能なノイズに対して異常スコアの強度を小さく，無視できないノイズに対しては異常スコアの強度が高くなる. 結果として, Fig. 30 で見られた正常データ内の環境音によるノイズは軽減され，異常データでは一部の時間窓について強度の高い異常スコアが残ることを確認できた．これにより，相関係数と音響 IDF を組み合わせた提案手法 2 によって，異常検知性能が向上しただけでなく，突発的なノイズの異常検出が抑制され，解釈性についても向上したと言える．

7.4 まとめ

本論文では，第 6 章で残った解釈性を考慮した異常検知モデルという課題に対し，適合する異常検知手法の検討と，異常検知手法に合わせた特徴抽出法の検討およびスコアリングの改良を行った．

実験の結果から，提案手法 1 におけるトポロジー特徴量の抽出は，一般的に用いるフーリエ変換などの特徴量と比較して，周波数情報の系列が時系列モデル以外にも学習しやすい形で表現されており，時間窓ごとの異常検知において有効であることが示された．さらに，提案手法 2 は，時系列情報のうち時間窓間の依存関係を異常スコアに組み込むことで，既存の深層学習ベースの異常検知モデルと同等の性能を持ち，かつ解釈性に優れることを示した．

本章で扱ったデータでは Fig. 30，Fig. 31 より明らかな解釈性の改善はみられるものの，3.1 のデータにおいて正しい異常領域が示されたデータが存在しないため，本章で述べた解釈性については著者らの評価に留まる．今後は，本章の内容を臨床応用可能な CAD 技術として確立するために実務者による評価が必要である．

第 8 章　おわりに

8.1 本論文のまとめ

本論文では，CAD 技術の中で比較的技術開発が遅れている肺聴診音に着目し，限られた数の肺聴診音データと機械学習・深層学習を用いた CAD 技術の研究を行った．研究全体で得られる共通の知見としては，「周波数帯域と時間の特徴量の組合せ」が少数の肺聴診音データの CAD 技術（分類モデル，異常検知モデル）の構築に有効であるという点である．本論文では肺聴診音の CAD 技術に適用される可能性のある上流タスク（分類・異常検知など）から検討しつつ，CAD 技術の発展に向けた示唆として「周波数帯域と時間の特徴量の組合せ」が有用であり，解釈性のある手法の展開可能性も示した．

ここで，本論文に示した第 4〜7 章の各章で得られた結果，および章間の比較から，次の結論を得た．

(1) 肺聴診音の CAD 技術はどのようなタスクが適切か

CAD 技術で深層学習のような複雑なモデルを扱う場合，AI モデルのタスク設定は実用性の点で重要である．本論文では，実際の医師の診断に準拠し，分類タスク（第 4 章，第 5 章）と異常検知タスク（第 6 章，第 7 章）の 2 つのタスクを扱った．実用的な CAD 技術において必要となる，「精度」と「解釈性」の 2 つの側面から検証を実施したところ，少数のデータにおいて異常検知のほうが高い精度が得られること（第 6 章）と，解釈性においても異常検知タスクのほうが優れたモデルを構築できること（第 5 章，第 7 章）が確認された．正常データは異常データと比較して収集が容易であることも加味すると，本論文で扱った診断環境に特化した肺聴診音の CAD 技術としては，異常検知タスクを対象としたモデル構築が向いていると言える．

ただし，タスクやモデルによって用途が限定され得ることに留意する必要がある．分類モデルは患者の予備診断や，スマートフォンなどのデジタル診療などが想定される．本論文で扱った分類モデルは精度や解釈性に課題は残っているが，ラベルを限定できる点や軽量さから事前診断の用途には向いている．一方，異常検知モデルは，患者のスクリーニングや診断補助などが想定される．第 6 章で扱った深層学習による異常検知は，本論文において高い AUC を達成しており，医師が診るべき患者に注力できるようスクリーニングなどの用途で使用することが可能である．第 7 章で扱った異常検知手法は，異常時刻を可視化できることから，肺音のより詳細な解析や患者への説明など，診断補助に活用することが可能である．

(2) 少数のデータにおいてもタスクの精度を高める方法

肺聴診音は，医師の実務において音声データの収集やラベリング（クラスラベルの付与）を行うハードルが高く，多量のデータを用意することは難しい．そのため，少数のデータでも精度を高める手法の検討が必要であった．本論文では，分類モデルにおいては事前学習による訓練（第 4 章）が，異常検知モデルにおいては GMM（第 6 章）の仕組みが有効であることが示された．これらの手法は，いずれも少ない特徴量でドメイン内のデータを表現することを重視しており，これによって音声時系列の過学習や，肺聴診音内のノイズの悪影響が抑制され，精度の向上に繋がっている．

(3) 肺聴診音の特徴量エンジニアリング

肺聴診音を機械学習や深層学習で扱うための特徴抽出法の検討も，データが少ない場合のモデル構築はもちろん，データが増えた場合においても，パラメータ数に対する学習効率の向上が望めるため価値がある．本論文においては，第 4 章と第 5 章において，複数の深層学習アーキテクチャの比較を行い，以下の知見を得た．

- 肺聴診音の分類タスク・異常検知タスクのどちらも，肺聴診音の「周波数情報（周波数帯域の情報）」と「時系列情報（周波数帯域の時系列変化情報）」が性能の向上に効果的

である．特に，周波数帯域の情報は，それ単体で学習を行った場合に，時系列情報よりも高い精度を示した（第 5 章）ことから，時系列情報と比較して深層学習アーキテクチャが識別し易い情報であると言える．また，第 5 章より，前述の周波数情報と時系列情報は，どちらも特徴量として取り込むことによってより高い性能となることが示された．

- ノイズの軽減を前処理，もしくは深層学習アーキテクチャ内で行うことが重要である．肺聴診音の識別において，周波数情報と時系列情報はいずれもノイズの影響を受けている．周波数情報では，MFCC 変換後の周波数情報や，さらに情報量を落としたトポロジー特徴量と比較して，フーリエ変換のみの周波数情報では精度が落ちている（第 7 章）ことから，周波数帯域の情報についてもノイズの軽減を前処理として行うことが必要である．時系列情報についても，性能の向上に寄与することを本論文中一貫して述べている（第 4〜7 章）が，試行によっては，時系列に対する過学習を要因とする極端な精度の低下が見られる．これらのノイズを含んだ状態で精度を上げるためには，(2)で述べた学習方法の他に，深層学習アーキテクチャによるノイズの軽減（例：畳み込み，プーリングなど）が有効であることが示された（第 4 章）．

一般的な音声認識では用いられない次元数の MFCC についても，識別性能向上に効果があったこと（第 4 章），MFCC の転置によって時系列ニューラルネットワークの性能に大幅な改善が見られたこと（第 5 章）などから，一般的ではない音声信号の加工方法や，医学的な根拠とは異なる前処理についても，肺聴診音の識別性能の向上に寄与する可能性があることを示した．音声認識で一般的に用いられる特徴量である MFCC は，離散コサイン変換によって微細な包絡情報が失われる（Fig. 1）が，これによって肺聴診音の識別に有用な情報が一部欠落している可能性がある．このような，一般的な音声認識の前処理で生じたデータの意図的な欠損は，通常モデルの学習を妨げるノイズを減らすように機能するが，肺聴診音向けの用途において負の影響が発生するケースが確認された．4.2.2（第 4 章の提案手法 2）では，MFCC に微細な包絡情報を残した場合の分類モデルの検証を行っ

たが，畳み込みを持つニューラルネットワークではMFCCの次元数が少ない場合，つまり情報の欠損が多いほど，識別性能が低下した．一方で，LSTMなど一部のニューラルネットワークではこの傾向が見られなかったことから，前処理による情報の欠損の影響はアーキテクチャにも依存することが確認された．本論文で得られた上記の(1)〜(3)の示唆は，今後の肺聴診音のCAD技術の精度の向上や実用化に寄与するものである．

8.2 今後の課題

本論文で得られた結果は，肺聴診音について固有の環境で有用性を示しているが，実際の診断現場においては，医師の技量によるノイズの増加や，聴診環境で生じる環境音，聴診機器の変更に伴う音声などの外的要因により，予期しない形で正常データの分布がシフトする可能性がある．これは，特に少ないデータで学習したAIモデルにおいてリスクが大きく，発生した場合にはAIモデルの挙動が予期しないものとなり得る．そのため，より汎用的なCAD技術の確立に向けては，周囲の雑音をはじめとする外的な環境要因に左右されない，さらに頑健なAIモデルを構築することが重要である．このような課題に対して，AIモデルの安定性を高めるためには，信号処理技術の 1 つである音源分離を応用することなどが考えられる．肺聴診音は，正常・異常な肺音に加えて，多種類のノイズを含み得る音声であるため，音源分離技術による異常肺音の抽出や，診断に無関係なノイズを分離するなどのアプローチにより，さらに汎化的なAIモデルが期待できる．もしくは，2.3.6 にて言及した大規模言語モデル，大規模音声認識モデル，マルチモーダル大規模言語モデルなど，会話や環境音などの多様な音声，または音声データに限らない大量のデータによって大規模な事前学習が行われたモデルを活用することで，意味のないノイズに対して頑健な特徴抽出が期待できる．

最後に，実用化に向けては目標性能の設定や，医師による定性評価を実施することも必要となる．本論文で述べた精度や解釈性の他にも，医師の実務において重要な評価は存在する可能性があり，今後調査を継続しそれらを多面的に評価したい．

参考文献

[1] Y. LeCun, Y. Bengio, and G. Hinton：“Deep learning”, Nature, Vol. 521, No. 7553, pp.436-444 (2015)

[2] D. Shen, G. Wu, and H.-I. Suk：“Deep Learning in Medical Image Analysis”, Annual Review of Biomedical Engineering, Vol. 19, pp. 221-248 (2017)

[3] K. Doi, H, MacMahon, S. Katsuragawa, R. M. Nishikawa and Y. Jiang：“Computer-aided diagnosis in radiology: potential and pitfalls”, European Journal of Radiology, Vol. 31, Issue 2, pp. 97–109 (1999)

[4] 石原恒夫・川城丈夫・阿部直・菊丸功次・米丸亮：CD による聴診トレーニング, 南江堂 (1993)

[5] 桑原道義：「総論―医用画像処理の歴史と展望」, BME, Vol. 3, No. 8 (1989)

[6] H. K. Huang and R. K. Taira：“Infrastructure design of a picture archiving and communication system”, American Journal of Roentgenology, Volume 158, Issue 4 (1992)

[7] 奥山文雄・橋本則男・河村徹郎：「電子カルテの動向と課題」, MEDICAL IMAGING TECHNOLOGY, Vol. 25, No.3, pp. 143-148 (2007)

[8] 井神佳明・庄野逸・木戸尚治：「隣接する波形間隔の統計を用いた肺聴診音データの識別」, 第 9 回 IEEE 広島支部学生シンポジウム(HISS2007), B53 (2007)

[9] D. Bardou, K. Zhang, and S. M. Ahmad：“Lung sounds classification using convolutional neural networks”, Artificial Intelligence in Medicine, Vol. 88, pp. 58-69 (2018)

[10] C. E. Shannon：“Communication in the Presence of Noise”, Proceedings of the IRE, Vol. 37, pp. 10-21 (1949)

[11] J. W. Cooley and J. W. Tukey：“An algorithm for the machine calculation of complex Fourier series”, Mathematics of Computation, Vol. 19, pp. 297-301 (1965)

[12] S. Davis and P. Mermelstein：“Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences”, in IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28, No. 4, pp. 357-366 (1980)

[13] 篠田浩一：音声認識，講談社 (2017)

[14] L. R. Rabiner, B. Gold and C. K. Yuen：“Theory and Application of Digital Signal Processing”, in IEEE Transactions on Systems, Man, and Cybernetics, Vol. 8, No. 2, pp. 146 (1986)

[15] G. E. Carlsson：“Topology and data”, Bulletin of the American Mathematical Society, Vol. 46, pp. 255-308 (2009)

[16] 大林一平：「位相的データ解析の現在」，数理解析研究所講究録，Vol. 2057, pp. 34-50 (2017)

[17] F. Takens ： “Detecting strange attractors inturbulance”, LectureNotes in Mathematics, Vol. 898, pp. 366-381 (1981)

[18] M. B. Kennel, R. Brown, and H. D. I. Abarbanel：“Determining embedding dimension for phase-space reconstruction using a geometrical construction”, Phys. Rev. A 45, pp. 3403–3411 (1992)

[19] H. Edelsbrunner, D. Letscher and A. Zomorodian：“Topological Persistence and Simplification”, Discrete & Computational Geometry, Vol. 28, pp. 511-533 (2000)

[20] M. Rucco, F. Castiglione, E. Merelli, and M. Pettini：“Characterisation of the idiotypic immune network through persistent entropy, ” in Proceedings of ECCS 2014, pp. 117–128 (2014)

[21] 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, Vol. 27, No. 6, pp. 588-595 (1979)

[22] 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, Vol. 2, pp. 845-848 (1990)

[23] W. S. McCulloch and W. Pitts：“A Logical Calculus of the Ideas Immanent in Nervous Activity”, Bulletin of Mathematical Biology, Vol. 52, pp. 99-115 (1990)

[24] V. Nair and G. E. Hinton：“Rectified Linear Units Improve Restricted Boltzmann Machines”, International Conference on Machine Learning (2010)

[25] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel：“Backpropagation Applied to Handwritten Zip Code Recognition”, in Neural Computation, Vol. 1, No. 4, pp. 541-551 (1989)

[26] J. S. Bridle：“Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition”, NATO ASI Series, Vol. 68 (1990)

[27] D.E. Rumelhart, G.E. Hinton and R.J. Williams：“Learning Representations by Back-Propagating Errors” Nature, Vol. 323, pp. 533-536 (1986)

[28] 岡谷貴之：深層学習改訂第２版，講談社 (2022)

[29] Y. LeCun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard and L. D. Jackel：“Handwritten digit recognition with a back-propagation network”, In Advances in Neural Information Processing Systems, pp. 396–404 (1990)

[30] M. Lin, Q. Chen, S. Yan：“Network In Network”, Proceedings of the International Conference on Learning Representations 2014 (2014)

[31] S. Hochreiter and J. Schmidhuber：“Long short-term memory”, Neural computation, Vol. 9, No. 8, pp. 1735–1780 (1997)

[32] A. Graves and J. Schmidhuber：“Framewise phoneme classification with bidirectional LSTM networks”, Proceedings. 2005 IEEE International Joint Conference on Neural Networks, Vol. 4, pp. 2047-2052 (2005)

[33] S. Xingjian, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo：“Convolutional LSTM network: A machine learning approach for precipitation nowcasting”, In Neural Information Processing Systems (2015)

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez and I. Polosukhin：“Attention is all you need”, 31st Conference on Neural Information Processing Systems, Vol. 30 (2017)

[35] K. He, X. Zhang, S. Ren and J. Sun : “Deep Residual Learning for Image Recognition”, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778 (2016)

[36] A. Baevski, H. Zhou, A. Mohamed, and M. Auli : “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”, In Advances in Neural Information Processing Systems, Vol. 33, pp. 12449–12460 (2020)

[37] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen : “Deep autoencoding gaussian mixture model for unsupervised anomaly detection”, International Conference on Learning Representations (2018)

[38] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar : “Efficient gan-based anomaly detection”, International Conference on Learning Representations (2018)

[39] J. Donahue, P. Krahenbuhl, and T. Darrell : “Adversarial feature learning”, Proceedings of the 5th International Conference on Learning Representations (2016)

[40] F. T. Liu, K. M. Ting and Z. H. Zhou : “Isolation Forest”, 2008 Eighth IEEE International Conference on Data Mining, pp. 413-422 (2008)

[41] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell : “Caffe: Convolutional Architecture for Fast Feature Embedding”, Proceedings of the 22nd ACM international conference on Multimedia, pp.675-678 (2014)

[42] N. Srivastava, E. Mansimov, and R. Salakhutdinov : “Unsupervised learning of video representations using LSTMs”, Proceedings of the 32nd International Conference on Machine Learning, Vol. 37, pp. 843-852 (2015)

[43] B.A. Hanson, and T.H. Applebaum : “Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with Lombard and noisy speech”, Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing , Vol. 2, pp.857-860 (1990)

[44] J. Devlin, M. Chang, K. Lee, and K. Toutanova : “BERT: Pre-training of deep bidirectional transformers for language understanding”, In North American

Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171

[45] S. M. Lundberg and S. I. Lee : “A unified approach to interpreting model predictions”, in Proceedings of the Advances in Neural Information Processing Systems, pp. 4765–4774 (2017)

[46] M. T. Ribeiro, S. Singh, and C. Guestrin : ““Why should I trust you? ”: Explaining the predictions of any classifier” , In Proceedings of the ACM Conf. on Knowledge Discovery and Data Mining (KDD) (2016)

[47] L. McInnes, J. Healy, and J. Melville : “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”, arXiv preprint arXiv:1802.03426 (2018)

[48] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner : “Gradient-based learning applied to document recognition”, Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278-2324 (1998)

[49] A. Radford, L. Metz, and S. Chintala : “Unsupervised representation learning with deep convolutional generative adversarial networks”, In Proceedings of the 4th International Conference on Learning Representations (ICLR) (2016)

[50] T. Dissanayake, T. Fernando, S. Denman and S. Sridharan : “Understanding the Importance of Heart Sound Segmentation for Heart Anomaly Detection”, unpublished (2020)

[51] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner : “A Statistical Interpretation of Term Specificity and Its Application in Retrieval”, Journal of Documentation, Vol. 28, pp. 11–21 (1972)

[52] J. MacQueen : “A Statistical Interpretation of Term Specificity and Its Application in Retrieval”, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297 (1967)

謝辞

本論文の研究を遂行するにあたり，指導教員として多大なるご指導とご助言を賜りました山口大学大学院創成科学研究科の間普真吾教授に，心より感謝申し上げます．先生のご指導のもと，データサイエンティストとしての礎を築く貴重な学びを得ることができました．また，副査を務めてくださった山口大学大学院創成科学研究科の中村秀明教授，田村慶信教授，藤田悠介准教授，佐村俊和准教授にも，学位審査の過程で有益かつ適切なご助言を賜りました．この場を借りて深く感謝申し上げます．

研究の初期から多くのご指導を賜りました，大阪大学大学院の木戸尚治招へい教授，日本工業大学の呉本尭教授にも心より御礼申し上げます．

生体情報システム工学研究室の諸先輩方，そして共に励んできた同僚の皆様にも，温かいご支援を賜りました．多くの議論や助言を通じて，新たな視点を得ることができたことに，深く感謝いたします．

最後に，これまで長年にわたり私を支えてくれた家族，そして社会人博士の取り組みを理解し励ましてくれた妻に，心より感謝いたします．

本研究を支えてくださったすべての方々に，深甚なる感謝の意を表します．