

# Bolasso を用いたびまん性肺疾患画像の特徴選択

遠藤 瑛泰<sup>1,a)</sup> 永田 賢二<sup>2,3</sup> 木戸 尚治<sup>4</sup> 庄野 逸<sup>1,b)</sup>

受付日 2019年5月8日, 再受付日 2019年7月10日,  
採録日 2019年8月30日

**概要:** びまん性肺疾患は肺全体に広がる病態を持ち, 早期の発見と適切な治療が求められている. 本研究では, びまん性肺疾患の X 線 CT 画像に対して複数の特徴量を算出し, これらの特徴量群から, 病態の特異性を示す特徴を選択しつつ, 診断支援を行う学習機械の構築を試みた. このような学習機械の構築は, 病態の特異性を持つ特徴量をデータに基づいて選択しているため, どのような特徴の組合せに病理診断のための情報がのっているのかを発見することが期待できる. 本研究では, 特徴選択において Bolasso と呼ばれる手法を適用している. Bolasso はスパース推定の枠組みで特徴選択を行う LASSO とブートストラップ法を組み合わせた手法である. 通常 LASSO を用いて特徴選択を行うと, データに含まれるノイズなどの影響により特徴を過剰に選択する傾向にある. Bolasso では, この欠点を克服するために, ブートストラップ法による再標本データと LASSO の適用を繰り返し実行し, 得られた組合せ集合から有効な特徴を推定する, この Bolasso を, びまん性肺疾患の画像セットに対して適用することにより, 異常陰影の解釈に適した特徴を絞り込めることを確認した.

**キーワード:** 特徴選択, スパース推定, Bolasso, びまん性肺疾患, テクスチャ特徴

## Feature Selection for Diffuse Lung Disease Image by Bolasso

AKIHIRO ENDO<sup>1,a)</sup> KENJI NAGATA<sup>2,3</sup> SHOJI KIDO<sup>4</sup> HAYARU SHOUNO<sup>1,b)</sup>

Received: May 8, 2019, Revised: July 10, 2019,  
Accepted: August 30, 2019

**Abstract:** Diffuse Lung Disease (DLD) are observed wide spreading in lung, and preventing spread of the disease requires early detection and proper medical treatment. In this research, we try to construct a computer aided diagnosis (CAD) system which selects features of DLD in X-ray CT image for classification. Combination of selected features, which are determined by data, would make clues of finding information carrier diagnosing task. In order to select features, we apply a method called “Bolasso” for this purpose. The Bolasso is a combination of bootstrapping method and a sparse modeling method called LASSO. LASSO requires sparse solution for representation, however, it tends to overestimate the features in selection process. Overcoming this problem, we introduce Bolasso, which integrates many LASSO solutions for resampled data by boot strapping. We confirm the Bolasso solution shows reasonable and stable for the DLD classification task.

**Keywords:** feature selection, sparse estimation, Bolasso, diffuse lung disease, texture feature

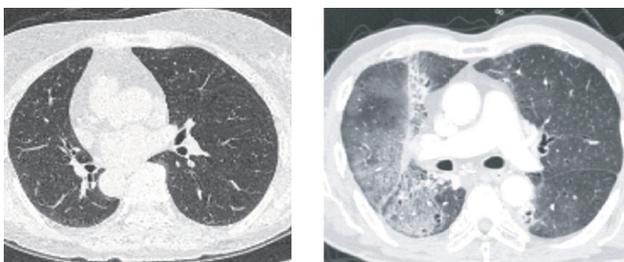
<sup>1</sup> 電気通信大学大学院情報理工学研究所  
Graduate School of Informatics and Engineering, The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan  
<sup>2</sup> 物質・材料研究機構統合型材料開発・情報基盤部門材料データプラットフォームセンター  
Materials Data Platform Center, Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba, Ibaraki 305–0047, Japan

<sup>3</sup> 科学技術振興機構さきがけ  
Japan Science and Technology Agency, PRESTO, Kawaguchi, Saitama 332–0012, Japan  
<sup>4</sup> 大阪大学大学院医学系研究科  
Graduate School of Medicine, Osaka University, Suita, Osaka 565–0871, Japan  
a) e1830017@edu.cc.uec.ac.jp  
b) shouno@uec.ac.jp

## 1. はじめに

びまん性肺疾患は、肺の広範囲に病変が分布している病態を持ち、日本国内において難病指定されている病気である。びまん性肺疾患は根治の方法が確立しておらず、病態の拡大することを抑制するために、早期の発見と症状に適した加療が求められている [12]。図 1 に肺で見られる X 線 CT 画像の陰影例を示す。図中、左側の図は正常な肺画像を示し、右側の図の右肺の淡く白い領域がびまん性肺疾患の患部を示している。正常な肺は、空気で満たされているため、黒みがかった領域で覆われているのに対し、びまん性肺疾患患者の肺では肺泡領域の線維化による蜂巢状影、肺泡内の様々な充填物による浸潤性陰影やすりガラス状影、網状影などの異常陰影が観測される [12]。これらの異常陰影は病変の性状を示しており、びまん性肺疾患の疾患の特定や進行の確認といった診断の手がかりとなる。そのため、肺 X 線 CT 画像を用いた画像所見はびまん性肺疾患の診断に有効である。このような肺 CT 画像に対して計算機診断支援システムのような計算機による画像解析を行うことは、びまん性肺疾患で現れるような異常陰影を効率的かつ早期の発見のために重要な役割を持つ。

びまん性肺疾患を対象とした先行研究では様々な画像診断支援手法が提案されている [4], [9]。近年の画像認識の分野において大きな成果を示している深層畳み込みニューラルネットワーク (Deep Convolutional Neural Network; DCNN) による判別では、それほど大規模な数のデータ収集が見込めない医用画像においても転移学習手法を導入すれば、十分な精度での陰影分類が可能であることが報告されている [7]。しかし、その一方で、DCNN の分類における判断根拠は、いまだ不明瞭であり、DCNN 内部においてどのような特徴表現に着目しているのかといった解釈は十分な結論がでていないわけではない。予測結果やモデルに対する解釈可能性は、予測における信頼性を担保する役割を担っており、医用分野などにおいては非常に重要である。



正常な肺CT画像

びまん性肺疾患CT画像

図 1 X 線 CT 画像における肺画像例。左図が正常な肺画像を示し、右図中の右肺がびまん性肺疾患陰影を示している

Fig. 1 Example of lung image in X-ray CT image. The left figure shows normal lung image, The right lung in the right figure shows a diffuse lung disease shadow.

このような解釈をより容易にする手法として、与えられた特徴から必要な特徴を特定する特徴選択が考えられる。これまでびまん性肺疾患の陰影分類に対して、組合せ探索やスパース推定による特徴選択が適用されているが [5], [10], 選択の妥当性に関しては分類性能に依存しているため、十分に解釈に適した特徴を絞り込めていないことが懸念される。

本論文では、Bach によって考案された特徴選択手法である Bolasso (Bootstrapped Lasso) を適用し [1], びまん性肺疾患の陰影分類に関する特徴解析を行った。Bolasso はスパース推定の枠組みで特徴選択をする LASSO (Least Absolute Shrinkage and Selection Operator) とブートストラップ法を組み合わせた手法である [2], [8]。LASSO は、後述するように特徴選択において  $L_1$  正則化をあたえることで必要変数の刈り込みを行う手法である。この LASSO による特徴選択は、データセットに対してとても敏感であり、観測データに含まれるノイズによって選択される特徴の組み合わせが異なる場合が多い [3], [11]。このような LASSO の欠点を克服するために、Bolasso では、ブートストラップ法により、複数回の LASSO を並列的に実行し、得られた選択特徴のセットを統合することで、データセットに対してより頑健で安定した特徴を推定する手法である。この Bolasso を識別手法に対して適用することで、識別結果に影響を与える特徴 1 つ 1 つに対して重要性を調べることが可能になり、これまで困難であった識別判断の解釈に適した特徴セットを特定できることが期待できる。

本論文の構成は、2 章において LASSO と Bolasso の説明を行う。3 章では、識別課題における特徴選択問題に対して、人工データを用いて Bolasso の有効性を示す。続く 4 章では、びまん性肺疾患から得られる画像特徴量を用いて Bolasso による特徴選択を行い、5 章において議論を行う。

## 2. 手法

本論文では、2 値分類を行う線形識別器としてロジスティック回帰モデルを考える。

ここでは線形識別器に与える入力データを  $P$  次元のベクトル  $\mathbf{x} = (x_1, x_2, \dots, x_P)$  とし、1 か 0 で表されたラベル  $t$  の予測を行う。ただし  $P$  は  $\mathbf{x}$  の特徴量の数を表す。また、 $N$  個の入力データを  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} = \{\mathbf{x}_n\}_{n=1}^N$ 、対応するラベルを  $\mathbf{t} = \{t_1, t_2, \dots, t_N\} = \{t_n\}_{n=1}^N$  と表すことにする。入力データの各次元に対応する重みパラメータを  $\mathbf{w} = (w_1, w_2, \dots, w_P)$  としたとき、ある入力データ  $\mathbf{x}_n$  が  $t = 1$  である確率を  $y_n = \sigma(\mathbf{w}^\top \mathbf{x}_n + w_0)$  とする。 $w_0$  は定数項であり、 $\sigma(z)$  は  $\sigma(z) = 1/(1 + \exp(-z))$  で表されるロジスティックシグモイド関数である。一方、ある入力データ  $\mathbf{x}_n$  が  $t = 0$  である確率は  $y_n$  を用いて、 $1 - y_n$  と求めることができる。 $\mathbf{w}$  は、入力データ  $\mathbf{X}$  から  $t$  の予測誤差を最小にするように、以下の目的関数の最小化を行い求める：

$$E(\mathbf{w}) = - \sum_{n=1}^N \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\}. \quad (1)$$

構築した識別器を用いたある入力データ  $\mathbf{x}_n$  に対する予測ラベル  $\hat{t}_n$  は,  $y_n > 1 - y_n$  ならば 1,  $y_n \leq 1 - y_n$  ならば 0 とする.

### 2.1 LASSO による特徴選択

特徴選択の問題はある特徴を使うか使わないかを定める, 組合せ最適化の問題である. ある特徴を使うか使わないかを 1 か 0 かの 2 値で表現した  $P$  次元のインジケータベクトル  $\mathbf{s}$  で表現する:

$$\mathbf{s} = (s_1, s_2, \dots, s_P), \quad (2)$$

$$s_i \in \{0, 1\} \quad (i = 1, \dots, P). \quad (3)$$

ある特徴  $x_i$  を使うときは  $s_i = 1$  であり, 使わないときは  $s_i = 0$  となる. 最適な組合せを求めるための最も確実な方法はすべての組合せを探索することであるが, 特徴の数  $P$  に対して組合せの数は  $2^P$  で考えられるため, 一般には計算困難な問題となる.

LASSO [8] は, 目的関数  $E(\mathbf{w})$  に  $L_1$  正則化項を追加することで効率的に有効な組合せを見つけるスパース推定手法である:

$$E(\mathbf{w}) = - \sum_{n=1}^N \{t_n \log y_n + (1 - t_n) \log(1 - y_n)\} + \lambda \sum_{p=1}^P |w_p|. \quad (4)$$

ただし,  $\lambda$  は, 正則化項の効果の強さを制御するハイパーパラメータである. この正則化項を加えることにより,  $\mathbf{w}$  の一部の要素  $w_i$  が 0 となりやすくなるため, 疎な解が選択されやすくなる. ある重み  $w_p$  が 0 であるとき,  $i$  番目の特徴量は  $\mathbf{y} = \{y_1, y_2, \dots, y_N\} = \{y_n\}_{n=1}^N$  の結果に寄与しないため,  $w_p = 0$  となる特徴を選択しない結果と考えることができる. そこで  $\mathbf{w}$  から推定するインジケータベクトル  $\hat{\mathbf{s}}$  を以下のように定義する:

$$\hat{\mathbf{s}} = (\hat{s}_1, \hat{s}_2, \dots, \hat{s}_P), \quad (5)$$

$$\hat{s}_p = \text{sign}(|w_p|), \quad (6)$$

$$= \begin{cases} 1, & |w_p| > 0 \\ 0, & |w_p| = 0 \end{cases}. \quad (7)$$

$\text{sign}(u)$  は符号関数であり,  $u > 0$  ならば  $\text{sign}(u) = 1$ ,  $u = 0$  ならば  $\text{sign}(u) = 0$ ,  $u < 0$  ならば  $\text{sign}(u) = -1$  を返す関数である.  $\hat{\mathbf{s}}$  は  $w_p \neq 0$  となる特徴を使い,  $w_p = 0$  となる特徴を使わないことを表したインジケータベクトルとなる.

### 2.2 Bolasso による特徴選択

LASSO による特徴選択は様々な分野において利用され

---

**Algorithm 1** Feature selection algorithm for classifier using Bolasso

---

**Input:** Input data:  $\mathbf{X}$ , Label:  $\mathbf{t}$ , Number of replicas:  $M$ , Hyper parameter:  $\lambda$

**Output:** Indicator vector:  $\hat{\mathbf{s}}$

Replica:  $(\mathbf{X}^m, \mathbf{t}^m) \leftarrow \text{bootstrap sampling}(\mathbf{X}, \mathbf{t}) \quad (m = 1, \dots, M)$

**for**  $m = 1$  to  $M$  **do**

Optimize  $\mathbf{w}^m$  to fit  $(\mathbf{X}^m, \mathbf{t}^m)$  by LASSO with  $\lambda$

Estimate  $\hat{\mathbf{s}}^m$  from  $\mathbf{w}^m$

**end for**

Indicator vector set:  $\mathbf{S} = \{\hat{\mathbf{s}}^1, \dots, \hat{\mathbf{s}}^M\}$

**for**  $p = 1$  to  $P$  **do**

Compute  $\hat{s}_p = \bigcap_{m=1}^M \hat{s}_p^m$

**end for**

Estimate  $\hat{\mathbf{s}} = (\hat{s}_1, \dots, \hat{s}_P)$

---

た手法であるが, LASSO は, 正則化項の強さや与えるデータのゆらぎに敏感で, これらを少し変化させると  $\hat{\mathbf{s}}$  が変化してしまい, 最適な組合せとは異なる  $\hat{\mathbf{s}}$  が得られる場合がある. この問題を解決するために, 適した特徴のみを推定するための枠組みとして, Bach によって Bolasso が考案されている [1]. Bolasso とは, LASSO による特徴選択から頑健な組合せを求めるために, ブートストラップ法 [2] を取り入れた特徴選択手法である.

Bolasso は, 複数回の LASSO によって得られるインジケータベクトルの集合  $\mathbf{S}$  から  $\hat{\mathbf{s}}$  を求める. Bolasso の手順を Algorithm 1 に示す. はじめに, レプリカと呼ばれる複製データセットを入力データ  $\mathbf{X}$  と対応するラベル  $\mathbf{t}$  から  $M$  個準備する. レプリカはブートストラップ法に基づき, 重複を許した再標本によって作成される.

作成したレプリカに対する LASSO による特徴選択を考える.  $m$  番目のレプリカが持つ入力データを  $\mathbf{X}^m$ , 対応するラベルを  $\mathbf{t}^m$  とする. このとき,  $m$  番目のレプリカで得られるインジケータベクトル  $\mathbf{s}^m$  を,  $\mathbf{X}^m$  と  $\mathbf{t}^m$  から LASSO によって求まる  $\mathbf{w}^m$  から決める. 同上の手続きを作成したすべてのレプリカに対して行い,  $\mathbf{S} = \{\hat{\mathbf{s}}^1, \dots, \hat{\mathbf{s}}^M\}$  を求める. すべてのレプリカの試行において,  $\lambda$  は事前に設定した同一の値を用いる. 得られた  $\mathbf{S}$  から, Bolasso で推定される  $\hat{\mathbf{s}}$  を求める:

$$\hat{\mathbf{s}} = (\hat{s}_1, \dots, \hat{s}_P), \quad (8)$$

$$\hat{s}_p = \bigcap_{m=1}^M \hat{s}_p^m. \quad (9)$$

$\hat{s}_p$  は各要素に対して論理積を求めた結果を示しており, すべてのレプリカにおいて必ず選択された特徴は 1 をとり, 1 度でも選択されなかった特徴は 0 となる.

Bolasso を適用する際, レプリカ数  $M$  を十分に大きな値で設定することにより, 適切に不要な特徴を取り除くことができる. そのため, すべてのレプリカにおいて選択される特徴は, データに対して選択性が頑健であり, 解釈にお

いて適した特徴であることが期待できる。

### 3. 人工データによる実験

本章では実データを想定して生成した人工データに対して、LASSO および Bolasso を適用した特徴選択実験を行う。Bach が提案した Bolasso は、線形回帰モデルに対しての提案であるが、本研究では識別モデルを対象としているため、Bolasso が十分に機能するかに関しては検証する必要がある。そこで人工データを生成して、Bolasso が識別問題に対して機能するかを確認する。

#### 3.1 人工データの生成

人工データを  $\mathbf{X}$  として、 $P$  次元の空間上の各次元において 0 から 1 の範囲で一様な乱数を特徴とするデータを  $N = 100$  個生成した：

$$x_p = \text{Uniform}(0, 1) \quad (1 \leq p \leq P). \quad (10)$$

生成したデータ  $\mathbf{X}$  に対応するラベル  $t$  は  $P$  次元の特徴空間上において、設定した  $\mathbf{w}$  に従う分離境界をもとに決定した。本実験では、 $w_1 = 1, w_4 = -1$ 、とし、それ以外の特徴の重みが 0 となるように分離境界を設定し、ラベル  $t$  を  $\mathbf{y} = \sigma(\mathbf{X}\mathbf{w})$  から決定した。

この人工データは  $x_1, x_4$  のみがラベルに関する情報を含んでいるため、推定したい真のインジケータベクトル  $\mathbf{s}$  は以下のような組合せとなる：

$$s_p = \begin{cases} 1, & (p = 1, 4) \\ 0, & (\text{otherwise}) \end{cases}. \quad (11)$$

この人工データに前処理として各特徴に対して平均 0、分散 1 となるような線形変換を適用し、特徴選択実験を行った。

#### 3.2 人工データに対する特徴選択

次元数  $P$  が 20 の人工データとさらに不要な特徴を 20 次元追加した次元数  $P$  が 40 の人工データを用いて実験を行った。以降の実験では、 $L_1$  正則化の強さを制御するハ

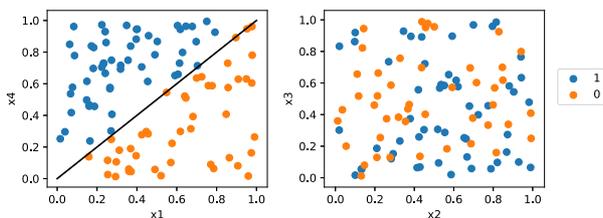


図 2 生成した人工データの分布。左図は、第 1 要素  $x_1$  と第 4 要素  $x_4$  にとった散布図で、右図が第 2 要素  $x_2$  と第 3 要素  $x_3$  にとった散布図

Fig. 2 Distribution of generated synthetic data. The left figure shows scatter plot with  $x_1$  and  $x_4$ . The right figure shows scatter plot with  $x_2$  and  $x_3$ .

イパーパラメータは、 $\lambda$  の逆数  $C = 1/\lambda$  を用いて記述することにする。すなわち  $C$  が、大きいほど正則化項の効果が弱くなり、小さいほど正則化項の効果が強くなる。LASSO を適用するに際して、 $C$  は 10 分割交差検証法で最も交差検証誤差の小さい値を採用した。Bolasso 内で用いる LASSO のハイパーパラメータ  $C$  は、同上の手法で決定した。表 1 に用いた  $C$  の値とそのときの交差検証誤差を示す。また Bolasso では、レプリカ数  $M$  を 1,000 として推定を行った。

選択された特徴の結果を図 3 と図 4 に示す。選択された特徴の次元を黒、選択されなかった特徴の次元を白で表している。LASSO を用いた場合、 $P = 20$  のときは推定している結果が設定した特徴のみを選択することができている。しかし、 $P = 40$  のときでは設定した特徴を選択しているが、推定結果には余計な特徴が選択されてしまっている。人工データによる実験から、不要な特徴が含まれている場合に、これらがデータのとり方によって識別に有効に働いてしまうことが考えられる。一方 Bolasso を適用した場合は、 $P$  が 20, 40 のどちらの場合においても  $s_1$  と  $s_4$  のみを選択することができている。

図 5 に Bolasso における各特徴ごとの選択頻度を示す。横軸に次元、縦軸にそれぞれの特徴が選択された頻度を表している。赤で示している特徴は Bolasso で選択される特徴であり、設定した特徴と一致している。 $P = 40$  のときに LASSO で選択されていた特徴は、レプリカによって

表 1  $C$  と交差検証誤差  
Table 1  $C$  and cross validation error.

$P$	$C$	交差検証誤差
20	0.091	0.037
40	213.8	0.029

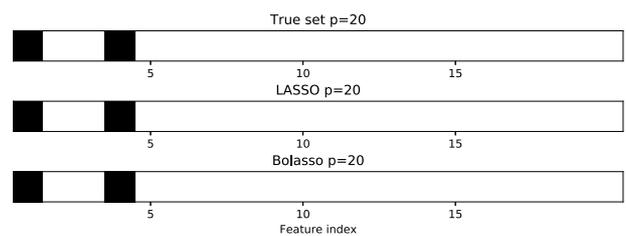


図 3 人工データを用いた  $P = 20$  のときの結果。選択された特徴が黒く示されている

Fig. 3 The result of using synthetic data, when  $P = 20$ . Black regions show selected features.

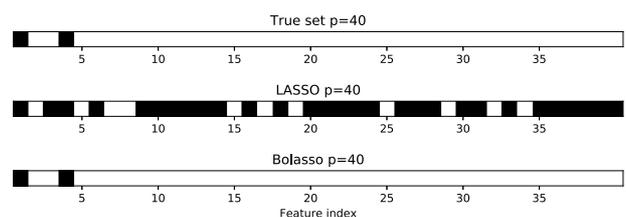


図 4 人工データを用いた  $P = 40$  のときの結果

Fig. 4 The result of using synthetic data, when  $P = 40$ .

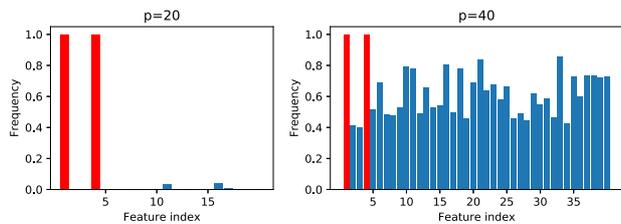


図 5 LASSO による選択. 横軸は特徴番号を表し縦軸が選択された頻度を表す. 左右の図は, それぞれ  $P = 20$  と  $P = 40$  のときの結果を示す

Fig. 5 Selection by LASSO. The horizontal axis represents feature numbers and the vertical axis represents the frequency of selection. The left and right figures show the results for  $P = 20$  and  $P = 40$ , respectively.

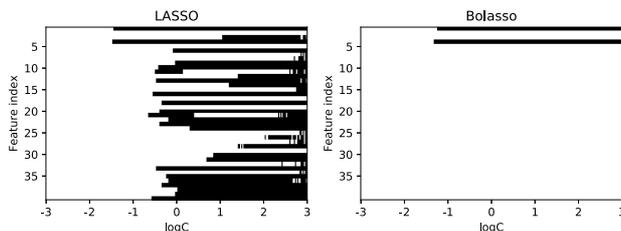


図 6 ハイパーパラメータ  $C$  に対する LASSO と Bolasso の選択特徴の結果. 横軸は  $\log C$  を表し, 縦軸は要素番号を表す

Fig. 6 Results of LASSO and Bolasso selection features for hyperparameter  $C$ . The horizontal axis represents  $\log C$ , and the vertical axis represents feature index.

選択結果が変動しているが, 設定した特徴は確実に選択されているため, Bolasso においては正しく選択が行われている.

図 6 に,  $P = 40$  のときの  $C$  に対する LASSO と Bolasso が選択する特徴の変化を示す. LASSO では  $C$  の値によって推定される組合せが変化しており,  $C$  の決め方が推定結果に強く影響していることが分かる. さらに,  $C$  の値を大きくしていくとともに特徴が増えていくのではなく, 使われていた特徴が使われなくなるような状況を確認できた. 一方で Bolasso は,  $C$  の値が大きく, 正則化項の効果が弱いような場合であっても, 設定した特徴を選択することができている.

これらの結果は, 分類に寄与する特徴が存在する場合, Bolasso による特徴選択が安定して動作することを示唆している.

#### 4. びまん性肺疾患に対する実験

本章ではびまん性肺疾患を含む, 肺で見られる陰影パターンの分類における特徴選択の結果とその考察を行う.

##### 4.1 データセット

肺疾患画像のデータセットは, 大阪大学より提供されたびまん性肺疾患を含む肺 X 線 CT データを用いた. 提供された CT データには, 16 [bit] の CT 値が記録されているた

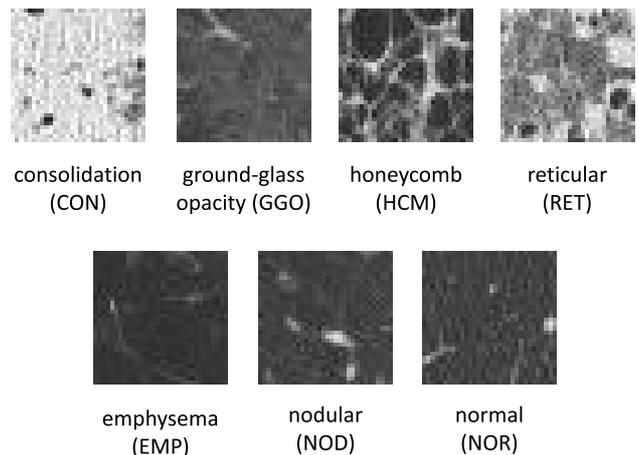


図 7 びまん性肺疾患を含む陰影パターンの例. 上段 4 種がびまん性肺疾患で見られる陰影である

Fig. 7 Example of shading pattern including diffuse lung disease. The upper four types are the shadows seen in diffuse lung disease.

表 2 各クラスの ROI 画像の枚数

Table 2 Number of ROI images in each classes.

クラス	学習用	テスト用
Consolidation (CON)	26 枚	26 枚
Ground-Glass Opacity (GGO)	50 枚	46 枚
Honeycomb (HCM)	72 枚	73 枚
Reticular (RET)	70 枚	66 枚
Emphysema (EMP)	294 枚	296 枚
Nodular (NOD)	65 枚	65 枚
Normal (NOR)	359 枚	355 枚
計	936 枚	927 枚

め, 肺野条件にあわせて階調変換を行い 8 [bit] の CT 画像として扱った. CT 画像にはあらかじめ医師による陰影パターンに基づいた領域指定が行われているため, 指示された領域が  $32 \times 32$  [pixels] に 80% の領域が含まれるような関心領域画像 (Region of Interest: ROI) を切り出し, データとして用いた. 陰影パターンは正常 (Normal; NOR) な陰影に加え, びまん性肺疾患における陰影として浸潤影 (Consolidation; CON), すりガラス状陰影 (Ground Glass Opacity; GGO), 蜂巣状陰影 (Honeycomb; HCM), 網状影 (Reticular; RET) の 4 種, その他の病気が呈する異常陰影として肺気腫 (Emphysema; EMP), 粒状影 (Nodular; NOD) の 2 種の計 7 クラスを用いた. 図 7 に 7 クラスの陰影ごとに切り出された画像の一例を示す. 最終的に学習用に 936 枚, テスト用に 927 枚の画像を準備した. 表 2 に学習用とテスト用データ内におけるクラスの内訳を示す. 学習用データセットと訓練用データセットは医師の指示のもと, 患者が異なるようなデータセットとなっている.

これらのデータセットから, Sugata らの手法に基づき [6], 同時生起行列 (Co-Occurance Matrix; COM), ランレングス行列 (Run Lenght Matrix; RLM), 濃淡ヒストグ

表 3 テクスチャ特徴の解析と要素番号

Table 3 Analysis and index of texture features.

要素番号	テクスチャ解析 (略称)
0-5	同時生起行列 (COM)
6-10	ランレングス行列 (RLM)
11-17	濃淡ヒストグラム (GLH)
18-24	差分統計量 (GLD)
25-31	動径方向フーリエパワースペクトル (Fr)
32-38	方位フーリエパワースペクトル (Ft)

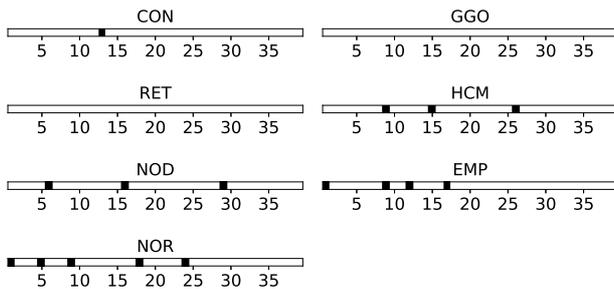


図 8 Bolasso によって得られた選択特徴の結果

Fig. 8 Results of selected features by Bolasso.

ラム (Gray Level Histogram; GLH), 差分統計量 (Gray-Level Differential; GLD), 動径フーリエパワースペクトル (Fourier Power Spectrum of  $r$ ; Fr), 方位フーリエパワースペクトル (Fourier Power Spectrum of  $\theta$ ; Ft), といった 6 種のテクスチャ解析によって特徴抽出を行い, 39 種のテクスチャ特徴を算出した. 表 3 に各テクスチャ解析から得られる特徴が対応する要素番号を示す. 算出したテクスチャ特徴ごとに, 前処理として各特徴量に対して平均 0, 分散 1 となるような線形変換を適用し, 実験を行った.

#### 4.2 Bolasso による特徴選択

はじめに, Bolasso を用いた特徴選択を行った. 肺の陰影パターンは 7 クラスあるため, 多クラスの分類として各クラスごとに 1 対他の識別器の構築を行った. Bolasso を適用するにあたり, レプリカ数  $M = 1,000$  とした. また, LASSO のハイパーパラメータ  $C$  は各クラスごとに  $10^{-3}$  から  $10^3$  の範囲で探索を行い, 10 分割交差検証法による交差検証誤差が最も小さくなるような値を設定した.

図 8 に各クラスにおいて Bolasso によって推定された組み合わせを示す. CON, HCM, EMP, NOR, NOD のクラスにおいて Bolasso を適用したところ, いくつかの特徴が選択された. 推定された組み合わせはクラスごとに異なる特徴の選び方を表しており, それぞれの陰影を表現する特徴を選び出すことができたと考えられる. 一方, GGO と RET のクラスに関しては Bolasso によって選択された特徴は存在していなかった. この結果から, GGO と RET の 2 クラスにおいては有効な特徴が存在していないことが考えられる. これは先行研究 [10] で指摘されている結果と一

表 4 LASSO と Bolasso, Bolasso-S で選択した特徴数

Table 4 Number of selected features by LASSO, Bolasso and Bolasso-S.

	CON	GGO	HCM	RET	EMP	NOD	NOR
LASSO	3	5	30	2	14	24	17
Bolasso	1	0	3	0	4	3	5
Bolasso-S	2	2	14	0	7	8	11

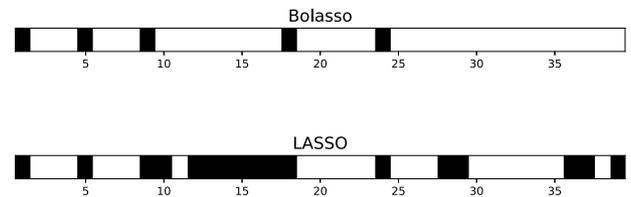


図 9 NOR クラスにおける選択特徴の結果

Fig. 9 Results of selected features for NOR class.

致しており, この 2 クラスを十分に表現することができる特徴が存在していないことを Bolasso から確認することができた.

LASSO を適用した場合の選択結果と Bolasso による選択結果を比較した. 表 4 に各クラスごとに LASSO と Bolasso で選択した特徴の数を示す. Bolasso において特徴の選択が行われた 5 つのクラスでは, LASSO で選択される特徴数と比べて, より限られた特徴を選択していることが分かる. 一方で, GGO, RET の 2 クラスは Bolasso では特徴が選択されていなかったが, LASSO では特徴が選択されている. これはハイパーパラメータ  $C$  に非常に小さい値を用いているため, Bolasso の選択基準を満たす特徴が現れなかったと考えられる. また, このときの  $C$  の値に対して求めた交差検証誤差は, すべての入力に対して目的クラス以外であるとするような値であったため, GGO, RET クラスを十分に分類することができないと考えられる.

表 4 中の Bolasso-S [1] は, Bolasso の選択基準を 0.9 以上の頻度で選択される特徴に緩和した場合の選択結果である. Bolasso-S のとき, GGO クラスでは特徴が 2 つ選択されており, 解釈に適しているかもしれない特徴が存在している. しかし, RET クラスでは Bolasso-S でも選択される特徴はなく, 解釈に適した特徴が存在していないと考えられる.

図 9 は NOR クラスにおける LASSO と Bolasso の特徴選択結果である. LASSO で選択される特徴と比較を行うと, LASSO で選択された 17 個の特徴から, Bolasso ではさらに 5 個まで特徴数を削減することができた. このように単に LASSO を適用した場合, 最適な特徴として多めに見積もられてしまっている. Bolasso では, テクスチャ特徴の中でもデータに対して選択が頑健な特徴を選ぶことができています.

図 10 の特徴ごとの選択確率を確認すると, LASSO で

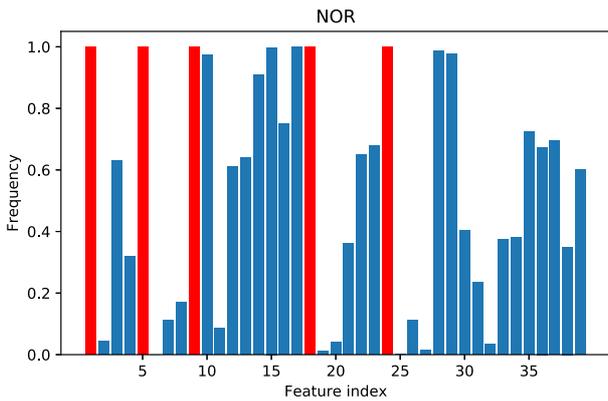


図 10 LASSO による NOR クラスの選択. 横軸が特徴の要素番号を表し縦軸が頻度を表す. 図中赤い部分は Bolasso によって選択された特徴を表す

Fig. 10 Selection of NOR class by LASSO. The horizontal axis represents feature element numbers and the vertical axis represents frequency. In this figure, the red parts represent the features selected by Bolasso.

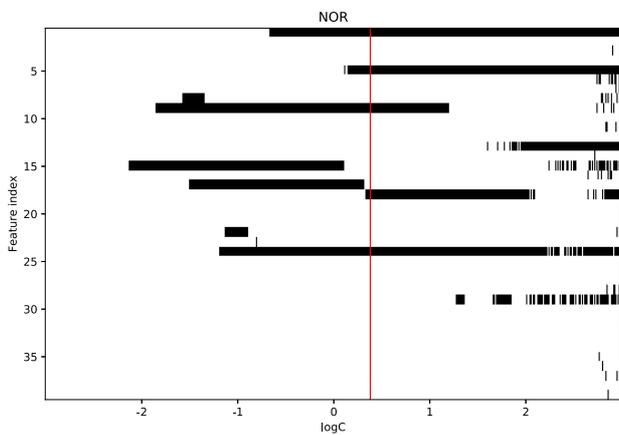


図 11 ハイパーパラメータ  $C$  に対する NOR クラスにおける Bolasso の選択特徴の結果. 横軸は  $\log C$ , 縦軸は要素番号を表す. 赤線は交差検証誤差を最小とする  $C$  の値を示す

Fig. 11 Results of Bolasso selection features of NOR class for hyperparameter  $C$ . The horizontal axis represents  $\log C$ , and the vertical axis represents feature index. Red line shows value of  $C$  which minimizes cross validation error.

選択されているような特徴は Bolasso の過程においても比較的选择されているものの, 生成されたレプリカによっては選択されていないことが確認できる. 高い頻度で選択されている特徴と Bolasso で選択された特徴との相関を調べたところ, いくつかの特徴の組において高い相関関係があることを確認した.

図 11 はハイパーパラメータ  $C$  の値に対する Bolasso の選択特徴を示している. 赤線で示した値は交差検証誤差が最小となった  $C$  である. 図 10 から,  $x_{17}$  と  $x_{18}$  はどちらも非常に高い頻度で選択される特徴であるが,  $C$  の値を増減させることで, 切り替わるように Bolasso の選択が変化していることが分かる. この 2 つの特徴は相関係数が  $-0.9$

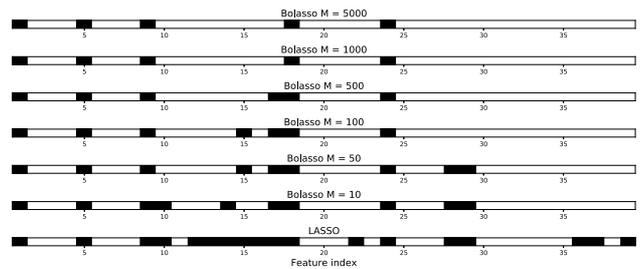


図 12 NOR クラスにおけるレプリカ数  $M$  に関する Bolasso の選択特徴

Fig. 12 In NOR class, Selected features of Bolasso on  $M$ .

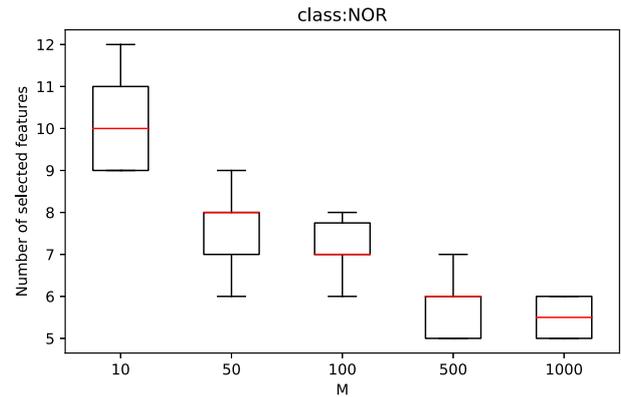


図 13 レプリカ数  $M$  に関して Bolasso で選択される特徴数の箱ひげ図

Fig. 13 Boxplot of number of selected features with Bolasso on  $M$ .

以下である負の相関関係があり,  $C$  の値に合わせて適した特徴を選択されていると考えられる.

図 12 は NOR クラスにおいて, レプリカ数  $M$  を変化させて Bolasso を適用したときの選択した特徴を示している.  $M$  が 10 や 50 のように小さいとき,  $M$  を十分に大きくしたときに選択されない特徴が選択されていることが確認できる. 図 13 はレプリカ数  $M$  に関して選択された特徴数の箱ひげ図である.  $M$  が 10, 50, 100, 500, 1,000 の場合に関して, それぞれ Bolasso を 10 回試行し, 選択された特徴の数を確認した. 図 13 中の箱は四分位を表し, 箱中の赤線は中央値を表している.  $M$  が小さいとき, とりうる特徴の数のばらつきがあることが分かる. 一方で,  $M$  を大きくしていくことで, 選択される特徴が徐々に少なくなり, とりうる特徴の数のばらつきも小さくなっている. これは,  $M$  が小さいときには選択される特徴が用いたレプリカに強く依存してしまうため, 結果にばらつきが生じてしまっていると考えられる. そのため, 十分な数のレプリカを用いることで, より安定して選択される特徴を特定することができる. また, 本実験では  $M = 1,000$  として Bolasso を適用したが, 図 12 の  $M = 5,000$  の結果と一致していることから, 安定した特徴を選択するために十分なレプリカ数であったと考えられる.

4.3 テストデータによる評価

Bolasso で選択された特徴を用いた識別器のテストデータによる評価を行った。対象とするクラスは Bolasso において特徴の選択が行われた CON, HCM, EMP, NOR, NOD の 5 クラスとした。比較対象として、テクスチャ特徴全 39 種を利用した場合と LASSO で選択された特徴を利用した場合を考えた。

表 5 は各クラスごとの特徴選択によるテストデータの誤答率である。テストデータによる評価を行ったすべてのクラスにおいて、Bolasso で選ばれた特徴によって誤答率 10%以下の識別器が構築できた。テストデータによる性能評価の観点からも、テクスチャ特徴から各クラスの解釈において有効な特徴が選択されたと考えられる。CON, HCM, NOD, NOR の 4 クラスでは、Bolasso を適用した場合がすべての特徴や LASSO で選ばれた特徴を用いた場合よりも誤答率が低く、不要な特徴を取り除くことによる汎化性能の向上が確認できる。一方で EMP クラスでは、Bolasso の適用によって LASSO などの結果と比較して誤答率が約 6%程悪化しており、他のクラスに比べ識別性能の差が大きかった。

図 14 は EMP クラスの選択特徴の結果の比較である。EMP クラスでは LASSO による特徴選択の結果が最もテストデータの誤答率が低く、LASSO で選択された特徴が分

表 5 テストデータによる性能評価

Table 5 Evaluation of performance by test data.

	CON	HCM	EMP	NOD	NOR
全特徴	0.004	0.016	0.044	0.088	0.102
LASSO	0.003	0.016	<b>0.033</b>	0.082	0.093
Bolasso	<b>0.002</b>	<b>0.009</b>	0.092	<b>0.070</b>	<b>0.081</b>

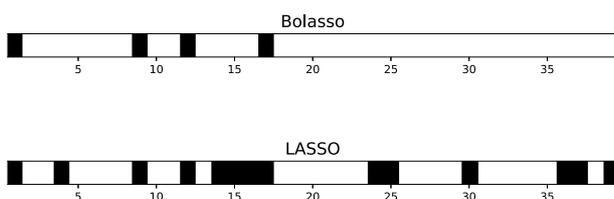


図 14 EMP クラスにおける選択特徴の結果

Fig. 14 Results of selected features for EMP class.

類に必要な特徴であると考えられる。Bolasso では LASSO で選択された特徴の一部が選択されており、Bolasso によって最も分類に寄与している軸が選ばれているといえる。そのため、LASSO が選択し、Bolasso が選択しなかった特徴の中に、識別性能に差を付けるような表現が存在していると考えられる。

4.4 選択された特徴による陰影の解釈

表 6 に、Bolasso によって選ばれた特徴を示す。本研究では、びまん性肺疾患の陰影である CON, HCM の 2 クラスに対して、陰影の解釈に関する考察を行った。

CON クラスでは GLM のコントラストが選択されている。GLM のコントラストは画素濃淡値がどの程度偏っているのかを示している。この特徴から、CON クラスは他の陰影に比べて関心領域内において暗い領域が極端に少なく、明るい領域が広がっている画像であると解釈することができる。CON クラスは CT において均一な透過性の低下が見られるような陰影を指しており [12]、均一な透過性の低下を高い階調値への偏りが示していると考えられる。

HCM クラスにおける特徴選択では、RLM の Gray Level Nonuniformity, GLM の歪度, Fr の平均が選択された。HCM クラスの陰影は、肺の構造が病気の進行によって変形し、蜂の巣のような構造が陰影として確認することができる [12]。特徴選択によって選ばれた Fr の平均は画像に含まれる周波数の期待値であり、HCM クラスで生じる特有の構造をとらえていると考えられる。また、特有の構造とそれ以外の領域で明暗がはっきりとしており、階調値の分布の歪みを示す GLM の skewness が特徴として選択されていると考えられる。RLM の Gray Level Nonuniformity は、ランと呼ばれる同一の階調値連続が現れる階調値の多さを示している。RLM の Gray Level Nonuniformity によって、HCM クラスが持つ特有の構造がつながっていることや、暗い領域が広がっていることが特徴であることを示唆している。

5. まとめ

本研究では、びまん性肺疾患における陰影の特徴表現を

表 6 Bolasso で選択された特徴の一覧

Table 6 List of selected features by Bolasso.

テクスチャ解析	クラス				
	CON	HCM	NOD	EMP	NOR
COM	-	-	Inverse Difference Moment	energy	energy, entropy
RLM	-	Gray Level Nonuniformity	-	Gray Level Nonuniformity	Gray Level Nonuniformity
GLH	contrast	skewness	kurtosis	mean, energy	entropy
GLD	-	-	-	-	energy
Fr	-	mean	kurtosis	-	-
Ft	-	-	-	-	-

明らかにするために、Bolasso を用いた特徴選択を行った。Bolasso は、ブートストラップ法と LASSO を組合せ手法であり、よりデータやハイパーパラメータに頑健な特徴を選択することができる。人工データを用いた実験では、LASSO による特徴選択では不要な特徴が含まれてしまう可能性と、Bolasso を適用することで適切に有効な特徴を推定できることを確認した。びまん性肺疾患の特徴選択において Bolasso を適用したところ、LASSO によって選ばれる特徴よりも少数の特徴を選択することができ、より陰影の表現として有効な特徴の絞り込みができた。Bolasso によって選択された特徴はテストデータの評価でも十分な識別性能を示し、文献 [12] に記述された解釈と対応がとれるような特徴が選択されていることを確認した。

これまでの研究で、特徴選択の問題において同程度に有効な組合せが数多く存在していることが確認され、解釈として妥当な特徴を絞り込むことが困難であったが、Bolasso を適用させることによって、候補となる組合せからさらに特徴を絞り込むことに成功した。この結果から、解釈性を求める手法においてもアンサンブル学習のような、複数の解を統合する取り組みが有効であることが確認できた。

また、EMP のように今回選択された特徴以外に、識別に有効な特徴が存在していることが考えられる。このような表現を特徴の組合せで分解を行うことによって、クラスに内在する複数のパターンなどを明らかにすることができると思われる。

謝辞 本研究の実験にて使用した X 線 CT データを提供していただいた、大阪大学附属病院に深く感謝いたします。また、本研究の一部は、科学研究費 (JSPS) 16K00328, 19H04982 のサポートを受けて行われた。

## 参考文献

- [1] Bach, F.R.: Bolasso: Model consistent Lasso estimation through the bootstrap, *Proc. 25th International Conference on Machine Learning, ICML '08*, pp.33-40, ACM (online), DOI: 10.1145/1390156.1390161 (2008).
- [2] Efron, B. and Tibshirani, R.J.: *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, No.57, Chapman & Hall/CRC (1993).
- [3] Hara, S. and Maehara, T.: Enumerate Lasso Solutions for Feature Selection, *AAAI* (2017).
- [4] Inagaki, T., Shouno, H. and Kido, S.: Classification of Idiopathic Interstitial Pneumonia CT Images using Convolutional-net with Sparse Feature Extractors (2012).
- [5] Ono, S., Koiwai, M. and Shouno, H.: Comparison of Feature Selection method for Diffuse Lung Disease (2017).
- [6] Sugata, Y., Kido, S. and Shouno, H.: Comparison of two-dimensional with three-dimensional analyses for diffuse lung diseases from thoracic CT images, *Medical Imaging and Information Sciences*, Vol.25, No.3, pp.43-47 (online), DOI: 10.11318/mii.25.43 (2008).
- [7] Suzuki, A., Sakanashi, H., Kido, S. and Shouno, H.: Feature Representation Analysis of Deep Convolutional Neural Network using Two-stage Feature Transfer—An Application for Diffuse Lung Disease Classification, *情報処理学会論文誌数理モデル化と応用 (TOM)*, Vol.11, No.3, pp.74-83 (2018) (オンライン), 入手先 (<https://ci.nii.ac.jp/naid/170000149956/>).
- [8] Tibshirani, R.: Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol.58, No.1, pp.267-288 (1996).
- [9] Xu, R., Hirano, Y., Tachibana, R. and Kido, S.: Classification of Diffuse Lung Disease Patterns on High-Resolution Computed Tomography by a Bag of Words Approach, *Medical Image Computing and Computer Assisted Intervention, MICCAI 2011*, Fichtinger, G., Martel, A. and Peters, T. (Eds.), Berlin, Heidelberg, Springer Berlin Heidelberg, pp.183-190 (2011).
- [10] 遠藤瑛泰, 永田賢二, 木戸尚治, 庄野 逸: びまん性肺疾患診断における階層的特徴選択アプローチ, *技術報告 33* (2018).
- [11] 川端大貴, 市川寛子, 永田賢二, 永福智志, 田村了以, 岡田真人: ES-SVM の解空間の解析, *人工知能学会全国大会論文集*, Vol.JSAI2016, pp.2L51in2-2L51in2 (2016).
- [12] 酒井文和: 画像から学ぶびまん性肺疾患, 克誠堂出版 (2018).



遠藤 瑛泰 (学生会員)

1995 年生。2018 年電気通信大学情報理工学部総合情報学科卒業。現在、同大学大学院修士課程在学中。



永田 賢二 (正会員)

2004 年東京工業大学工学部情報工学科卒業, 2008 年同大学大学院総合理工学研究科博士課程修了。博士 (工学)。その後、東京大学特任研究員, 特任助教, 助教, 産総研主任研究員, 研究チーム長を経て, 2019 年より物質・材料研究機構主任研究員。ベイズ推定, データ駆動科学, マテリアルズインフォマティクス等の研究に従事。IEEE, 電子情報通信学会, 日本物理学会各会員。



木戸 尚治

1988年大阪大学医学部医学科卒業。  
1992年同大学大学院博士課程修了。  
同附属病院，八尾市立病院，西宮市立  
中央病院，大阪府立成人病センター勤  
務。1999年山口大学工学部知能情報  
システム工学科教授。2006年同大学

院医学系研究科応用医工学系学域教授。2016年同大学院創  
成科学研究科（工学系学域）知能情報工学分野教授。2019  
年より大阪大学大学院医学系研究科人工知能画像診断学共  
同研究講座特任教授。博士（医学）・博士（工学），日本医  
学放射線学会認定放射線科専門医。専門：人工知能画像診  
断学，胸部画像診断学。所属学会：日本医学放射線学会，  
日本医用画像工学会，電子情報通信学会，日本生体医工学  
会，北米放射線学会等の各会員。



庄野 逸（正会員）

1968年生。1992年大阪大学基礎工学  
部生物工学科卒業。1994年同大学大  
学院修士課程終了。1994年同大学助  
手。2001年奈良女子大学助手。2002  
年山口大学助教授。2008年電気通信  
大学准教授を経て2015年同大学教授。

機械学習，ニューラルネットワーク，画像処理の研究に従  
事。IEEE，電子情報通信学会，日本神経回路学会各会員。