全層転移学習によるプロテオミクス解析

Proteomics Analysis using All-Transfer Deep Learning

澤田好秀 Yoshihide Sawada
佐藤佳州 Yoshikuni Sato
中田透 Toru Nakada
氏本慧 Kei Ujimoto
林宣宏 Nobuhiro Hayashi

要　旨

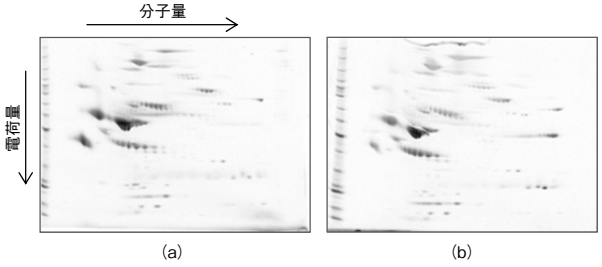
本稿では，少数の学習データを用いて高精度な深層学習を実施するための新たな転移学習法を提案し，それをプロテオミクス解析に適用した結果を報告する．プロテオミクス解析とは，たんぱく質を分子量と電荷量に基づいて分離し画像化した二次元電気泳動像を利用してたんぱく質の解析を行う技術であり，生物学者や医学者より注目を集めている．疾病の識別や創薬の分析，テーラーメイドの化粧品などに役立つと言われている一方，検体の収集が困難であり，高精度な判定が難しいという課題がある．そこで本稿では，転移学習と深層学習を組み合わせることで小規模データにおける識別精度の向上を試みた．従来と異なり，ニューラルネットワークの出力層も含めた全層を再利用する手法を提案し，有効性を敗血症識別によって確認できたので報告する．

Abstract

We have proposed a novel method of transfer learning for deep neural networks and applied it to proteomics analysis. Proteomics analysis uses two-dimensional electrophoresis images split off from proteins on the basis of the degree of charge and molecular weights, and it is widely attracting attention from biological researchers and doctors. However, it is difficult for proteomics analysis to accurately classify diseases, lead to drug discoveries, and produce tailor-made cosmetics because it is not easy to collect the images needed for the analysis. In this article, we describe our efforts to improve the classification performance of the analysis by combining transfer learning and deep learning when we only have a small-scale dataset. The main difference between our proposed method and the conventional method is that we transfer all layers of neural networks including the output layer. We confirmed the effectiveness of our method by applying it to sepsis classification.

1. はじめに

近年，プロテオミクス解析が注目を浴びている[1]．プロテオミクス解析とは，たんぱく質を分子量と電荷量に基づいて分離し画像化した二次元電気泳動像（Two Dimensional Electrophoresis Images, TDEIs）を利用して，体内のたんぱく質の状態を解析する技術のことである．TDEIsの例を**第1図**に示す．TDEIsはX軸が分子量，Y軸が電荷量での分離結果を表し，画像中の黒領域がたんぱく質を表す．これにより，複数のたんぱく質の状態を一度に，かつ包括的に確認することができ，疾病の識別や創薬の分析，テーラーメイドの化粧品などに役立つと言われている[1].



第1図　二次元電気泳動像の例
(a)　敗血症患者の二次元電気泳動像
(b)　非敗血症患者の二次元電気泳動像

Fig. 1　Examples of two-dimensional electrophoresis images
(a)　Images of patients suffering from sepsis
(b)　Images of patients free from diseases.

通常，たんぱく質の状態を計測し，その計測結果より診断を行う場合には，バイオマーカーを利用する．バイオマーカーは特定のたんぱく質が明確に発現しているか否かを判定する際に利用され，たんぱく質との生物学的な相関関係が明確な疾病を診断する場合に広く利用されている．そのため，たんぱく質との相関関係が不明瞭な場合や複数のたんぱく質の微細な変化が診断に有用な場合には，十分な性能を発揮することができない．

そのような疾病の一例として敗血症が挙げられる．敗血症とは，体内に入った細菌が原因で多臓器不全を起こす，全身性炎症反応症候群のことであり，患者の約25％が死に至る病である．この重篤性を考慮して，敗血症をターゲットとした診断支援システムについて報告する．

本稿では，診断支援システムとして，教師有学習[2]を採用する．教師有学習では，事前に正解ラベル付きの学習用のデータを複数用意しておき，そのデータを利用してモデルを構築する．そして，新規データが入力された際には，そのモデルを利用して新規データの正解ラベルを推定する．一方，TDEIsを対象とした場合，TDEIsを識別に利用した先行研究は存在しないため，識別に有効な

画像特徴量は未解明である．そのため，自動で画像特徴を学習できる深層学習[2]の適用が望ましい．しかし，TDEIsの生成機器のスループットの低さ，および患者データであることに起因するデータ収集の困難さより，学習用データを十分に収集することは難しい．そのため，一般に大量のデータを利用する深層学習はそのままでは適用できない．もし，そのまま適用した場合，システムの識別率は低下し，さらには過学習[3]に陥る．

そこで本稿では，高精度な識別を行うために深層学習に転移学習を適用する．転移学習とは，対象とする識別タスクに関するデータ（転移先データ）とは異なるデータ（転移元データ）を用いてモデルを構築し，そのモデルを再利用することで識別精度を向上させる手法であり，深層学習への適用も広く研究されている[4][5]．これらの手法は，まず転移元データを用いてモデルを構築する．次に，転移元データで構築したモデルを転移先データの識別に利用する．この際，転移先と転移元データの教師ラベルが異なり，直接利用することが困難であるため，出力層のみ，もしくは出力層を含む中間層の一部（上位層）を取り除く．そして，取り除いた層を置き換え，置き換えた層，もしくは全層を転移先データで再学習する．しかし，このアプローチでは，少数の学習データで上位層を再学習する必要があるため，それによる識別精度の低下や過学習を防ぐことは難しい．これを防ぐには，転移元データで構築したモデルの出力層を含む，すべての層を再利用する必要があると考える．そこで本稿では，出力層の情報も含めてすべて再利用する新たな手法である，全層転移学習を提案する．

以降では，提案手法について説明した後に，TDEIsを用いた敗血症識別を対象とした実験結果を示し，提案手法の有効性を示す．

2. 全層転移学習

提案手法の概要を**第2図**に示す[6]．以降では，**第2図**に従って提案手法の各ステップについて説明する．まず，転移元データを用いて深層ネットワーク（Deep Neural Network, DNN)を下式に基づいて構築する（**第2図 (a)**）．

$$l(\{y^s, x^s\}) = \frac{1}{N^s}\sum_j \left| y_j^s - f(h_L \mid x_j^s) \right|^2 \quad \cdots\cdots\cdots \quad (1)$$

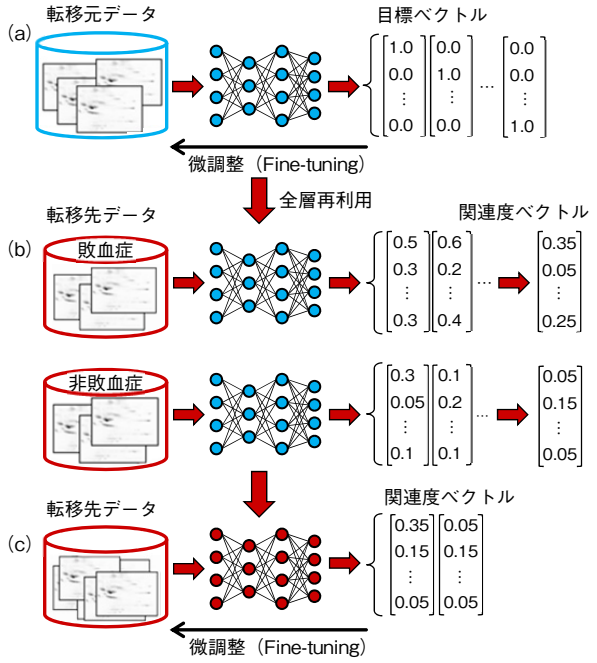
ただし，y_j^sは下式のようにj番目の転移元データの正解ラベルに対応するone-hotベクトルを表す．

$$y_j^s = [y_{j,1}^s, y_{j,2}^s, \ldots, y_{j,D_{L+1}}^s]^{\mathrm{T}} \quad y_i^s = \{0,1\} \quad \cdots\cdots\cdots \quad (2)$$

D_{L+1}は転移元データのカテゴリ数，x_j^sはj番目の転移元データ，N^sは転移元データ数を表し，$f(h_L|x_j^s)$はL+1番目の層（出力層）の次元のベクトルを表す．

$$f(h_L \mid x_j^s) = h_{L+1} = W_{L+1}h_L + b_{L+1} \quad \cdots\cdots\cdots\cdots \quad (3)$$

なお，h_LはL番目の層を表し，W_LはL番目の層の重み，b_Lはバイアスを表す．



第2図　提案手法の概要[6]
(a) 転移元データを利用してDNNを学習
(b) 転移先データのカテゴリごとに関連度ベクトルを推定
(c) モデル全体に微調整をかけ，転移元データで構築したモデルを転移先データ用のモデルへと転移

Fig. 2　Overview of our proposed method [6]
(a) Training DNN for source task
(b) Computing target vectors of each target label
(c) Tuning all parameters to transfer

次に，構築したDNNにl番目のカテゴリに対応する転移先データx_j^tを直接入力することで，転移先データのl番目のカテゴリに対応する関連度ベクトルr_lを（4）式に従って算出する（**第2図 (b)**）．関連度ベクトルは，転移元データのカテゴリによって構成させるD_{L+1}次元の基底ベクトル（（2）式）で表現される空間上で，転移先データの各カテゴリを表現する代表ベクトルのことである．

$$r_l = \arg\max_{h_{L+1}} p(h_{L+1} \mid x_j^t) \quad \cdots\cdots\cdots\cdots\cdots \quad (4)$$

本稿では，確率分布$p(h_{L+1}|x_l^t)$に正規分布を仮定する．これは，転移元DNNを（1）式のように回帰モデルによって構築しているため，自然な仮定である．すると，関連度ベクトルr_lは下記のように，l番目のカテゴリに対応する転移先データの出力ベクトルの平均で表現することができる．

$$r_l = \frac{1}{N^t(l)}\sum_j^{n^t(l)} f(h_L \mid x_j^t) \quad \cdots\cdots\cdots\cdots\cdots \quad (5)$$

ただし，$N^t(l)$はl番目のカテゴリに対応する転移先データ数を表す．このステップにより，転移元データで構築したDNNの出力層を取り除く必要なく，出力層の結果も再利用可能となる．

最後に，下式に基づいて，モデル全体を微調整（Fine-tuning）することでモデルを転移させる（**第2図（c）**）．

$$l(\{y^t, x^t\}) = \frac{1}{N^t}\sum_j \left| r_l - f(h_L \mid x_j^t) \right|^2 \quad \cdots\cdots\cdots \quad (6)$$

ただし，x_j^tはj番目の転移先データであり，N^tは転移先データ数を表す．なお，x_j^tの正解ラベルy_j^tに対応する関連度ベクトルがr_lである．

この微調整のステップは，モデルの中間層で算出される特徴量も変更しつつ，各カテゴリの関連度ベクトルを中心として，転移先データを各カテゴリに集約させることと等しい．すなわち，転移先データのクラス内分散が小さくなり，Fisherの評価基準[7]が大きくなる効果がある．

これらのステップにより，出力層を含めたすべてのパラメータに転移元データに関する制約を加えることができ，パラメータの初期値を乱数によって決定する際に起こり得る，不適切な局所解を避けることが可能となる．これにより，識別率は低下や過学習を避けることができる．

新規データxに対する識別では，構築したDNNに新規データを入力した際の出力ベクトル$f(h_l|x)$と各カテゴリの関連度ベクトルr_lを比較し，新規データxを最も近い関連度ベクトルに対応するカテゴリのデータとして判定する．

3. 実験結果

TDEIsを用いた敗血症識別を対象として，提案手法と他手法との比較実験を行った．比較手法としては，DNNを利用せずに主成分分析を行った後にロジスティック回帰[3]を実施する方法, 転移なし, 半教師有DNN, 文献[4], および文献[5]の手法を用いた．なお，半教師有DNNは転移先データだけでなく転移元データも含めて各層の事前学習を行う手法のことを指す．DNNの学習アルゴリズムとして積層雑音除去自己符号化器（Stacked Denoising Autoencoders, SdA）[2]を採用し，微調整には確率的勾配降下法[2]を利用した．

本実験では，敗血症患者のTDEIsを30例，非敗血症患者のTDEIsを68例用意した．画像サイズは53 pixel×44 pixelのグレースケールであり, 2分割交差検定にて精度を評価した．

3.1 他手法との比較結果

本節の実験では，転移元データとして，敗血症／非敗血症の正解ラベルが付いたTDEIsとは異なる別のTDEIsのデータセットを用いた．具体的には，TDEIsの生成の違い[6]を識別するDNNを構築した. 詳細を**第1表**に示す. 二次元電気泳動像はさまざまな方法で生成することが可能である[1]．本実験では，10種類の生成方法で二次元電気泳動像を生成し, そのうちの9種類を転移元データとして利用し（**第1表**），これら生成方法を転移元データの正解ラベルとして9種類のカテゴリ, 合計180症例用意した. また, 残りの1種類の方法で生成された二次元電気泳動像を転移先データとして利用し，敗血症／非敗血症の正解ラベルを付した．すなわち，転移元と転移先データは，正解ラベルも画像の作成方法も全く異なるデータを利用していることに注意する．

第1表　二次元電気泳動像の転移元データのリスト
Table 1　List of source two-dimensional electrophoresis images

# of images	Type of protocol
$N^s(1)$=25	Change an amount of protein
$N^s(2)$=4	Change a concentration protocol
$N^s(3)$=30	Unprocessed
$N^s(4)$=49	Only top-2 abundant protein removal
$N^s(5)$=11	Top-2 abundant protein concentrated
$N^s(6)$=15	14 abundant protein concentrated
$N^s(7)$=12	Plasma sample instead of serum
$N^s(8)$=19	Sugar chain removal
$N^s(9)$=15	Other protocol

第2表に実験結果を示す．なお，本実験では，中間層の数Lを1, 2, 3, 4と変化させ，各手法の性能を評価し，識別精度（classification accuracy）が最高となった場合を表中に示している．PCAが主成分分析を行った後にロジスティック回帰を実施した手法，NTLが転移なし（Non-Transfer Learning），SSL（Semi-supervised Supervised Learning）が半教師有DNNを表す．また，DNNの中間層の次元数は188次元としており, これは転移元および転移

第2表　中間層の次元数を変更した場合の各手法の精度
Table 2　Classification performance as function of the number of hidden layers

	PPV	MCC	ACC
PCA	0.875	0.545	0.816
NTL(L=1)	0.725	0.755	0.878
SSL(L=1)	0.682	0.736	0.857
[4](L=1)	0.732	0.783	0.888
[5](L=1)	0.750	0.8	0.898
proposed(L=3)	**0.875**	**0.859**	**0.939**

先データを主成分分析し，寄与率99.5 %以上となる次元数である．これらは，コンパクトなDNNを構築することで，転移元データの過学習を抑えるために設定した．また，転移元データにおいても交差検定を行い，転移元データで構築したDNNに過学習が発生していないことは事前に確認している．

表中のACCが識別精度を表している．また，参考指標として，PPV，MCCも求めた．PPVは陽性的中率（Positive Predictive Value），MCC（Matthews Correlation Coefficient）とは下記の式で求まる指標である．

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \quad \cdots\cdots (7)$$

ただし，TPは真陽性（True Positive），TNは真陰性（True Negative），FPは偽陽性（False Positive），FNは偽陰性（False Negative）を表す．PPVは医療現場にてよく利用される指標であり，この値が高いほど確定診断に向いている．一方，MCCは敗血症および非敗血症のデータ数のアンバランスを考慮した精度指標である．なお，どの指標においても，数値が1に近い方が精度が高いことを示していることに注意する．

第2表より，中間層を3層に設定した際の提案手法がPPV，MCC，ACCにおいて最高値を示していることが確認できる（表中の太字）．このことから，提案手法が最も精度が高く，特に確定診断に有用であることがわかる．

3.2 転移元データを変更したときの精度比較

次に，転移元データを変更したときの提案手法の識別性能の変化を示す．本実験では，MNIST [7]とCIFAR-10 [8]の2種類のデータセットを転移元データとして利用し，TDEIsを転移元データとして利用した場合と比較した．MNISTとは，手書き文字の認識のためのデータセットであり，画像サイズは28 pixel×28 pixel，各データには「0」から「9」までの10カテゴリの正解ラベルが付されている．一方，CIFAR-10とは，「飛行機」，「乗用車」，「鳥」，「猫」，「鹿」，「犬」，「蛙」，「馬」，「船」，「トラック」の10カテゴリの正解ラベルが付されたデータセットであり，画像サイズは32 pixcel×32 pixelである．なお，MNISTとCIFAR-10は画像サイズを第3.1節の実験と同様に53 pixcel×44 pixcelへと変更し，かつCIFAR-10はカラー画像であるため，グレースケールへと変換した．また，MNISTとCIFAR-10のデータ数はN^S =50000であり，中間層の数Lは前節で最適であった値に固定した（L=3）．また，MNISTとCIFAR-10のデータ数はTDEIsよりも多いため，PCA基準でTDEIsと同様に次元数を圧縮すると，DNNの表現能力が低下し，識別性能が低下する可能性がある．そこで本稿では，中間層の次元数Di（i=1, 2, 3）は188, 500, 1000のなかから最適な次元数を選択した．なお，TDEIsにおいても同様に，中間層の次元数を188, 500, 1000と変更して実験を行い，188次元の場合が最適であったことは事前に確認している．

第3表に比較結果を示す．この表より，MNISTおよびCIFAR-10を転移元として利用することで，転移学習を実施しない場合よりも識別性能が向上することが確認できる．一方，転移元データ数が少ないにも関わらず，TDEIsを転移元として利用した場合がMCC，ACCは最も精度が高かった．これは，精度向上のためには，例えデータ数が少なくとも，転移先データの識別に有効な基底ベクトル（(2) 式）を利用すべきであることを示唆している．基底ベクトルは，転移元データに依存しており，本手法では転移先データを用いて変更することはできない．そのため，基底の選択方法，すなわち，転移元データの選び方が高精度な識別には重要となる．そこで，次節にて，関連度ベクトル間の距離に基づいた転移元データの選択に関する一考察を述べる．

第3表　転移元データを変更したときの敗血症識別精度の比較
Table 3　Classification performance of actual sepsis data classification for different source tasks

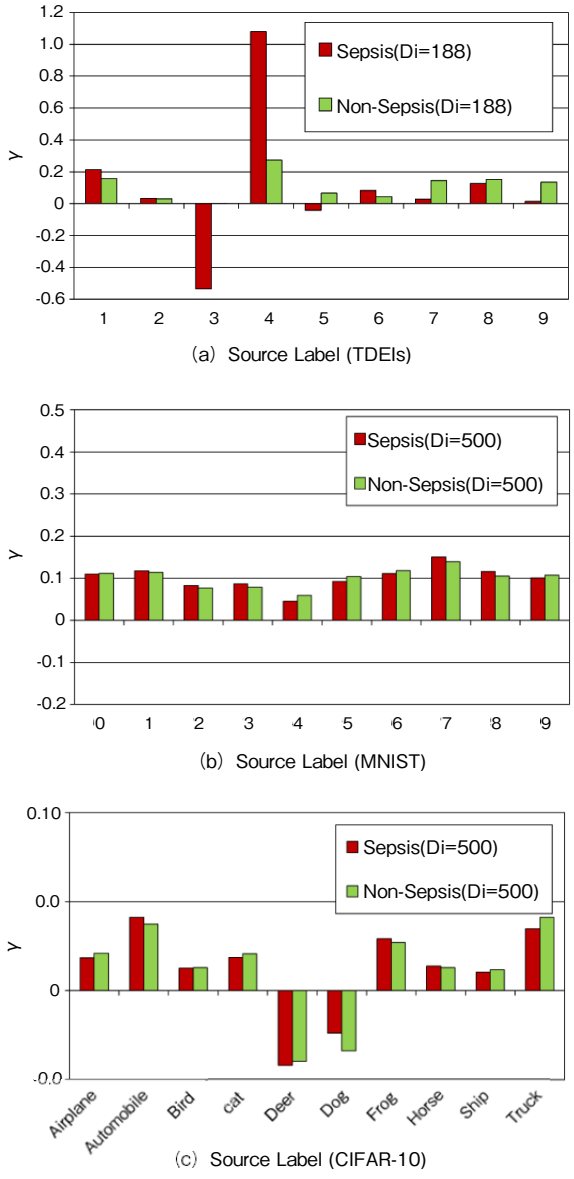
	PPV	MCC	ACC
NTL(Di=188)	0.725	0.755	0.878
CIFAR-10(Di=500)	**0.923**	0.804	0.918
MNIST(Di=500)	0.839	0.786	0.908
TDEIs(Di=188)	0.875	**0.859**	**0.939**

3.3 関連度ベクトル間の距離と認識精度の関係

提案手法は，各カテゴリの関連度ベクトルを算出し，それを利用することで微調整を実施する．そのため，算出した各カテゴリの関連度ベクトル間の距離が近ければ，識別性能は十分に向上しないことが予想される．そこで本節では，識別性能と関連度ベクトル間の距離との関係性についての一考察を示す．

第3図に第3.2節において，転移元データをMNIST，CIFAR-10，TDEIsと設定したときの関連度ベクトルを示す．**第3図（a）**がTDEIs，**(b)**がMNIST，**(c)**がCIFAR-10のときの関連度ベクトルである．図より，認識精度が最も高かったTDEIsを転移元データとして選んだ場合が，敗血症および非敗血症の関連度ベクトルの値が大きく異なっていることが見て取れる．これは，関連度ベクトル間の距離によって，認識精度を微調整前に予測できる可能性を示唆している．したがって，事前に複数の転移元データセットで学習したDNNを用意しておくことができれば，関連度ベクトル間の距離に基づいて，疾病や他

の識別タスクに適切なDNNを選択できると考える．これにより，適切な基底ベクトルの変更が実現できると考える．これに関する詳細な解析は今後の課題である．



第3図　関連度ベクトルの比較
（a）TDEIs，（b）MNIST，（c）CIFAR-10
横軸は各転移元データのラベルを表し（TDEIsは第1表の番号，MNISTは数値），縦軸は関連度ベクトルの各変数の値を表す．

Fig. 3　Comparison of relation vectors
（a）TDEIs，（b）MNIST，（c）CIFAR-10
The X-axis represents the label of source domain (TDEIs is the number in Table 1, and MNIST is the digit), and the Y-axis represents the value of the relative vector.

4. まとめ

本稿では，DNNのすべてのパラメータを転移可能とする新たな転移学習法を提案し，TDEIsを用いた敗血症識別へと適用した．その結果，提案手法が従来法の識別精度を上回ることを確認した．

今回提案した出力層の結果も転移させるという発想はこれまでになく，加えて，機械学習をTDEIsに適用し，疾病診断を行った例もこれまでにない．そのため本稿は，機械学習分野だけでなく，生物学や医学分野に対する貢献も高いと考える．

提案手法によって得られた認識精度は，現在，敗血症のバイオマーカーとして利用されているプレセプシンと比較しても遜色ないものと考える（PPV：86.6 %，MCC：66.5 %，ACC：86.0 %）[8]．しかし，現在のデータ数はまだ十分であるとは言えないため，広範囲なデータの拡充と，さらなる分析が必要である．

本稿で提案した手法は，DNNの構造が大きくなれば（googLeNet [9]など），最後の微調整（**第2図（c）**）では転移先への転移が十分に実施しきれない可能性がある．加えて，本手法は転移元の出力次元を基底とするため，転移元の出力次元数が転移先データ数よりも遥（はる）かに大きい場合には「次元の呪い」と呼ばれる現象によって過学習が起き，遥かに小さい場合には表現能力が低下するために識別精度が減少する可能性が高い．これらの内容に加え，関連度ベクトルの生物学的意味合いを含めた精査や中間層の次元数などのネットワークの最適構造の詳細解析は今後の課題とする．また，softmax関数[2]を利用した場合の関連度ベクトルの推定手法の検討や，複数転移元データからの選択や異なる疾病への適用も行っていく予定である．

参考文献

[1] 林宣宏, “バイオサイエンス 生体の高精度かつハイスループットなプロファイリング--改良型2次元電気泳動法を基盤技術に用いる高性能プロテオミクス,” 未来材料, vol.11, no.1, pp.42-49, 2011.
[2] 麻生英樹 他, 深層学習 Deep Learning, 神嶌敏弘（編）,（株）近代科学社, 東京, 2015.
[3] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2010.
[4] M. Oquab et al., “Learning and transferring mid-level image representations using convolutional neural networks,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1717-1724, June 2014.
[5] P. Agrawal et al., “Analyzing the performance of multilayer neural networks for object recognition,” Proceedings of the European Conference on Computer Vision, pp.329-344, Sept. 2014.
[6] Y. Sawada et al., “All-Transfer Learning for Deep Neural Networks and Its Application to Sepsis Classification,” Proceedings of the European Conference on Artificial Intelligence,

pp. 1586-1587, Aug. 2016.
[7] 石井健一郎 他, わかりやすいパターン認識,（株）オーム社, 東京, 1998.
[8] 金子守 他, “新規敗血症マーカーとしての「プレセプシン」,” 生物試料分析, vol.37, no.5, pp.311-320, 2014.
[9] C. Szegedy et al., “Going deeper with convolutions”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1-9, June 2015.

執筆者紹介

澤田　好秀　Yoshihide Sawada
ビジネスイノベーション本部
AIソリューションセンター
AI Solutions Center, Business Innovation Div.
博士（工学）
（2017年3月まで先端研究本部に所属）

佐藤　佳州　Yoshikuni Sato
ビジネスイノベーション本部
AIソリューションセンター
AI Solutions Center, Business Innovation Div.
博士（工学）
（2017年3月まで先端研究本部に所属）

中田　透　Toru Nakada
先端研究本部　基盤技術研究部
Fundamental Technology Research Department, Advanced Research Div.

氏本　慧　Kei Ujimoto
東京工業大学　生命理工学研究科
School and Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology

林　宣宏　Nobuhiro Hayashi
東京工業大学　生命理工学院生命理工学系
Dept. of Lift Science and Technology, Tokyo Institute of Technology
博士（理学）