情報処理学会論文誌 Vol.63 No.2 379–387 (Feb. 2022)

機械学習による画像診断の差分画像による解析

綾塚 祐二[1,2,a)] 雅樂 隆基[1,b)] 安川 力[3,c)] 吉高 淳夫[2,d)]

受付日 2021年5月6日, 採録日 2021年9月9日

概要：機械学習を用いた画像分類は医療を含め幅広い分野で多くの成果をあげている．しかしその分類は何を根拠としてなされているかが人間には分かりにくい場合も多く，どこを見て分類を行ったかを可視化するためのさまざまな研究が行われている．疾患の分類のような目的のためには，画像中の「それらしい」正の寄与部分（所見）だけでなく「そぐわない」，すなわち負の寄与部分となる所見も診る必要があるが，既存の研究では負に寄与する部分は可視化の対象として注目されていない．我々は，機械学習モデルに対し，画像に微小な差異を加えた画像を入力した場合の確信度の出力の大きさの変化を画像化し提示する手法，DiDA を提案する．提案手法ではグリッド単位で区切りマスクした画像を用いて出力の差異をとらえ，複数のグリッドサイズを用いることで，正負の寄与領域を的確に描出する．DiDA を光干渉断層計による眼底の断層画像からの疾患分類に適用し眼科専門医の見解と照合した結果，DiDA による解析画像は正負の寄与を的確にとらえていることが分かった．また，眼底の断層画像の疾患分類において画像中の正負の寄与領域を既存手法よりも的確に描出することを確かめた．

キーワード：機械学習，医用画像，画像分類，解析，所見

Differential Image Analysis for Image Diagnosis by Machine Learning

Yuji Ayatsuka[1,2,a)] Takaki Uta[1,b)] Tsutomu Yasukawa[3,c)] Atsuo Yoshitaka[2,d)]

Received: May 6, 2021, Accepted: September 9, 2021

Abstract: Image classification by machine learning has already established a variety of excellent results. The classifications of machine learning models are not always understandable for human, so that many researches are trying to visualize areas positively cotributing a classification in an image. However, negative contributions, which are important to diagnose an medical images, have not been paid attention. We propose a new analysis technique named DiDA to visualize positive and negative contribution areas in an image, using sets images with small difference to the original. An image masked along the grid is input to a target machine learning model and its result is compared with the result of the original. Positive and negative differences are represented as different colors in an analyzed grid image. Multiple grid sizes are used for visualizing contributing areas precisely. We applied the DiDA to classifications of optical coherence tomography (OCT) images by mechine learning. The results were checked by a doctor and we found colored areas in resulted images shows appropriate points for diagnoses.

Keywords: machine learning, medical imaging, image classification, analysis, clinical findings

1. はじめに

畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）をはじめとする機械学習を用いた画像分類は幅広い分野で多くの成果をあげている．その 1 つが

[1] 株式会社クレスコ技術研究所
Technology Laboratory, CRESCO LTD., Minato, Tokyo 108–6026, Japan
[2] 北陸先端科学技術大学院大学先端科学技術研究科
Division of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923–1292, Japan
[3] 名古屋市立大学大学院医学研究科視覚科学
Department of Ophthalmology and Visual Science, Nagoya City University Graduate School of Medical Science, Nagoya, Aichi 467–8601, Japan

a) ayatsuka@acm.org
b) t-uta@cresco.co.jp
c) yasukawa@med.nagoya-cu.ac.jp
d) ayoshi@jaist.ac.jp

医療の分野であり，著者らのグループでも光干渉断層計（Optical Coherence Tomography, OCT）による眼底の断層画像から眼底疾患の種類を分類する研究 [1] などを行っており，こうした分類が精度良く行えることを確認している．

一方で機械学習による分類は，何を根拠として分類されたかが人間には分かりにくい場合も多く，それを改善するために「説明可能な AI」と呼ばれる取り組みが行われている．画像分類においても画像中の注目した領域を可視化する方法の研究がすでに数多くなされてきている．しかし，既存の研究ではある分類に属するオブジェクトや領域を適切に抽出することを主眼としており，眼底の断層画像からの眼底疾患の分類のような，症状や状態などの所見の組合せにより総合的に判断すべきような事例には不十分である．特に「この疾患だとするとここに見える異常所見は合わない」というような負の寄与が扱えないという問題がある．

「疾患を正確に診断するために，症状が他の要因で起こっていることを否定し，可能性のある疾患を絞り込む」ことは鑑別診断と呼ばれるが，画像分類においても鑑別の補助となる情報を提供することで医師の診断をサポートできると考えられる．そのためには機械学習による画像の各分類に対する，画像中の正の寄与部分・負の寄与部分をそれぞれ明確に視覚化する必要がある．

我々はその具体的な手段として，ある画像を分類した機械学習モデルに対し，その画像に微小な差異を加えた画像を入力した場合の出力値（分類に対する確信度）の変化の方向と大きさを画像化し提示する手法，DiDA（Differential Image Diagnostic Analysis）を提案する．具体的には，画像をグリッドに区切り，1 カ所ずつグリッドをマスクした画像を機械学習モデルに分類させ出力の変化の方向（正負）で色分けしたものをそのグリッドの色とし，出力の変化の大きさを輝度に反映させた画像を生成する．複数のサイズのグリッドに対してそのような画像を生成し，重ね合わせることで画像中のさまざまな大きさの特徴をとらえ可視化することができる．

本稿はインタラクション 2021 でのデモ発表の内容 [2] に，他の手法との眼底断層画像での比較を加えたものである．

2. 関連研究

画像分類を行う機械学習モデルの分類根拠の可視化を行う既存の研究のアプローチは大別して

(1) CNN の中間層のデータから画像を生成する，
(2) 入力に撹乱を与えた場合の出力値から画像を生成する，
(3) 分類モデルとは別の説明生成モデルを構築する，

に分けられる．以下に述べるように，どのアプローチでも分類ターゲットはオブジェクトあるいは領域として表されるものであることが（暗黙の）前提となっており，ターゲットがどこに存在するかを可視化しようとしている．すなわち，たとえば画像中のどこに「犬」がいるのかを可視化しようとしており，それがなぜ（「猫」ではなく）「犬」と判断されたのかというような情報を提示するものではない．

2.1 CNN の中間層のデータから画像を生成する

DeepLIFT [3] は CNN の分類結果を出力するセルから入力に向かって backpropagate することによって元の画像上へ結果を反映するような形で画像を生成する．CAM [4] や Grad-CAM [5]，Grad-CAM++ [6] では畳み込み層の最終段の出力のエッジの重みや，activation map の変化による出力の変化から画像（attention map）を生成している．Score-CAM [7] は activation map のチャネルを考慮し attention map の精度の改善を試みている．

これらは CNN というアーキテクチャに依存した手法であり，また正の寄与部分のみが扱われている．Attention mining と呼ばれる，attention map の改善を試み，分類の精度自体も向上させる研究（文献 [8], [9], [10] など）でも分類対象の領域を取りこぼしなく抽出することが主眼となっており，負の寄与に関しては扱われていない．

Bach らの研究 [11] では，負の寄与部分が可視化されている例があげられている．しかし，手書き文字の分類を対象としており，分類対象の文字として想定される形状と合致しない部分が示されているにすぎない．

2.2 入力に撹乱を与えた場合の出力値から画像を生成する

Petsiuk らの RISE [12] は，画像をグリッドに区切りランダムに数カ所マスクしたものを多数用意し，それぞれのマスク画像を学習済みの機械学習モデルに対し入力した際の（対象となる分類の）出力を重みとしてそれらのマスクを足し合わせたものを説明画像とする手法である．入力をグリッド単位でマスクするのは我々のアプローチと同じであるが，複数箇所を同時にマスクした画像に対する出力値を重みとしており，結果として基本的には正の寄与部分しか取り出すことができず，可視化されるのは対象の存在する領域となっている．

グリッド単位で正の寄与部分を集めるのはいわば積分的なアプローチといえるのに対し，我々の手法はグリッド単位での正ならびに負の変化を調べる微分的なアプローチであると表現することもできる．RISE では複数のグリッドサイズを用い特徴の大きさをとらえることは想定されておらず，解像度が限定される点も我々の手法との差異である．

2.3 分類モデルとは別の説明生成モデルを構築する

LIME [13] や SHAP [14]，Anchors [15] などは，分類する対象を「特徴のセット」として表し，各特徴の有無が分類にどう影響するかを調べることにより説明モデルを構築する．これらの研究では画像も扱われているが，画像をどのように「特徴のセット」として表すかには触れられておら

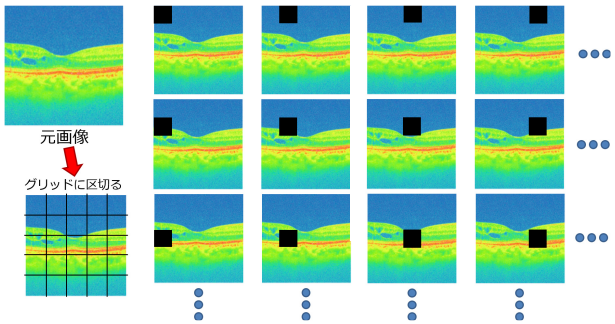

図 1 元の入力画像とグリッドに区切り 1 カ所ずつマスクした画像
Fig. 1 Original Image and Masked Images along to Grids.

ず，あらかじめ何らかのセグメンテーションが行われていることを前提としている．我々の手法では，セグメンテーションが困難な，境界の曖昧な特徴なども探索可能である．

Residual Attention Network [16] や Attention Branch Network [17] のように，CNN による分類モデルから分岐させたネットワークにより，分類と同時に attention map を生成させるというアプローチもある．これも，正の寄与部分だけを扱っているという点は他のアプローチと変わらず，負の寄与に関しては考慮されていない．また，我々の手法は CNN などの分類モデルの仕組みに依存しない．

3. DiDA のアルゴリズム

我々の提案する DiDA は，分類器への入力と，それに対する出力値のみから解析画像を生成する．そのため，出力値がそれぞれの分類に対する確信度やその分類らしさを表すものであれば，機械学習によるものを含め分類の仕組みを問わず適用が可能である．以下，具体的な解析画像の生成手順を説明する．まず，グリッドサイズを設定して DiDA Grid Image を生成し，複数のグリッドサイズの DiDA Grid Image から，DiDA Mixed Image を生成する．

3.1 DiDA Grid Image

図 1 左上の画像を，分類および解析を行いたい画像 I_o とする．これに対して，同図右側の画像群のように，任意のサイズ（図の例では 5×5 個に分けるサイズ）のグリッドに区切り 1 カ所ずつマスク[*1]した画像を用意する．(i, j) の位置のグリッドをマスクした画像を M_{ij} とする．画像 I をある分類器 C で分類したときのカテゴリ k に対する確信度の出力の値を $C_k(I)$ と表す．出力値は個々の分類に対する出力そのままの値でも，Softmax などの正規化を行った後の値でもよい．

このとき，画像 I_o の (i, j) の位置のグリッドをマスクすることによるカテゴリ k に対する出力値の変化 d_{ij} は次のように表される．

[*1] ここでは黒 1 色の塗りつぶしを採用しているが，対象画像により他のマスクパターンを使うことが考えられる．

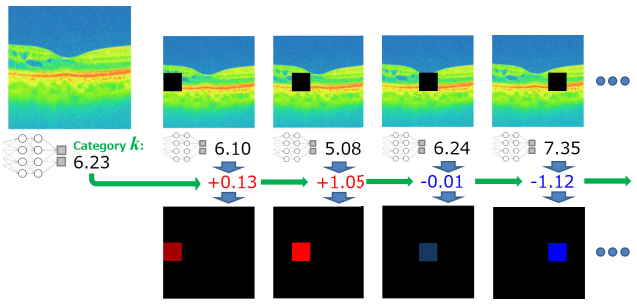

図 2 元画像を分類したときとマスクした画像を分類したときの出力値の差分を色と輝度にマップする
Fig. 2 Mapping Color and Intensity to Each Grid According to the Difference of Confidence Values of the Original Image and Masked Images.

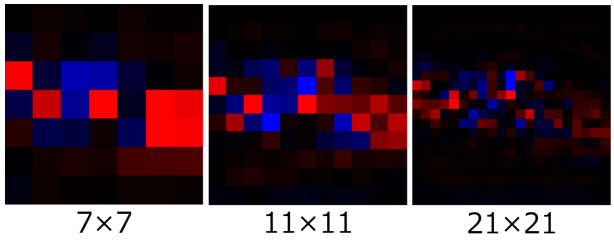

図 3 同じ入力画像，同じ分類モデルに対して生成された，グリッドサイズだけが違う 3 つの DiDA Grid Image（赤いグリッドは正の寄与部分，青いグリッドは負の寄与部分である）
Fig. 3 Three DiDA Grid Images in Different Grid Sizes to One Input Image (Red and Blue Grids Represent Positive and Negative Contributions).

$$d_{ij} = C_k(I_o) - C_k(M_{ij})$$

この値は，マスクした箇所にカテゴリ k に対し正に寄与する特徴が存在するならば正の値となり（すなわち，マスクした画像を入力した際の確信度は下がる），負に寄与する特徴が存在するならば負の値となる（マスクした画像を入力した際の確信度は上がる）はずである．

生成する画像を G とし，これを入力画像に対するグリッドと同じ数のグリッドに区切る．(i, j) の位置のグリッドは d_{ij} から，$|d_1| > |d_2|$ ならば $a(d_1) > a(d_2) \geq 0$ となるような性質を持つ関数 a により $a(d_{ij})$ で表される輝度（最大値を超える場合は最大値）の単色で塗りつぶす．色はたとえば，$d_{ij} \geq 0$ ならば赤，$d_{ij} < 0$ ならば青のように設定する（**図 2**）．このようにしてすべてのグリッドに彩色したものが DiDA Grid Image（以下，Grid Image と略す）である．

図 3 にあげたのはある同じ入力画像，同じ分類モデルに対して生成された，グリッドサイズだけが違う 3 つの Grid Image である．正の寄与部分，負の寄与部分がはっきりと分けられ可視化されているのが分かる．細かいグリッドサイズのほうが特徴の位置や形状を解像度良くとらえることができることが見てとれる．

しかし，粗い 7×7 のグリッドで正の寄与度が高い明るい表示色になっている画像右側，中央やや下の領域が，細

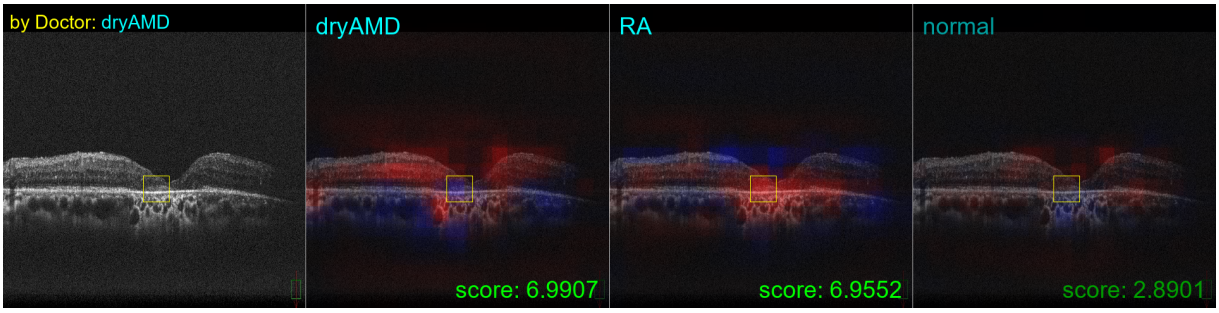

図 5 DiDA View による表示例 (1)
Fig. 5 DiDA Mixed Images for Fundus Images Obtained by OCT in DiDA View (1).

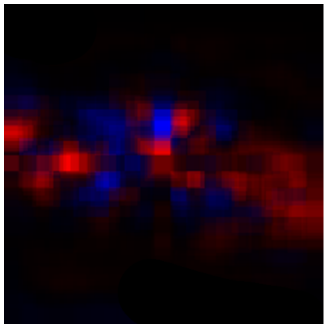

図 4 8 つのグリッドサイズの DiDA Grid Image から生成された DiDA Mixed Image
Fig. 4 DiDA Mixed Image: Combination of Eight Grid Images in Different Grid Sizes.

かい 21 × 21 のグリッドではすべて暗い色となっている．これは，細かなグリッドでは当該部分の正の寄与度を持つ画像中の特徴が一部しか隠されず，分類の確信度があまり下がらないためと解釈できる．すなわち，複数のサイズの Grid Image を比べることで，画像中の特徴のスケールを推定することができると考えられる．

3.2 DiDA Mixed Image

前述のように，Grid Image は提示できる情報がグリッドサイズに応じて変わる．そこで複数の Grid Image を統合し，1 枚の画像に情報をまとめて提示することを考える．その画像を，DiDA Mixed Image と呼ぶ（以下では Mixed Image と略す）．

まとめ方にはさまざまなバリエーションが考えられるが，ここでは Grid Image を重み付けしつつ重ね合わせる方法を採用する．すなわち，Grid Image G の座標 (x, y) のピクセル値を $p(G, x, y)$ と表すと n 枚の Grid Image G_i $(i = 1, 2, 3, \ldots, n)$ から重み w_i で生成される Mixed Image D の座標 (x, y) のピクセル値 $p(D, x, y)$ は

$$p(D, x, y) = \sum_{i=0}^{n} w_i \cdot p(G_i, x, y)$$

となる．

細かなグリッドのほうが重めになるようにして[*2]，7×7〜21 × 21 の 8 つの奇数サイズの Grid Image から生成された Mixed Image を図 4 に示す．重みの付け方は一例であり，対象や抽出すべき情報に合わせて調整することを想定している．

4. DiDA の適用例：眼底断層画像の分類

我々は DiDA を用いて，OCT で得られた眼底の断層画像を分類する機械学習モデルが妥当な分類を行っているかどうかを検証するためのアプリケーション DiDA View を試作した．対象とした機械学習モデルは，ニデック社の OCT 機器，RS-3000 Advance および RS-330 で撮影された画像を，32 種類の診断名と健常，合わせて 33 種に分類するものであり，垂直・水平断面あわせて 3,173 枚の教師画像を，横幅を 1/2 に縮め[*3] 224 × 224 ピクセルにフィットするようにスケーリングしたものを EfficientNet(B0) [18] に学習させたものである．学習済みのモデルは，与えられた画像（学習データと同様，224 × 224 ピクセルに調整されたもの）に対する各分類の確信度のスコアをそれぞれ出力する．793 枚のテスト画像を分類させた場合の，正答率（確信度が最も高くなるものが正答である率）は 72.1%，確信度上位 3 つ以内に正答が入る率は 88.9%である．

図 5 は DiDA View の画面例である．4 つの画像が表示されており，最左は元の画像および熟練した医師による診断名（この場合は Dry AMD（萎縮型加齢黄斑変性））である．その右に並ぶのは，元の画像の上に Mixed Image を半透明で重ね合わせた画像であり，機械学習モデルが出力した確信度の高い上位 3 つの分類（この場合は順に，Dry AMD，RA（網膜萎縮），normal（健常））に対応するものである．それぞれ右下部にスコアが表示されている．中央の小さな正方形は，画像中の同じ部分を比較しやすいよう，ユーザの操作したマウスカーソルの位置に合わせて表示される．

[*2] 粗い画像から順に，2 枚め以降を $1/(\mathrm{i}+1)+0.1$ の透明度で重ねている．全体の比率は約 4.4%：約 6.7%：約 8.5%：約 10.5%：約 13.0%：約 15.6%：約 18.8%：約 22.5%となる．
[*3] 眼底の OCT 画像は元々縦（奥行き）方向が強調されたスケールになっており，この比率でも医師は違和感なく読影が行える

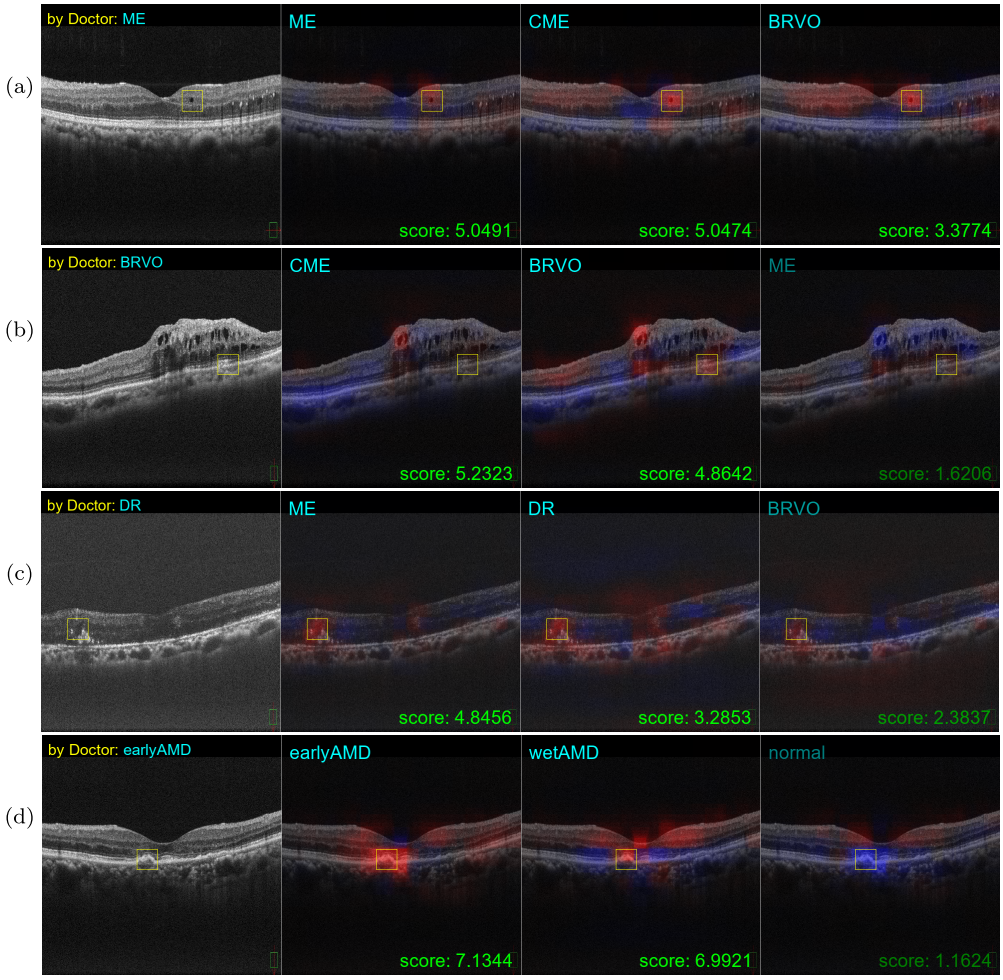

図 6 DiDA View による表示例 (2)

Fig. 6 DiDA Mixed Images for Fundus Images Obtained by OCT in DiDA View (2).

それぞれの分類に対し，赤（正の寄与部分）と青（負の寄与部分）が，共通する部分も違う部分もあるパターンで分布していることが分かる．図 5 の画像中の正方形で囲われた部分は，Dry AMD の分類には負に寄与しており，「Dry AMD らしくない」と機械学習モデルが判断していると解釈できる．RA に対してはその周辺を含め強めの赤，正の寄与の表示が出ており「RA の特徴的所見のように見える」ことを示している．normal に対しては暗めの青色となっている．「『健常』に対する負の寄与部分」とはすなわち「何らかの病変」と解釈することができる．

図 6 に他の 4 つの例を示す．図 5 の例も含め，機械学習の判断が複数の診断名に対して拮抗したスコアを出している症例を抽出した．眼底疾患に詳しくない者にとっても，注目すべきであろう箇所が明確に分かるので，医師のための診断補助だけでなく，患者への説明用途や，初学者に対する教育用途として応用することも考えられる．

5. 眼科専門医の見解との照合

前章で提示したような DiDA Mixed Image は，熟練した眼科専門医の読影によりラベルを付けたデータにより学習した機械学習モデルの出力から構築されている．学習により，同じラベルがつけられた画像に共通して存在する特徴が見出され，それに基づき分類を行っているはずであり，DiDA はそれを可視化しているはずである．しかし，医学的に見て妥当な情報が DiDA により示されているかどうかは慎重に検証する必要がある．

その端緒として，図 5，6 であげた例に関して，学習データのラベル付けを行った眼科の医師である筆者らの 1 人が検証した．その結果 DiDA で提示している情報は医師が判断する際のポイントとおおよそ合致していることが確認できた．ここではそのうち 3 つの症例について説明する．

5.1 萎縮型加齢黄斑変性（Dry AMD）の症例

図 5 の症例に対し医師は Dry AMD と判断し，機械学習は Dry AMD と RA のスコアが拮抗している．この画像（最左）について医師が Dry AMD と判断したポイントは，画像中央付近の下部の明るくなっている部分である．これは地図状萎縮と呼ばれる部位の網膜色素上皮（正方形部分の中央やや下を左右に横切る明るい線状の部分に相当，**図 7** 参照）の萎縮により OCT の観察光がその下の脈絡膜

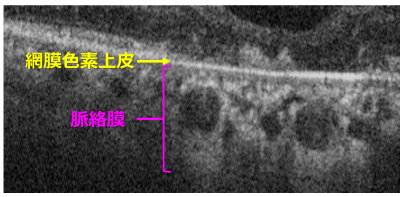

図 7 網膜色素上皮と脈絡膜
Fig. 7 Retinal Pigment Epithelium (RPE) and Choroid.

（図 7 参照）まで透過しやすくなっており，脈絡膜の輝度が高く写っているということである．Dry AMD の診断においては，この輝度が高くなる境界部位と，萎縮部位以外の網膜は正常に近い所見であるところを見ている．一方，RA の場合（右から 2 枚目），まさに網膜の萎縮による菲薄化を見ていると考えられる．

Dry AMD の画像では脈絡膜の明るくなっている部分の左側がやや青くなっている．この部分に関して医師は「意識していなかったが，Dry AMD であれば脈絡膜がもっと菲薄化している症例が多く，この症例の脈絡膜厚は例外的」と述べている．DiDA により無意識には見ているであろう部分を顕在化させている例と考えられる．

5.2 黄斑浮腫（ME）の症例

図 6 (a) は医師が ME と診断名をつけた症例である．ME は疾患名ではなく網膜の肥厚を示す異常所見の名称であり，浮腫の部分に内部が液体（黒色）の嚢胞様所見をともなうと CME（嚢胞様黄斑浮腫）と呼ばれる．医師の判断は，ME は認められるが嚢胞は顕著ではなく，疾患の特定もできない，というものである．機械学習では ME と CME で判断が割れ，次点で BRVO（網膜静脈分枝閉塞症）という疾患名を出力している．

正方形で囲われている，CME の画像（右から 2 枚目）で局所的に強く赤色で示されている黒い空洞が医師は明確でないと判断した嚢胞様腔である．ME の画像（左から 2 枚目）ではその周辺も含め赤色が示されている．これは，小さな嚢胞様所見よりも網膜全体の肥厚（浮腫）を見て ME と判断していると考えられる．これは単に「対象となるオブジェクトや領域を抽出する」のとは違う可視化が必要であることを示す例であるといえる．

最右の BRVO は，症状として CME をともなうことも多い．画像右側中央付近に薄く赤で示されている，輝度の高い小さないくつかの点は硬性白斑またはその前駆体を示唆する高輝度病変（hyper-reflective foci）と呼ばれるものであり，BRVO の症状として現れうるものである．また，BRVO は上下に分かれて分布する網膜静脈枝の閉塞であるため，上下（画像の左右）で浮腫の偏りがあることが特徴である．本画像は水平断であるが，画面左半分の網膜が正常であることを確認していることが赤色で示されている（図 6 (b) の症例の右から 2 枚目の画像も同様）．これらの所見を総合して機械学習は BRVO をあげていることが分かる．

5.3 加齢黄斑変性前駆病変（Early AMD）の症例

図 6 (d) は加齢黄斑変性（AMD）の前駆病変と分類される症例である．機械学習による Early AMD の判断の画像（左から 2 枚目）で顕著な赤い表示の部分（正方形で囲われた部分とその周囲）は網膜色素上皮の隆起が孤発性に存在し（凹凸周囲は異常がない），ドルーゼンと呼ばれる加齢性沈着物であるとして前駆病変と診断される所見である．

一方，Wet AMD（滲出型加齢黄斑変性）の特徴である脈絡膜新生血管も同様の網膜色素上皮の隆起を認めるため，右から 2 枚目の画像において赤い表示となっているが，その左右には青い領域が見られる．Wet AMD であれば隆起がより広範囲にわたるか，あるいは滲出性網膜剥離という症状をともなうことが多く，この青い領域はそれらが見られないことを示していると考えられる．最右の normal の判断は，異常所見に乏しい画像であることを示しているが，その中でも，網膜色素上皮の隆起は青色すなわち正常らしくない（異常）所見と判断されている．

6. 他手法との眼底断層画像分類での比較

DiDA を既存の他の手法のうち 3 種，RISE, Grad-CAM, Score-CAM との比較を行った．対象は前章と同じ OCT による眼底の断層画像からの診断名の分類であり，**図 8** がその結果の一部である．左端の列の画像の上位 3 つの分類結果に対する各手法による可視化結果をあげている．

CNN の畳み込み層の情報から画像を生成している Grad-CAM や Score-CAM は注目しているであろうポイントを含むおおよその領域しか可視化できていない．これらの手法はこの課題に適用するのは向いていないと判断できる．

RISE による画像は Grad-CAM や Score-CAM のものよりも詳細を可視化できており，DiDA による画像に最も近い．元の画像で黒い，ほとんど何も写っていない（判断に影響を与えない中立的な）部分が中間的な値になり，それよりも値が低くなる部分が相対的な「負の寄与」部分として把握できる場合も見て取れる．**図 9** に RISE による画像を，輝度 0〜255 の中間値 128 を境に赤と青に変換し DiDA 風の配色で表示したものをいくつかあげる．この「負の寄与」と解釈できる部分は必ずしも明確ではなく，また DiDA によるものとずれているものもある．それぞれが何を可視化しているかの分析は今後の課題である．

RISE の画像の解像度は固定されたグリッドのサイズに制限されるため，DiDA よりも解像度が粗い．RISE はグリッドをランダムにマスクした画像を多数用い確率的に全体をカバーするという方法であるため，解像度を上げるためにグリッドを細かくすると，カバーするのに十分なランダムパターンの数が指数的に増える．よって，RISE の手

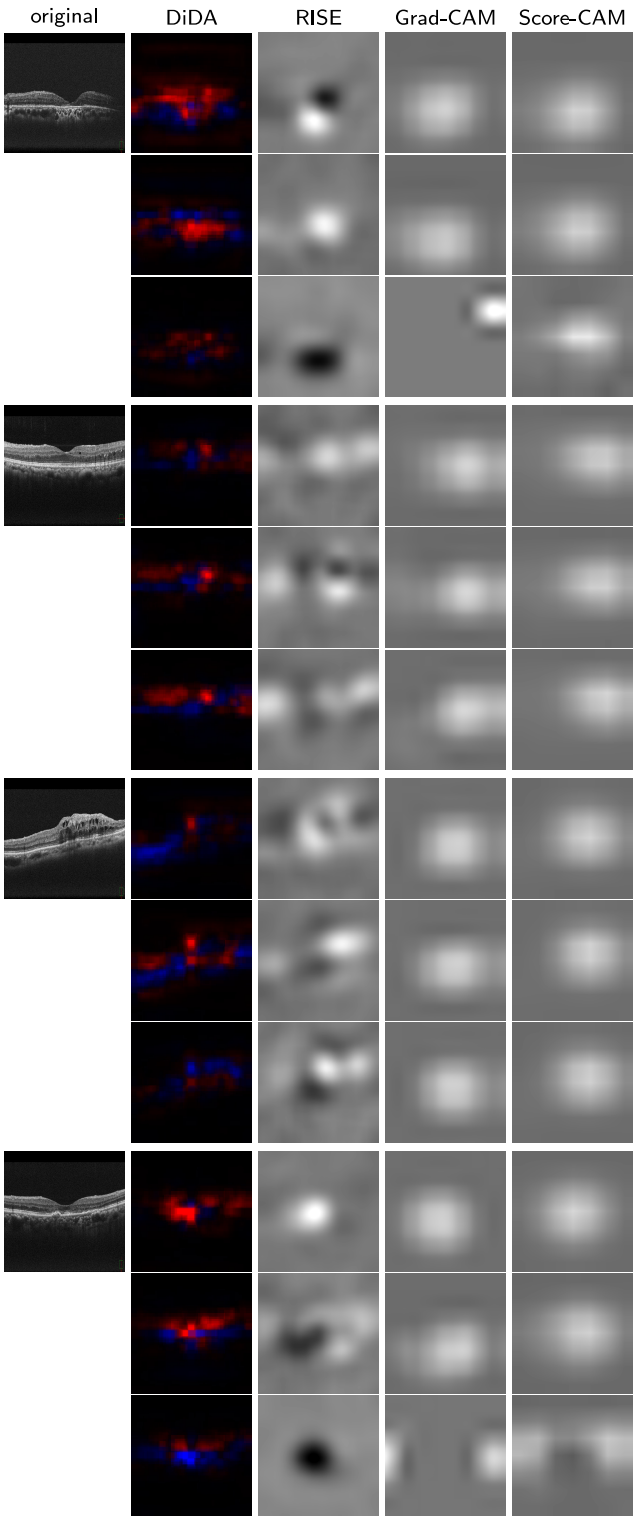

図 8 DiDA と他手法の比較
Fig. 8 Comparison of DiDA and Other Methods.

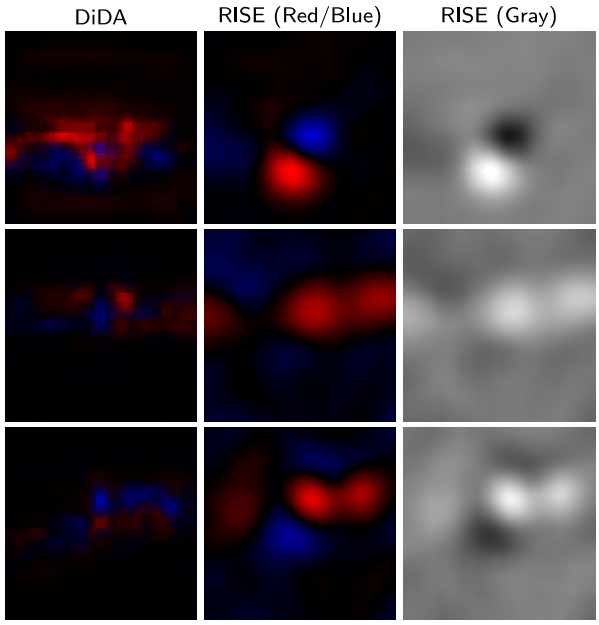

図 9 赤–黒–青の DiDA 風の表示にした RISE との比較
Fig. 9 Comparison of DiDA and RISE in DiDA-like Color Mapping.

法を用いて解像度を DiDA と同等に上げることは現実的ではない．

DiDA による画像は他の手法と比べて，元の画像の眼底領域とその構造が注目しているポイントとして全体的によく再現されている．疾患によっては，特定の部分以外に異常所見がないことが診断に重要な場合もあるが，DiDA はそれらを含め画像中の見るべきポイントを広く抽出できていると考えられる．正負どちらにも寄与しない中立的な領域が明確に黒で示されるのも DiDA の利点である．

これらから総合的に，DiDA の手法は眼底の断層画像から眼底疾患の種類を分類するような課題において，従来手法よりも適切に判断ポイントを可視化できていることが分かる．今後，より詳細な比較や検証を行ってゆく．

7. まとめ

本稿では，ある画像を分類するモデルに対し，その画像に微小な差異を加えた画像を入力した場合の出力の値の変化を画像化し提示する手法，DiDA を提案し，眼底断層像からの機械学習による眼底疾患の分類に応用した例を紹介した．正の寄与部分と負の寄与部分の双方を明確に区別して提示できることが大きな特徴であり，また，入力する画像と出力の関係だけを扱うため，分類モデルの種類によらず適用することができる汎用性の高い手法である．眼底断層像の分類に対し DiDA を適用した結果が，医師が判断する際のポイントを的確に示していることも確認した．また，既存の他手法と比べて，眼底断層像の分類においてより適切な可視化を行えていることも示した．

今後，DiDA により生成された画像が医学的に適切な情報を提示できているかなどをより詳細に検証し，有効性を確認する．DiDA の特性をより明らかにするために，この手法を他の分野の画像分類にも適用し，手法が有効に働く状況や条件の検討も行う．加えて，こうした情報を医師や患者などへの提示のしかたや，活用手法についても検討を行う．初学者の学習の補助としての応用も有望である．

機械学習をはじめとするAIによる診断などは「中身がブラックボックスだ」として不安がられることも多い．しかし，DiDAのような手法を通して判断のポイントなどを適切に可視化できれば，機械学習はむしろ，暗黙知を含む人間の持つ知識を外在化させ，つぶさに解析したり，他の人へ的確に伝えたりするための手段としても活用できると考えられる．我々は，そのような可能性を追求してゆく．

参考文献

[1] Kuwayama, S., Ayatsuka, Y., Yanagisono, D., Uta, T., Usui, H., Kato, A., Takase, N., Ogura, Y. and Yasukawa, T.: Automated Detection of Macular Diseases by Optical Coherence Tomography and Artificial Intelligence Machine Learning of Optical Coherence Tomography Images, *Journal of Ophthalmology*, Vol.2019 (online), DOI: 10.1155/2019/6319581 (2019).

[2] 綾塚祐二，雅樂隆基，安川　力，吉高淳夫：DiDA：機械学習による画像診断の判断ポイントを入力差分により解析・可視化する手法，インタラクション2021論文集，情報処理学会，pp.109–114 (1A02) (2021).

[3] Shrikumar, A., Greenside, P. and Kundaje, A.: Learning Important Features Through Propagating Activation Differences, arXiv:1704.02685 (2019).

[4] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A.: Learning Deep Features for Discriminative Localization, arXiv:1512.04150 (2015).

[5] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, *International Journal of Computer Vision*, Vol.128, No.2, p.336–359 (online), DOI: 10.1007/s11263-019-01228-7 (2019).

[6] Chattopadhay, A., Sarkar, A., Howlader, P. and Balasubramanian, V.N.: Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks, *2018 IEEE Winter Conference on Applications of Computer Vision* (*WACV*) (online), DOI: 10.1109/wacv.2018.00097 (2018).

[7] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P. and Hu, X.: Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks, arXiv:1910.01279 (2020).

[8] Zhang, X., Wei, Y., Feng, J., Yang, Y. and Huang, T.: Adversarial Complementary Learning for Weakly Supervised Object Localization, arXiv:1804.06962 (2018).

[9] Wei, Y., Feng, J., Liang, X., Cheng, M.-M., Zhao, Y. and Yan, S.: Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach, arXiv:1703.08448 (2018).

[10] Singh, K.K. and Lee, Y.J.: Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization, arXiv:1704.04232 (2017).

[11] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. and Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, *PLOS ONE*, Vol.10, No.7, pp.1–46 (online), DOI: 10.1371/journal.pone.0130140 (2015).

[12] Petsiuk, V., Das, A. and Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models, arXiv:1806.07421 (2018).

[13] Ribeiro, M.T., Singh, S. and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, arXiv:1602.04938 (2016).

[14] Lundberg, S.M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, *Proc. 31st International Conference on Neural Information Processing Systems*, *NIPS'17*, pp.4768–4777, Curran Associates Inc. (2017).

[15] Ribeiro, M.T., Singh, S. and Guestrin, C.: Anchors: High-Precision Model-Agnostic Explanations, *Proc. AAAI-18, the IAAI-18, and the EAAI-18*, pp.1527–1535, AAAI Press (2018).

[16] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. and Tang, X.: Residual Attention Network for Image Classification, arXiv:1704.06904 (2017).

[17] Fukui, H., Hirakawa, T., Yamashita, T. and Fujiyoshi, H.: Attention Branch Network: Learning of Attention Mechanism for Visual Explanation, *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*) (2019).

[18] Tan, M. and Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, *Proc. 36th International Conference on Machine Learning*, *PMLR*, Vol.97, pp.6105–6114 (2019).

綾塚 祐二　（正会員）

1993年東京大学理学部情報科学科卒業．1995年東京大学大学院理学系研究科情報科学専攻修士課程修了．1998年同博士課程中退，ソニー株式会社，株式会社ソニーコンピュータサイエンス研究所等を経て，現在，株式会社クレスコ技術研究所主席研究員．ユーザインタフェースや，機械学習を活用した医用画像の解析の研究に従事．日本ソフトウェア科学会，ヒューマンインタフェース学会，人工知能学会，IEEE，ACM各会員．

雅樂 隆基

1995年名古屋大学大学院理学研究科博士課程前期修了（専攻：宇宙理学），株式会社日立製作所中央研究所等を経て，現在，株式会社クレスコ技術研究所副所長．電子情報通信学会，日本メディカルAI学会各会員．

安川 力

1993 年京都大学医学部卒業．1994 年北野病院，2000 年京都大学大学院医学研究科視覚病態学，2004 年倉敷中央病院等を経て，2005 年より名古屋市立大学大学院医学研究科視覚科学．現在，名古屋市立大学大学院医学研究科視覚科学教授．日本眼科学会眼科専門医，PDT 認定，日本眼科学会評議員，日本網膜硝子体学会理事，日本眼科 AI 学会評議員，日本眼薬理学会評議員．

吉高 淳夫 （正会員）

1989 年広島大学工学部第 2 類（電気系）卒業．1991 年同大学大学院博士課程前期修了，1994 年同博士課程後期単位取得退学．現在，北陸先端科学技術大学院大学情報科学研究科教授．博士（工学）．マルチメディアデータ検索，感性情報等に基づいた動画像処理，映像を利用したインタラクティブシステムが主な研究分野．映像情報メディア学会，IEEE ComputerSociety 各会員．