F-002

胸部 X 線画像分類における判断根拠領域の可視化
Visualization of Decision Basis Areas in Chest X-ray Image Classification

深井 友貴 *　荒井 敏 *　長尾 智晴 *
Tomoki Fukai　Satoshi Arai　Tomoharu Nagao

1　はじめに

従来, 放射線画像診断において, 医師は医用画像を定性的に評価して診断を行ってきたが, 以下の 2 つの課題が指摘されている.

- 医師は膨大な量の医用画像に目を通さなければならないため, 医師にかかる負担が大きい [1]
- 診断結果が定性的評価に基づくことから, 同じ画像であっても医師によって, 診断の質にバラつきが生じる可能性がある [2]

このような課題を受けて, 近年, 人工知能（Artificial Intelligence: AI）を用いた医用画像診断支援システムに大きな注目が集まっている. その理由として, AI が大量の画像パターンを認識し, 画像特徴を定量的に評価することに優れていることが挙げられる. これにより, AI を用いた医用画像診断支援システムは「医師の負担軽減」と「診断の質におけるバラつきの解消」に繋がると期待されている.

しかし, AI を用いた医用画像診断支援システムの開発にはいくつかの課題が存在しており, その 1 つが、判断根拠の解釈可能性に欠けていることである. 医療現場では判断根拠を明確に説明することが求められるが, AI が導いた結果の導出過程や判断根拠を人間が理解することは, ブラックボックス性により困難である. これにより, システムの信頼性を担保することが難しいため, 大きな課題となっている. 解決策として, 近年, 導出過程や判断根拠を人間が理解できる形で提示する説明可能 AI（Explainable AI: XAI）が有力視されているが, 医師が求めるレベルでの判断根拠の説明には至っていないのが現状であり, さらなる発展が望まれている [1].

そのような中, 判断根拠を 2 次元マップとして可視化可能な画像分類モデルである Generative Contribution Mappings（GCM）[3] が提案されている。しかし, GCM の医用画像に対する検証は未だ十分に進んでいない. そこで本研究では, GCM を胸部 X 線画像に適用し, その有効性の検証を目的とする. 実験では「胸部 X 線画像の 4 クラス分類」に従来手法と GCM を適用し, その有効性を検証する. 具体的には, 従来手法と比較し, それぞれのクラスの ROC-AUC を定量的に評価することに加え, 「病変部位を判断根拠にすることができているか」を 2 次元マップから定性的に評価する.

2　関連研究

2.1　畳み込みニューラルネットワーク

畳み込みニューラルネットワーク （Convolutional Neural Network: CNN）[4] とは, 畳み込み（convolution）を用いた順伝播型ネットワークである. CNN は画像を入力とする様々な問題に適用可能であり, 画像分類の分野で広く利用されている.

CNN の特徴として, 畳み込み層（Convolution Layer）とプーリング層（Pooling Layer）の 2 つの層が挙げられる. 畳み込み層では画像の特徴抽出を行い, 画像特徴が反映されたマップ（特徴マップ）が出力される. また, プーリング層では情報の集約を行うことにより畳み込み層で抽出された特徴の次元数を削減する. この畳み込みとプーリングを繰り返すことで, 画像特徴を階層的に捉えて, 線形識別しやすい特徴マップを獲得することができる. これにより, CNN は画像を扱う問題に対して高い性能をもつ.

* 横浜国立大学
Yokohama National University

2.2 エンコーダ・デコーダモデル

エンコーダ・デコーダモデルとは, エンコーダによって入力 x を低次元な潜在表現 z に圧縮した後, デコーダによって z から x と同じサイズの出力 y を獲得するネットワークで, 画像データにも広く用いられている. エンコーダ・デコーダモデルの構造を図 1 に示す.

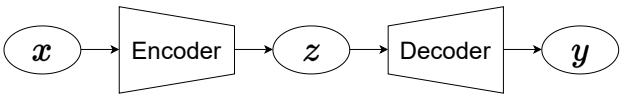


図 1: Encoder-Decoder モデル

エンコーダとして CNN を用いる場合, 特徴抽出の過程で空間サイズを縮小するため, 同じサイズの別画像を生成するには空間サイズの拡大が必要になる. 拡大処理にはアップサンプリング（upsampling）や転置畳み込み（transposed convolution）[5] がよく用いられる.

2.3 判断根拠の可視化手法

画像分類における判断根拠の可視化手法には, 主に以下の 3 つのアプローチが存在する.

1. ユニットの反応マップを生成する手法
2. 中間層の出力をそのまま可視化する手法
3. 注目クラスに関する物体の概略位置を示すマップを生成する手法

第 1 の「ユニットの反応マップを生成する手法」については, Zeiler ら [6] が転置畳み込みを用いて中間層のユニットの反応を可視化する手法を提案している. この手法は, ある入力画像を与えた場合の注目ユニットの反応を, 転置畳み込みの反復によって入力方向に逆伝播させることで, ユニットの反応マップを生成するというものである.

第 2 の「中間層の出力をそのまま可視化する手法」については, Lin ら [7] が特徴マップを各クラスの信頼度マップ（categorical confidence maps）として解釈可能であることを示し, 特徴マップをそのまま可視化マップとして利用することを提案している.

第 3 の「注目クラスに関する物体の概略位置を示すマップを生成する手法」については, Zhou ら [8] が, ある画像を学習済みの画像分類モデルに入力した場合の畳み込み層の出力（特徴マップ）を, 注目クラスに対応する全結合層の重みを用いて重み付き加算することで, 注目クラスに関する物体の概略位置を示すマップを生成し可視化する手法（Class Activation Mapping: CAM）を提案している. さらに, CAM を発展させた手法として, Selvaraju ら [9] は Grad-CAM を提案している. これは全結合層における注目クラスの出力を特徴マップで偏微分することで, CAM と同様に注目クラスに関する物体の概略位置を示すマップを生成し可視化するものである.

しかし, これまで提案された可視化手法には以下の 2 つの課題が存在することが指摘されている [3].

- 可視化マップの空間解像度が低下する
- 可視化結果が分類結果と直接対応していない

次節では, これら 2 つの課題を解決した手法である Generative Contribution Mappings（GCM）について述べる.

2.4 Generative Contribution Mappings

前節で述べた課題を踏まえ, 荒井ら [3] は, Generative Contribution Mappings（GCM）という新しい説明可能な画像分類モデルを提案している. GCM は, CNN と同程度の分類精度をもち, 「分類の際にどこに注目したか」を示す可視化マップを生成することができる.

GCM の基本構成を図 2 に示す. GCM はエンコーダ・デコーダモデルをベースにしており, 大きな特徴としては以下の 4 点が挙げられる.

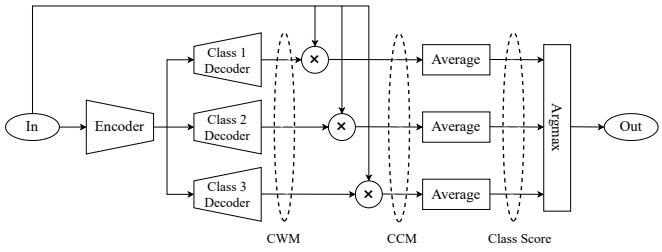


図 2: GCM の基本構成（分類クラス数=3 の場合）

1. エンコーダに続く処理として, 分類クラスごとにデコーダを有する.
2. デコーダの出力と入力画像との直接的な乗算経路を有する.
3. 2 の乗算後, パラメータを含まない単純な平均処理を用いてクラススコアを算出する.
4. GCM 全体としては, 重みを動的に生成しながら入

力画像との内積演算を行う処理に相当する．

入力画像はエンコーダによって次元数任意の特徴量に変換された後，デコーダによって元と同じサイズのマップに再構成される．このマップは入力画像の各位置が注目クラスに関してどの程度そのクラスらしいかを表す空間的な重みマップであり，荒井らはこれを Class Weight Map（CWM）と定義している．その後，入力画像と CWM を乗算することで Class Contribution Map（CCM）という新たな可視化マップを得る．CWM と CCM の例を図 3 に示す．

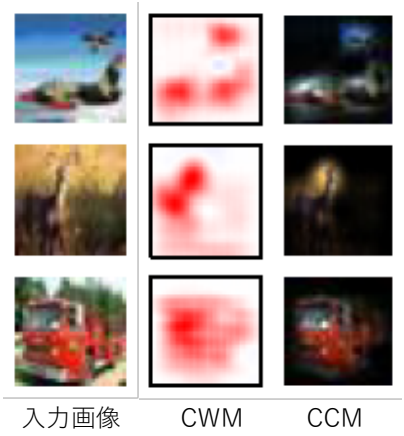


図 3: CWM と CCM[3]

CCM は入力画像からの情報と CWM からの情報の両方を合わせ持ち，入力画像のどの部位が注目クラスらしいかという情報を直感的に理解することが可能である．

GCM において注目すべきは，前節で述べた 2 つの課題を同時に解決する点である．まず，デコーダを用いて入力画像と同サイズの CWM を生成し，入力画像と CWM の乗算により，可視化マップである CCM を生成することで，第 1 の課題である「可視化マップの空間解像度の低下」を防いでいる．また，CCM を空間およびチャネルのすべての軸に関して平均したものをクラススコアとし，このクラススコアが最大となるクラスに分類することで，可視化結果と分類結果を直接的に対応させ，第 2 の課題を解決している．

しかし GCM は，一般物体画像に対しては有効性が確認されているが，医用画像に対する有効性の検証は未だ十分に進んでいない．小林ら [10] によって，マンモグラフィ画像を対象とした検証が行われているが，胸部 X 線画像に対する有効性は未検証である．

3 提案手法

3.1 概要

本研究で使用する GCM の構成を図 4 に示す．本研究では，エンコーダとして，ImageNet で事前学習済みの EfficientNet[11] を使用し，デコーダのみを学習させる．

EfficientNet は従来のモデルに比べて，少ないパラメータ数で高い精度を達成しているモデルである．本研究では事前学習済みのモデルをエンコーダとして使用することで，少ない学習枚数であっても過学習を起こさず学習できることが期待される．

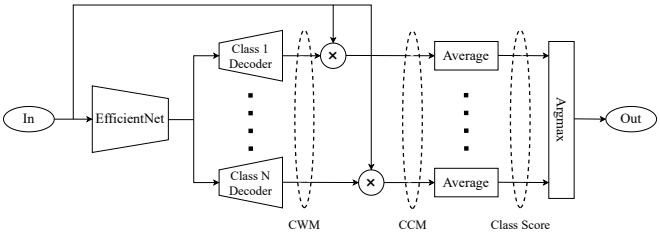
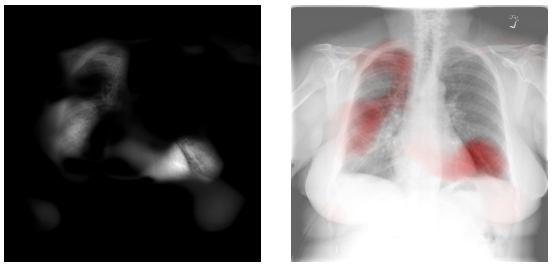


図 4: 本研究で用いた GCM の概要

3.2 可視化マップ

胸部 X 線画像において CCM を可視化マップとして使用する場合，判断根拠が視覚的に理解しにくいという課題があった (図 5(a))．そこで本研究では，「CWM をカラー化し，入力画像にオーバーレイ表示したもの」を可視化マップとして使用する (図 5(b))．これにより，「肺のどの部位が判断根拠となっているのか」を医師がより理解しやすくなることが期待される．



(a) CCM　　**(b)** オーバーレイ表示

図 5: 可視化マップの例

4 実験

4.1 概要

本実験では，胸部 X 線画像の 4 クラス分類に GCM を適用した．データセットは ChestX-ray8[12] を使用

し, 対象クラスは No Finding(正常), Cardiomegaly(心肥大), Pneumothorax(気胸), Effusion(胸水) とした. また, 評価指標として ROC-AUC を使用した. ROC-AUC は分類問題においてよく使用される評価指標であり, 値が 1 に近いほど優れた分類性能をもつことを示す. さらに比較手法としては, 同データセットにおいて SOTA を達成している手法である 121 層の DenseNet[13, 14] を採用した.

GCM の有効性を検証するため, DenseNet と比較し, 各クラスの ROC-AUC を定量的に評価した. また, 「病変部位を判断根拠にすることができているか」を可視化マップから定性的に評価した.

4.2 実験設定

4.2.1 データセット

学習データについては, ラベルが単一の画像のみを使用し, クラス間のデータの不均衡性を解消するため, 学習可能枚数が最小である心肥大の枚数 (963 枚) に揃えた. また, 疾患画像のテストデータについては, 病変部位の位置情報が提供されている画像を使用した. データの内訳を表 1 に示す.

表 1: データ内訳

	train	test
正常	963	100
心肥大	963	130
気胸	963	83
胸水	963	98

実験にあたっては内容の異なる 3 つのサブセットを, ランダム選択により作成し, 結果は「3 パターンの ROC-AUC の加算平均」とした. また, GCM による可視化に使用する重みは, 3 パターンの中で平均 AUC(各クラスの ROC-AUC の加算平均) が最大のものを採用した.

4.2.2 学習条件

本実験における学習条件を表 2 に示す. 入力サイズは $256 \times 256 \times 3$(空間サイズ: 256×256, チャンネル数: 3) とし, 最適化手法については, GCM では確率的勾配降下法 (Stochastic Gradient Descent: SGD), DenseNet では先行研究 [14] に倣い, Adam[15] を使用した. さらに, 初期学習率は実験的に決定し, GCM では 1e-1, DenseNet では 1e-5 とした.

表 2: 学習条件

	GCM	DenseNet
入力サイズ	$256 \times 256 \times 3$	
最適化手法	SGD	Adam
損失関数	Cross Entropy	
初期学習率	1e-1	1e-5
バッチサイズ	32	
エポック数	100	

4.2.3 モデル構造

ここでは, 本実験で使用した GCM のモデル構造について説明する. まず, 本実験における GCM のエンコーダ (事前学習済みの EfficientNet-B0 を部分的に使用) の構造を表 3 に示す. 表中の output の次元は「空間サイズ×チャンネル数」を表している . エンコーダでは, 畳み込みにより, $256 \times 256 \times 3$ の胸部 X 線画像を低次元に圧縮し, 画像特徴の抽出を行っている.

次に, 本実験における GCM のデコーダの構造を表 4 に示す. デコーダでは, エンコーダによって抽出された画像特徴から, Upsampling と畳み込みによって, 入力画像と同じサイズのマップを生成している. Upsampling においては, 双線形補間を用いて画像サイズを 2 倍に拡大している.

表 3: GCM におけるエンコーダの構造

	block	output
Encoder	Input	$256 \times 256 \times 3$
	Conv	$128 \times 128 \times 32$
	MBConv1	$128 \times 128 \times 16$
	MBConv6	$64 \times 64 \times 24$
	MBConv6	$64 \times 64 \times 24$
	MBConv6	$32 \times 32 \times 40$
	MBConv6	$32 \times 32 \times 40$
	MBConv6	$32 \times 32 \times 80$
	MBConv6	$32 \times 32 \times 80$
	MBConv6	$32 \times 32 \times 80$
	MBConv6	$16 \times 16 \times 112$
	MBConv6	$16 \times 16 \times 112$
	MBConv6	$16 \times 16 \times 112$

Conv: Convolution

MBConv: Mobile Inverted Bottleneck Convolution[16]

MBConv の数字は拡張率を表す

表 4: GCM におけるデコーダの構造

	block	output
Decoder	Input	$16 \times 16 \times 112$
	Conv	$16 \times 16 \times 112$
	Upsampling	$32 \times 32 \times 112$
	SepConv	$32 \times 32 \times 512$
	Upsampling	$64 \times 64 \times 512$
	SepConv	$64 \times 64 \times 256$
	Upsampling	$128 \times 128 \times 256$
	SepConv	$128 \times 128 \times 128$
	Upsampling	$256 \times 256 \times 128$
	SepConv	$256 \times 256 \times 64$
	Conv	$256 \times 256 \times 4$

Conv: Convolution

SepConv: Separable Convolution[17]

Conv 系の block は BatchNorm と ReLU を含む

5 結果と考察

5.1 ROC-AUC

実験の結果得られた ROC-AUC の値を表 5 に示す.

表 5: ROC-AUC

	GCM	DenseNet
正常	**0.860**	0.839
心肥大	0.878	**0.880**
気胸	**0.876**	0.874
胸水	**0.803**	0.795

考察

GCM はいずれのクラスにおいても, ROC-AUC が 0.8 を上回り, 同データセットの SOTA モデルである DenseNet に匹敵する ROC-AUC を記録した. このことから, 胸部 X 線画像の 4 クラス分類に対して, GCM は有効であると考えられる.

5.2 可視化マップ

GCM によって生成された可視化マップの例を次ページに示す. ここでは, 青色のバウンディングボックスで囲まれた領域が正解の病変部位を表しており, 赤色で着色されている領域が GCM による判断根拠領域を表している. 可視化マップの評価は非専門家が行い, バウンディングボックス内に赤色が明瞭に確認できる場合,「病変部位に注目できている」と判断した. 心肥大・気胸・胸水の図については, 左 3 枚が「病変部位に注目できている」と判断した例, 最右の 1 枚が「病変部位に注目できていない」と判断した例である.

クラス別の評価は以下のようになった.

- 心肥大 (図 6) においては, テストデータ 130 枚中 126 枚が病変部位に注目できていた.
- 気胸 (図 7) においては, テストデータ 83 枚中 59 枚が病変部位に注目できていた. 半数以上の画像が病変部位に注目できていたものの, 心肥大と比較すると, 過検出や見落としが多く見られた.
- 胸水 (図 8) においては, テストデータ 98 枚中 76 枚が病変部位に注目できていた. 気胸と同様, 半数以上の画像が病変部位に注目できていたものの, 心肥大と比較すると, 過検出や見落としが多く見られた.
- 正常画像 (図 9) においては, 肺全体が判断根拠領域となっている傾向が見られた.

考察

まず心肥大については, ほとんどの画像が病変部位に注目できていた. これは, 心肥大の特性上, 病変部位が心臓に限定されていることに加え, 病変の大きさも画像によってほとんど差がないことから, 病変が画像特徴として認識されやすかったためだと考えられる.

一方, 気胸と胸水については, 心肥大と比較すると, 過検出や見落としが多く見られた. これは, 病変の位置や大きさにバラつきがあり, 病変を画像特徴として捉えることがより困難だったためだと考えられる.

最後に, 正常画像については, 肺全体が判断根拠領域となっている傾向が見られた. このことから, GCM は肺全体に注目して異常が無いと判断した上で, 正常に分類していると考えられる.

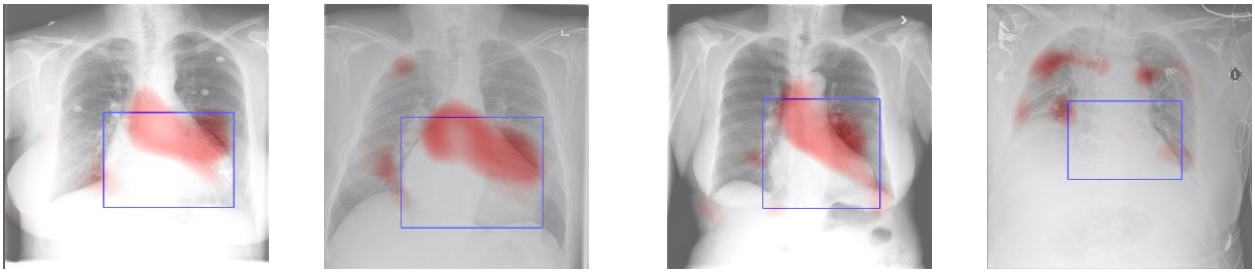


図 6: 心肥大の可視化マップ

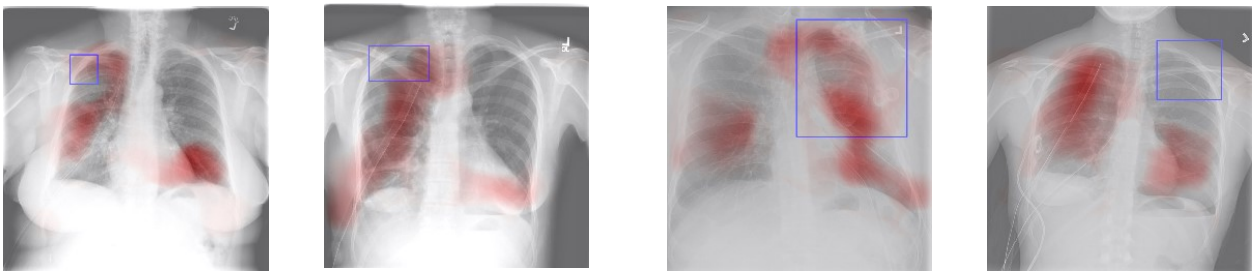


図 7: 気胸の可視化マップ

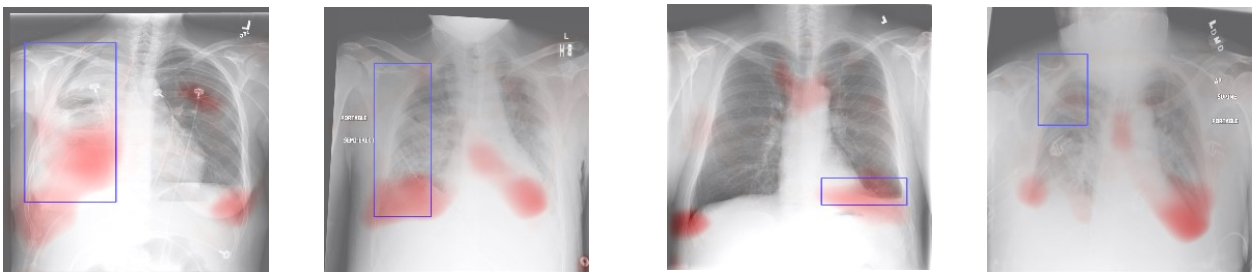


図 8: 胸水の可視化マップ

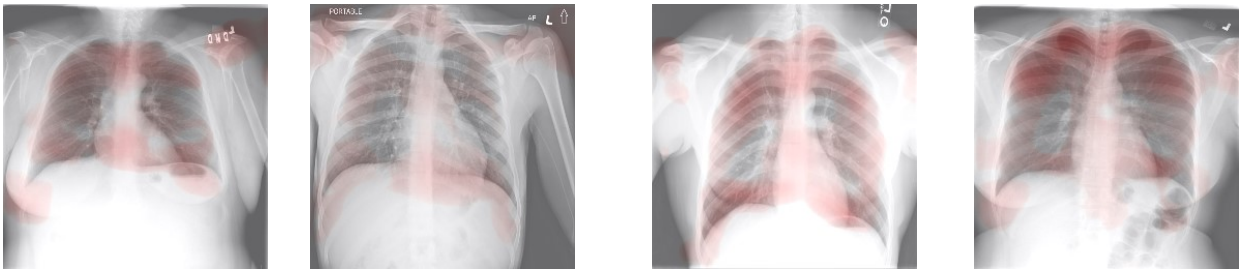


図 9: 正常の可視化マップ

6 まとめ

本研究では, 胸部 X 線画像分類に GCM を適用することにより, 胸部 X 線画像に対する GCM の有効性を検証した. また, 医師がより理解しやすい判断根拠領域の可視化マップを提案した.

実験の結果, GCM は, いずれのクラスにおいても 0.8 を上回る ROC-AUC を記録し, 胸部 X 線画像の 4 クラス分類に対する GCM の有効性が確認された. また, 可視化マップについては, 疾患によって差が見られたものの, GCM が病変部位を判断根拠として, 胸部 X 線画像を分類できる可能性があることを示した.

7 今後の課題

7.1 分類性能のさらなる向上

本研究では, 胸部 X 線画像の 4 クラス分類に対する GCM の有効性を確認したが, 実用性を考慮すると, 高い分類精度を維持しながら, 分類対象となるクラス数をさらに増加させることが望まれる.

現在 GCM は CNN をベースとしているが, 将来的には Vision Transformer[18] のような, より新しいモデルを組み込むことにより, 分類性能が向上することが期待される.

7.2 可視化マップの精度向上

本研究では, GCM が病変部位を判断根拠に胸部 X 線画像を分類できる可能性があることを示したが, 疾患によっては過検出や見落としが多く見られるため, 改善が望まれる. それぞれの疾患の画像特徴を解析し, モデルに反映させることに加え, 医学的知見を考慮した改良を行うことで, 可視化マップの精度が向上することが期待される.

謝辞

医学的知見に基づく貴重なご意見を賜りました, 横浜市立大学附属病院の酒井和也先生, 西井基継先生に厚く御礼申し上げます.

参考文献

[1] 杉野貴明, 中島義和. Ai を用いた医用画像診断支援. 電気学会誌, Vol. 143, No. 4, pp. 208–211, 2023.

[2] 小林和馬. 医用画像診断における深層学習モデルの開発—実臨床応用を志向した深層学習モデルの開発における課題と解決—. 人工知能, Vol. 35, No. 4, pp. 509–514, 2020.

[3] 荒井敏, 長尾智晴. 畳み込みニューラルネットワークを用いた画像分類タスクの直感的可視化方法. 情報処理学会論文誌数理モデル化と応用 (TOM), Vol. 10, No. 2, pp. 1–13, 2017.

[4] Shih-Chung B Lo, Heang-Ping Chan, Jyh-Shyan Lin, Huai Li, Matthew T Freedman, and Seong K Mun. Artificial convolution neural network for medical image pattern recognition. *Neural networks*, Vol. 8, No. 7-8, pp. 1201–1214, 1995.

[5] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. *arXiv preprint arXiv:1505.04366*, 2015.

[6] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

[7] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[8] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929. IEEE, 2016.

[9] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[10] Tatsuaki Kobayashi, Takafumi Haraguchi, and Tomoharu Nagao. Classifying presence or absence of calcifications on mammography using generative contribution mapping. *Radiological Physics and Technology*, Vol. 15, No. 4, pp. 340–348, 2022.

[11] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

[12] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[14] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pp. 232–243. World Scientific, 2020.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[17] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.