令和 6 年度

修士学位論文

敵対的手法による
医用画像分類モデルの説明性に関する研究

A Study on Explainability of Medical Image Classification Models Using Adversarial Techniques

1275107　　中嶋大雅

指導教員　　吉田真一

2025 年 2 月 28 日

高知工科大学大学院 工学研究科 基盤工学専攻

情報学コース

要 旨

敵対的手法による
医用画像分類モデルの説明性に関する研究

中嶋大雅

近年, Computer-Aided Diagnosis(CAD) に注目が集まっており, 中でも Convolutional Neural Network(CNN) を用いた画像認識による診断が広く採用されており, 病気を高精度に識別することが出来ている. しかし, 患者への診断結果の説明が重要である医学において, ニューラルネットのブラックボックス性が課題となり, 様々な説明手法の活用が試みられているものの十分な解決にはなっていない. これまでに CNN の分類過程の説明には学習済みのモデルから得られる特徴マップやパラメータの重み, 勾配を用いて CNN の分類に寄与する領域を特定する Class Activation Map(CAM) ベースの手法が用いられてきた. しかし, CAM 等の手法では CNN の分類に寄与する詳細な特徴を得ることができないという課題があり, 分類に寄与する領域とその領域内の形状やパターンの違いなどを獲得出来るような新たな説明性手法が必要になると考えられる.

そこで本研究では, 分類モデルの判定を反転させることのできる敵対的サンプルを活用して, 分類に寄与する領域とその領域内のパターンの違いを獲得する新たな説明性手法を提案する. 本手法では, 敵対的サンプルの生成に Projected Gradient Descent(PGD) を用い, PGD から得られる摂動から CNN の分類に寄与する領域やその形状のパターンを特定することで, 説明性の向上を目指す. 提案手法の有効性や信頼性を評価するため, 既存の説明性手法である Grad-CAM, Guided Backpropagation, CycleGAN との比較を行う. 実験には, 塵肺および心肥大のラベル付き胸部 X 線画像を対象とした CNN 分類モデルを用いる.

塵肺データセット, 心肥大データセットに対して実験を行った結果, 症状に沿った分類に寄

与する領域と領域内の形状のパターンの違いが得られた．また, 既存の手法と比較した結果から, 敵対的サンプルから得られる特徴が分類に寄与していることを示す根拠の信頼性も高いことが確認された.

キーワード　CAD, 畳み込みニューラルネットワーク (CNN), 敵対的サンプル, 説明可能性 AI

Abstract

A Study on Explainability of Medical Image Classification Models Using Adversarial Techniques

NAKAJIMA, Taiga

In recent years, computer-aided diagnosis (CAD) has attracted significant attention, with diagnostic methods based on image recognition using convolutional neural networks (CNN) being widely adopted. These methods have achieved high accuracy in disease classification. However, in the medical field, where explaining diagnostic results to patients is crucial, the black-box nature of neural networks remains a challenge. Although various explainability methods have been explored, a definitive solution has yet to be established.

Recently, class activation map (CAM)-based methods, which use feature maps, weight parameters, and gradients obtained from trained models to identify regions contributing to CNN classification, have been employed to explain the classification process of CNN. However, CAM-based methods have limitations in capturing detailed features that contribute to classification. Thus, there is a need for novel explainability methods that can capture not only the regions contributing to classification but also the differences in patterns and shapes within those regions.

In this study, we propose a new explainability method that leverages adversarial samples capable of reversing classification outcomes to identify both the contributing regions and the pattern variations within those regions. Our method utilizes projected gradient descent (PGD) to generate adversarial samples. By analyzing the perturbations

obtained from PGD, we aim to identify the patterns and shapes of regions contributing to CNN classification, thereby improving explainability. To evaluate the effectiveness and reliability of the proposed method, we compare it with existing explainability methods, including Grad-CAM, guided backpropagation, and CycleGAN.

Experiments were conducted using labeled chest X-ray datasets for pneumoconiosis and cardiomegaly classification. The results demonstrate that our method successfully captures both the classification-contributing regions and the differences in shape patterns within these regions, consistent with the symptoms. Furthermore, comparisons with existing methods confirm that the features obtained from adversarial samples provide highly reliable evidence of their contribution to classification.

key words Computer Aided Diagnosis(CAD), Convolutional Neural Network(CNN), Adversarial Example, Explainable Artificial Intelligence(XAI)

目次

図目次

表目次

第 1 章

序論

近年, 医師不足や患者数の増加などの要因から医者の負担が増加している. この問題を解決するために AI を用いた医用画像診断支援システム Computer Aided Diagnosis(CAD) に注目が集まっている. CAD とは医用画像に対して, コンピュータで解析を行い, 得られた結果を第 2 の意見として利用することで医師の診断の支援を行うシステムのことである. CAD にはさまざまな種類が存在するが, 中でも Convolutional Neural Network(CNN) を用いた画像認識による診断が現在広く行われている [1]. しかし, ニューラルネットのブラックボックス性が課題となり, 様々な説明手法の活用が試みられているものの十分な解決にはなっていない. 現在の CNN の分類過程の解析には CNN の分類に寄与する領域を特定する Class Activation Map(CAM)[2] ベースの手法や Guided Backpropagation[**?**] が用いられてきた. しかし CAM や Guided Backpropagation で解析することができるのは分類に寄与する領域のみであり, それだけでは結果に対する十分な説明にはならないため, 領域内の形状やパターンの違いも必要になると考えられる. 他にも, 筒井・吉田の研究 [3] では Generative Adversarial Network(GAN)[6] の一種である CycleGAN[7] を用いた説明性手法を提案していた. しかし, CycleGAN を用いた手法は心肥大などの心肥大などの分かりやすい疾患画像に対しては説明性が得られていたが, 塵肺といった特定の病気に対して有効的な効果を得ることが出来ないという課題があった. そこで本研究では, 敵対的サンプル [9] を用いた説明性手法を提案する. 敵対的サンプルとは機械学習モデルに誤った予測をさせるために、意図的に小さな摂動 (ノイズ) を持たせた画像などのことである. 本研究では, 敵対的サンプルを加えた画像を分類モデルで識別すると誤分類が発生するという事象に着目して敵対的サンプルから得られる特徴が CNN の分類に直接関与していると見なすこと

で, 敵対的サンプルから分類に寄与する領域とその領域内の形状やパターンの違いの獲得を目指す. また, 敵対的サンプルから得られる特徴パターンを活用し, Grad-CAM や Guided Backpropagation、CycleGAN を用いた手法との比較を通じて, 敵対的サンプルを用いたアプローチが説明可能性の向上に寄与するかどうかを検証する. 本研究ではこれらの検証のために, 肺野領域を抽出した塵肺ラベル付きの 2 次元胸部 X 線画像および心肥大ラベル付きの 2 次元胸部 X 線画像を対象に二値分類を学習させた CNN を用いる.

第 2 章

関連技術

本項では本研究で用いた技術やそれに関連する技術について述べる.

2.1 畳み込みニューラルネットワーク (Convolutional Neural Network)

畳み込みニューラルネットワーク (Convolutional Neural Network: CNN) は, いくつもの深い層を持ったニューラルネットワークであり, 主に画像認識の分野において優れた効果を発揮することが知られたいる. CNN は, 畳み込み層とプーリング層が組み合わさって構成されており, 畳み込み層では画像から特徴量を抽出して特徴マップを生成, プーリング層では生成された特徴マップから最大値や平均値を取ることでより重要な値を抽出する. 本研究では, CNN の代表的なモデルである ResNet を用いて画像の生成を行う.

2.2 Class Activation Mapping(CAM)

CAM をベースとした説明性手法では, 学習済みモデルの最終畳み込み層の特徴マップと全結合層の重みを使用して, 特定のクラスに寄与する領域を可視化する手法である. 制約として, 全結合層が Global Average Pooling(GAP) と結びついている必要がある.

また, CAM ベースの手法から得られる顕著性マップは, 特定のクラスに寄与する領域のみであるため, 具体的なクラス間の違いを得ることは困難であり, 説明可能性における課題として挙げられる. 本研究では CAM ベースの手法として Grad-CAM を使用する.

2.3 Grad-CAM

Grad-CAM は特定のクラスの損失に対する勾配を用いて, 特定のクラスに対する特徴マップの重要度を計算して可視化する手法である. 従来の CAM ベースの手法では GAP を含む特定の CNN にしか適用できなかったのに対して, Grad-CAM は勾配情報を活用するため, GAP を必要とせず, どんな CNN にも適用可能である. また, 顕著性マップの生成過程で GAP を介さなくてもいいため, 局所的な重要領域を保持することが可能である. その一方で, エッジや細かいピクセル情報などは得ることができない. Grad-CAM の計算手順を以下に示す.

1. **順伝搬**

   画像 X を CNN に入力して, 出力クラスのスコアを取得. また最終畳み込み層の特徴マップ A^k を取得.

2. **勾配計算**

   手順 1 から得られたスコアから特定のクラス c に対する損失 L^c を取得. 最終畳み込み層の特徴マップ A^k に対する勾配を計算 (式 (2.1)).

3. **重要度の計算**

   特徴マップ A^k のチャンネルごとの重要度 (重み)w_k^c を計算 (式 (2.2)). ここで Z は特徴マップのサイズ ($Z = Height \times Width$).

4. **顕著性マップの生成**

   特徴マップを重み付けして結合する. また, ReLU を用いて正の値のみを抽出し, 分類に寄与しない領域を除外することで, 顕著性マップ $L^c_{\text{Grad-CAM}}$ を取得 (式 (2.3)).

$$\frac{\partial L^c}{\partial A^k} \tag{2.1}$$

$$w_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial L^c}{\partial A^k_{i,j}} \tag{2.2}$$

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k w^c_k A^k\right) \tag{2.3}$$

2.4 Guided Backpropagation

Guided Backpropagation は勾配ベースの可視化手法であり, 特定のクラスに対する画像の重要なピクセルをハイライトする技術で, 通常の Backpropagation（誤差逆伝播法）を改良し, 視覚的に解釈しやすい勾配マップを生成することができる. CAM ベースの手法が分類に寄与する領域を可視化するのに対して, Guided Backpropagation は詳細なエッジやテクスチャといったピクセルレベルの特徴を可視化することができる. 一方で Grad-CAM のように全体的な領域を可視化することができず, 局所的な情報が強調される. また, Guided Backpropagatioin と Grad-CAM を組み合わせた Guided Grad-CAM という手法も存在する. 以下に Guided Backpropagation の計算手順を示す.

1. **順伝搬**

   入力画像をネットワークに通して出力を取得.

2. **勾配計算**

   通常の逆伝搬と同じように勾配を計算するが, Guided Backpropagation では ReLU などの非線形関数を通る際に, 勾配の符号が正の時のみ逆伝播を行う. 具体的には, 逆伝播中に ReLU 関数を通る際に次のように処理する.

   - 勾配が正であれば, その勾配をそのまま伝播させる.
   - 勾配が負であれば, その勾配をゼロにする.

   ReLU 関数を式 (2.4), 逆伝搬における勾配の計算式を式 (2.5) に示す. ここで, $\frac{\partial L}{\partial y}$ は CNN の出力に関する損失関数の勾配を表す.

3. **特徴マップの出力**

   CNN の中間層 (本研究では最終畳み込み層) の勾配マップを出力.

$$y = \mathrm{ReLU}(x) = \max(0, x) \tag{2.4}$$

$$\frac{\partial y}{\partial x} = \begin{cases} 1 & \text{if } \frac{\partial L}{\partial y} > 0 \\ 0 & \text{if } \frac{\partial L}{\partial y} \leq 0 \end{cases} \tag{2.5}$$

2.5 Generative Adversarial Network(GAN)

GAN[6] は Goodfellow らが提案した機械学習モデルであり, 生成モデル G と識別モデル D の 2 つの CNN モデルから構成されている. 生成モデル G は入力ノイズ z に基づいて元のデータ x と区別がつかないようなデータの生成が可能になるように学習する. GAN の損失関数は Zhu らの研究 [7] では Adversarial Loss と呼ばれており, 式 (2.6) で表される. ここで式 (2.6) の第 1 項は D が元データである x を正しく識別するために用いられる. 式 (2.6) の第 2 項は元のデータと G によって生成されたデータを識別するために用いられる. この G と D の関係はよく通貨の偽造者と偽装通貨を取り締まる警察に例えられる. 偽造者は警察を騙せるような偽札を作成し, 警察は偽札と本物の通貨を見分けることができるよう努力する. このような敵対的な関係によって偽造者の作成する偽装通貨が段々と本物に近づいていく. この関係性が GAN の G と D に対しても成り立っており, G と D が敵対的に学習することで G は元データである x に似たデータを生成するよう学習する. 図 2.1 は GAN の生成モデルと識別モデルの関係を表したものである.

$$L(G, D) = \mathbb{E}_{x \sim P_{data}}[\log D(x)] + \mathbb{E}_{z \sim P_Z}[1 - \log D(z)] \tag{2.6}$$

2.6 CycleGAN

CycleGAN[7] は Zhu らが提案したデータドメイン間の相互変換を教師なし学習で行う機械学習モデルである. 対象となるドメイン X, Y に対して生成モデル G_x, G_y, 識別モデル D_x, D_y の 4 つのモデルで構成されている. CycleGAN では教師なし学習で

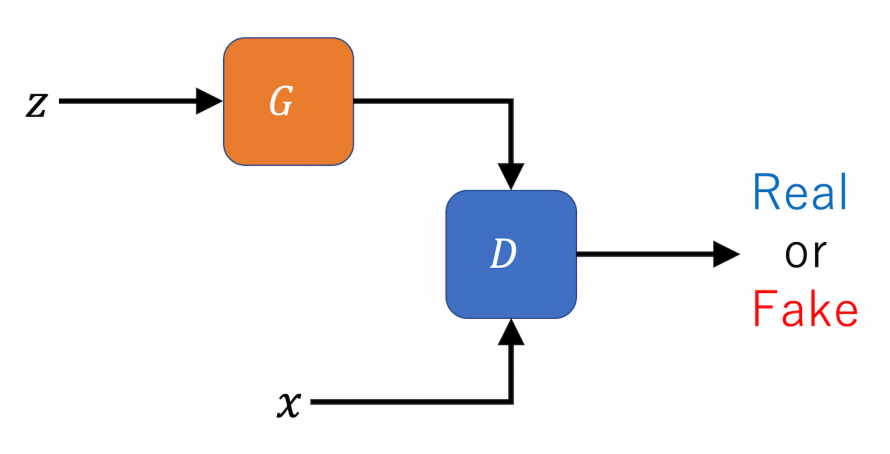


図 2.1: GAN 上でのデータ遷移

$G_x : Y \to X, G_y : X \to Y$ の相互変換を行うことから学習時のデータセットの制約が緩和された．これにより，男性から女性への変換といった現実では解の得られない，もしくはデータの獲得が困難な変換でもデータドメイン間の分布の違いから学習することができる．Zhu らの研究では損失関数を式 (2.7) のように定義している．式 (2.8) は式 (2.7) における L_{GAN}，式 (2.9) は式 (2.7) における L_{cyc} を表している．L_{GAN} は GAN の損失関数である Adversarial Loss のことであり，ドメイン X からドメイン Y へ変換する場合とドメイン Y からドメイン X へ変換する場合の 2 つの L_{GAN} が存在し，この 2 つの L_{GAN} により相互変換を学習する．L_{cyc} は Cycle Consistency Loss を示し，一方のドメインからもう一方のドメインへ変換した後，さらに元のドメインへと変換した場合に，元のデータの状態を保持しているような一貫性を持った変換を学習する．図 2.2 は CycleGAN 上でのデータ遷移を表したものである．

$$\begin{aligned} L(G_x, G_y, D_x, D_y) =& L_{GAN}(G_x, D_x, X, Y) \\ &+ L_{GAN}(G_y, D_y, Y, X) + \lambda L_{cyc}(G_x, G_y) \end{aligned} \tag{2.7}$$

$$\begin{aligned} L_{GAN}(G_x, D_x, X, Y) =& \mathbb{E}_{x \sim P_{data}(X)}[\log D_x(x)] \\ &+ \mathbb{E}_{y \sim P_{data}(Y)}[\log D_x(1 - G_x(y))] \end{aligned} \tag{2.8}$$

$$
\begin{aligned}
L_{cyc}(G_x, G_y) =& \mathbb{E}_{x \sim P_{data}(x)}[||G_x(G_y(x)) - x||_1] \\
&+ \mathbb{E}_{y \sim P_{data}(y)}[||G_y(G_x(y)) - y||_1]
\end{aligned} \tag{2.9}
$$

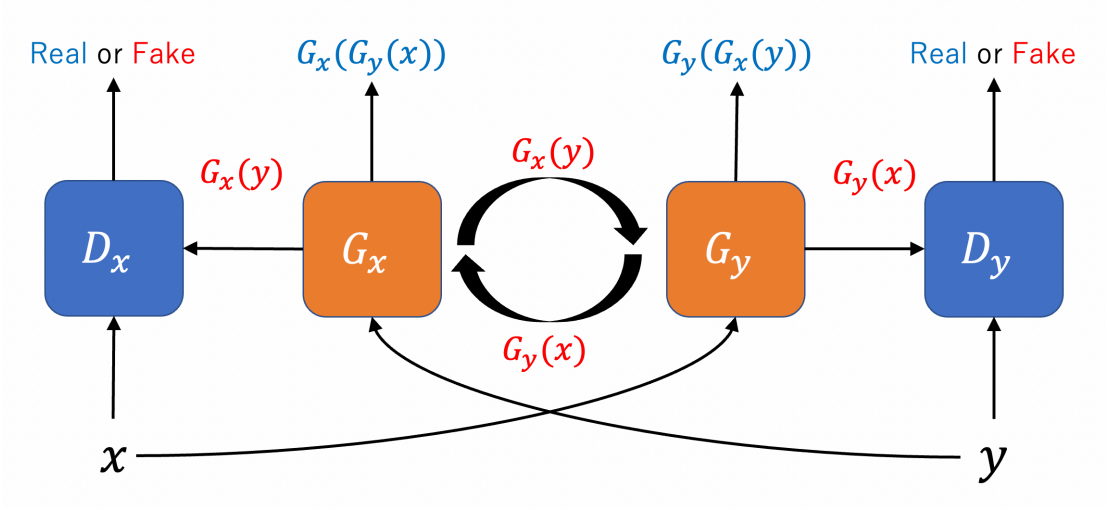


図 2.2: CycleGAN 上でのデータ遷移

2.7 Self-Attention

Self-Attention[8] は注意メカニズムの一種であり, 画像全体の情報の中から現在のタスク目標によって重要な情報を選択することにより, 特徴情報を有効に活用することができる. また, 他の付加的な情報を用いることなく入力からより関連性の高い情報を抽出し, 自己に注視することができる. 入力画像の文脈情報を把握することで, 画像中の長距離, 多階層の依存関係を上手く処理するとともに, 生成画像中の各位置の情報を調整することができる. Self-Attention の構造を図 2.3 と式 (2.10) に示す.

$$
O_j = \sum_{i=1}^{N} \frac{\exp(f(x_i)^T g(x_j))}{\sum_{i=1}^{N} \exp(f(x_i)^T g(x_j))} * h(x_i) \tag{2.10}
$$

畳み込みカーネルの受容野は局所的であるため, 画像中の異なる部分の関連付けを行う際には, 多くの層を積み重ねる必要がある. また, 畳み込みニューラルネットワーク (CNN) では, 各畳み込み層のカーネルのサイズには限りがあるため, 畳み込み演算は, 画素周辺の限ら

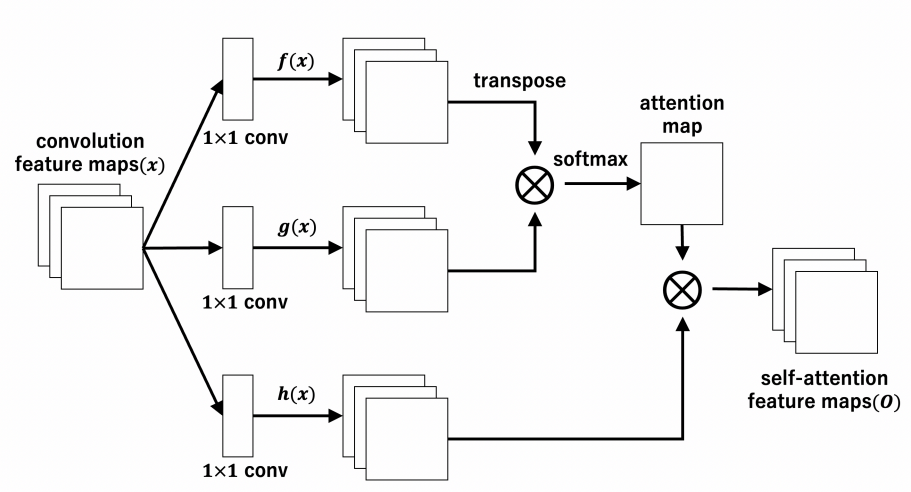


図 2.3: Self-Attention の構造

れた近傍領域しかカバーすることができない. そのため, 遠くの特徴をカバーするには, 多層の畳み込み演算とプーリング演算をする必要があり, 特徴マップのサイズが小さくなってしまう. 結果として, 元の画像に対応する領域にマッピングし直す際に, 畳み込みカーネルの後段がカバーする領域が大きくなるため, 一定量の特徴は簡単には捉えられないことになる. これに対して, Self-Attention では, 図 2.3 から特徴マップとその転置の掛け算と捉えることができるため, 任意の 2 箇所の画素間の依存関係を直接学習できる. そのため, 画像の大域的な幾何学的特徴を一度に取得でき, 特定の構造や幾何学的特徴の学習が容易となる.

2.8　敵対的サンプル (Adversarial Example)

敵対的サンプルとは機械学習モデルに誤った予測をさせるために, 意図的に小さな摂動 (ノイズ) を持たせた画像などのことである. 敵対的サンプルは分類問題において「意図的」に異なるクラスに誤分類させることが可能であり, AI のセキュリティや信頼性を評価するために用いられている. 以下の図 2.4 は敵対的サンプルを用いた攻撃の例で, 57.7%の確率でパンダと判定されている画像に摂動を加えることで 99.3%の確率で手長猿と判定されるようになった.

敵対的サンプルの生成手法の中でも代表的なものとして Fast Gradient Sign

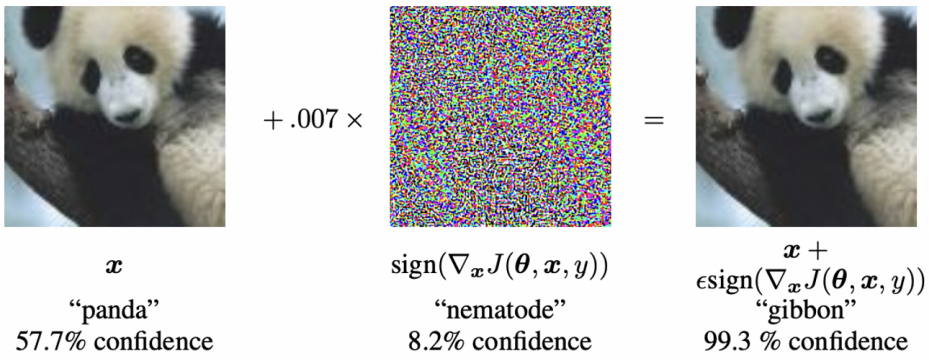


図 2.4: 敵対的サンプルを用いた攻撃の例
出典：Ian J. Goodfellow, et al. "Explaining and harnessing adversarial examples." *International Conference on Learning Representations (ICLR)*, 2015.

Method(FGSM) が挙げられる．FGSM はモデルの予測から得られる勾配の方向に沿って一度だけ摂動を加える単純な方法である．FGSM の計算式は式 (2.11) であり, 元の画像 x に対して, 損失関数 $J(\theta, x, y)$ の勾配を計算したのちに勾配の符号 sign に従って, 微小な摂動（ノイズ）を加えることで敵対的サンプルを作成する．

$$x_{adv} = x + \alpha * \mathsf{sign}(\nabla_{x^n} J(\theta, x^n, y)) \tag{2.11}$$

x_{adv}: 敵対的サンプル, θ: パラメータ, x: 入力, y: ラベル, J: 損失関数
α: 摂動の大きさ, sign: 勾配の符号

2.9 Projected Gradient Descent(PGD)

Projected Gradient Descent(PGD, 射影勾配降下)[10] とは, FGSM を拡張したより強力な攻撃手法でターゲット画像に対して複数回の勾配更新を行い, 敵対的サンプルを作成する．FGSM の勾配更新が 1 回だけなのに対して, PGD は任意の回数だけ攻撃を繰り返すため, より確実に誤分類を起こすことが出来る．PGD の計算式は式 (2.12) であり, 敵対的サンプル敵対的サンプル x^{n+1} がサンプル x^n から生成される．これを，入力画像 x^1 から始める．また, 画像のピクセル値が変更範囲 α を超えないように, 各ステップで P 関数を用いて元の画像からのズレを制限する．

2.9 Projected Gradient Descent(PGD)

$$x^{n+1} = P(x^n + \alpha * \mathsf{sign}(\nabla_{x^n} J(\theta, x^n, y))) \tag{2.12}$$

x^{n+1}: 敵対的サンプル, θ: パラメータ, x^n: 入力, y: ラベル, J: 損失関数
α: 摂動の大きさ, sign: 勾配の符号, P: 範囲の制限

第 3 章

提案手法

本研究では, 畳み込みニューラルネットワーク（CNN）が分類に寄与する領域と, その領域内の形状やパターンの違いを可視化することを目的として, 敵対的サンプルを用いた手法を提案する. 具体的には, 画像を CNN モデルへ入力し, その出力に基づいて勾配を用いて敵対的サンプルを生成する. 敵対的サンプルは元の画像とは異なるクラスに分類されるという性質を持つため, 敵対的摂動が分類に寄与する領域およびその中のパターンや形状を反映していると仮定し, 解析を行う. 敵対的サンプルの生成には, 敵対的攻撃手法として PGD（Projected Gradient Descent）を採用する. また, 既存の説明性手法である Grad-CAM, Guided Backpropagation, および CycleGAN との比較を行い, 提案手法の有効性を検証する. 提案手法を下記に示す.

1. 塵肺と心肥大データセットを用いて CNN 分類モデルを学習
2. 学習済み分類モデルから得られた予測値を用いて画像に対して勾配を計算し, 摂動・敵対的サンプルを作成
3. 出力された敵対的サンプルの傾向を検証し, 分類に寄与する領域と領域内の形状とパターンを評価
4. Grad-CAM や Guided Backpropagation などの既存の説明性手法と比較し, 今回の提案手法の有効性の検証

第 4 章

実験内容

ここでは提案手法を検証するために用いたデータセットや CNN 分類モデル, CycleGAN モデル, 実験手順について説明する.

4.1 使用したデータセット

本研究では, 塵肺 (pneumoconiosis) と心肥大 (Cardiomegaly) を扱った二つのデータセットを使用している. 本研究で扱う画像データは全てグレースケール画像として扱う.

4.1.1 塵肺データセット

塵肺は, 長期間にわたる粉塵や微粒子の吸入・蓄積によって引き起こされる肺疾患であり, 肺にさまざまな病理学的変化をもたらす [12]. 塵肺は 胸部 X 線画像を使用して, すりガラス陰影の存在を検出することで診断される. 塵肺データセットには, NIOSH(National Institute of Occupational Safety and Health) の胸部 X 線画像データセット [16], 高知大学医学部 (KM) で収集された胸部 X 線画像データセット, NIHCC(National Institutes of Health Clinical Center) の胸部 X 線画像データセットの 3 つを使用した. これらのデータセットには塵肺症状が見られない健常な肺である「検出なし」(NF: No Finding) と「塵肺」のいずれかにラベル付けされている. 画像の枚数は NIOSH データセットは「検出なし」画像 23 枚,「塵肺」画像 28 枚 の計 51 枚, NIHCC データセットは「検出なし」画像 90 枚,「塵肺」画像 0 枚. KM データセットは「検出なし」画像 4 枚,「塵肺」画像 91 枚となっている. 塵肺データセットの内訳と画像サイズを表 4.2 に示す. また, 胸部 X 線画像からの塵

肺検出において, Zhang らの研究によると入力画像の肺野領域の抽出を行うことで分類精度が向上することが報告されており, 特に U-Net を使った手法で高い領域抽出精度が報告されている [15]. そこで本研究では, 塵肺画像の分類精度を向上させるため, データセットに対して U-Net を用いたセグメンテーションによる肺野領域抽出を実施した (図 4.1). このデータセットにおける検出なし画像と塵肺画像の例を図 4.2 に示す. 図 4.2 を見ると「塵肺」画像は「検出なし」画像と比較してすりガラス状の陰影が存在することが確認できるのに加えて, 肺野領域の境界線が白く変化していることが確認できる. このデータセットを train データ 160 枚, validation データ 41 枚, test データ 35 枚に分割して使用する. 画像サイズは 512 × 512 にサイズ変更している.

表 4.1: 使用する画像データ

	検出なし	塵肺
NIOSH	23	28
NIHCC	90	0
KM	4	91

表 4.2: 塵肺データセットの構成

train データ	validation データ	Test データ	合計	画像サイズ
160	41	35	236	512 × 512

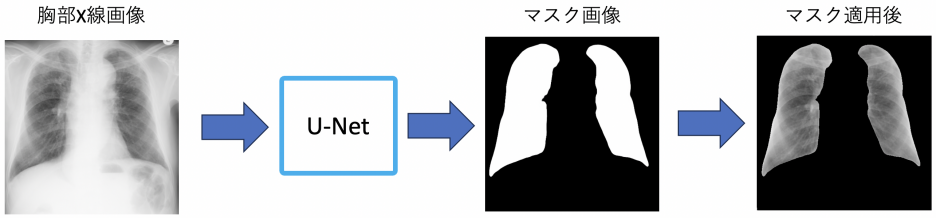


図 4.1: U-Net を用いた肺野領域抽出

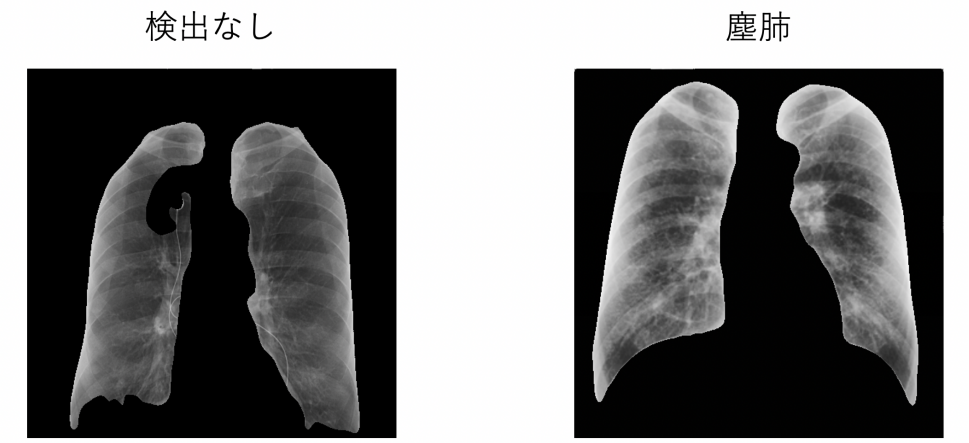


図 4.2: 塵肺データセットの画像の例

U-Net を用いた肺野領域抽出

U-Net の学習には Montgomery County X-ray Set [13][14] を用いた．このデータセットは, アメリカのメリーランド州モンゴメリー郡の保健福祉福祉省から収集されており, 800 枚の胸部 X 線画像と 704 枚のマスク画像から構成されている．今回はマスクの数に合わせた 704 枚の胸部 X 線画像を使用する．569 枚を Training データ, 71 枚を Validataion データ, 64 枚を Test データに分割して学習した．

4.1.2　心肥大データセット

心肥大データセットには Lee らが作成した心肥大の胸部 X 線画像データセットを使用する [11]．このデータセットはデータサイエンティストや機械学習実践者向けのオンラインコミュニティである Kaggle で入手可能な National Institutes of Health (NIH) の胸部 X 線データセットをもとにしており, 前後 (AP) ビュー、サポートデバイス、および画像に表示されている人工物 (ペースメーカー、心臓手術の傷跡、脊椎手術の人工物、肺と心臓の領域を覆い隠す IV ラインなど) を含む画像を除外している．また, 選別された胸部 X 線画像に対してトレーニングに必要な関連領域のみを含むように画像を切り取る処理を行い、高品質のデータセットを確保している。これらの一連の処理の結果, Lee らのデータセットは 368 枚の「検出なし」画像と 350 枚の「心肥大」画像で構成されている．データセットは訓練

用の train データと検証データに 7:3 の比率で分割され, さらに検証データは 2:1 の比率で validation データと test データに分割されている．画像サイズは学習の際に 224 × 224 にサイズ変更している．実際の心肥大データセットの画像の例を図 4.3 に示す．図 4.3 を見ると「心肥大」画像は「検出なし」画像と比較して心臓部分が肥大していることが確認できる．

表 4.3: 心肥大データセットの構成

train データ	validation データ	Test データ	合計	画像サイズ
516	130	72	718	224 × 224

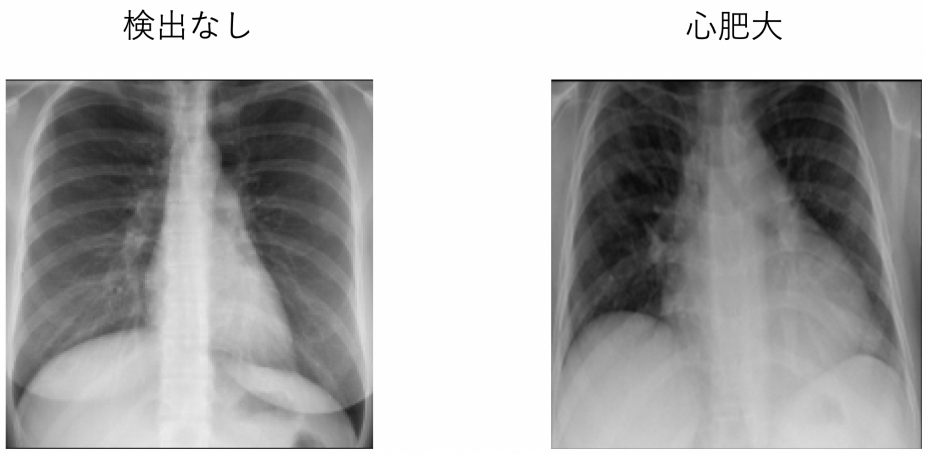


図 4.3: 心肥大データセットの画像の例

4.2　使用したモデル

4.2.1　分類モデル

今回使用する塵肺と心肥大の分類モデルは VGG16 をもとにしており, 学習の安定化を図るために Batch Normalization を導入している．これにより高い学習率を設定できるようになる．また, 本来の VGG16 モデルの全結合層の代わりに Global Average Pooling 層を採用しており, モデル全体のパラメータ数を削減して過学習を抑制することが出来る．今回使用する分類モデルの構造を図 4.4 に示す.

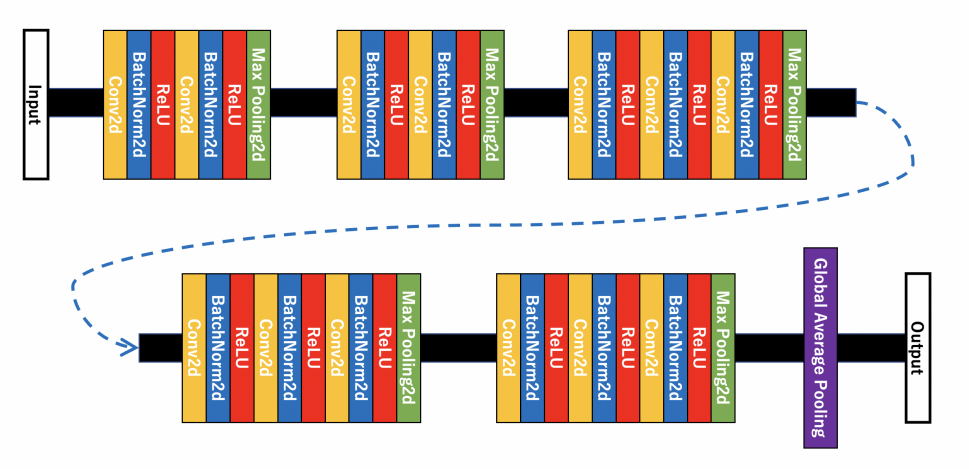


図 4.4: 分類モデルの構造

4.2.2　CycleGAN モデル

本研究では, 説明可能性手法の一つとして CycleGAN を採用する．ここでは今回用いた CycleGAN を構成する生成モデルと識別モデルの構造について述べる．

生成モデル

生成モデルはエンコーダ (特徴抽出部), 9 ブロックの residual block, Self-Attention 層, デコーダ (出力部) で構成されている．正規化には Instance Normalization, 活性化関数には ReLU を使用する．エンコーダでは, 入力画像は 3 つの畳み込み層を使用してダウンサンプリングされる．その後, 9 つの residual block を使用して画像のさまざまな特徴を抽出し, 画像の特徴ベクトルを「症状あり」の画像なら「検出なし」画像に, 「検出なし」画像なら「症状あり」の画像に変換する．識別器の構造を図 4.5 に示す．

識別モデル

識別モデルは画像から特徴を抽出し, 30 × 30 の特徴マップを出力します．正規化には Instance Normalizetion, 活性化関数には LeakyLeRU を使用する．識別器の構造を図 4.6 に示す．

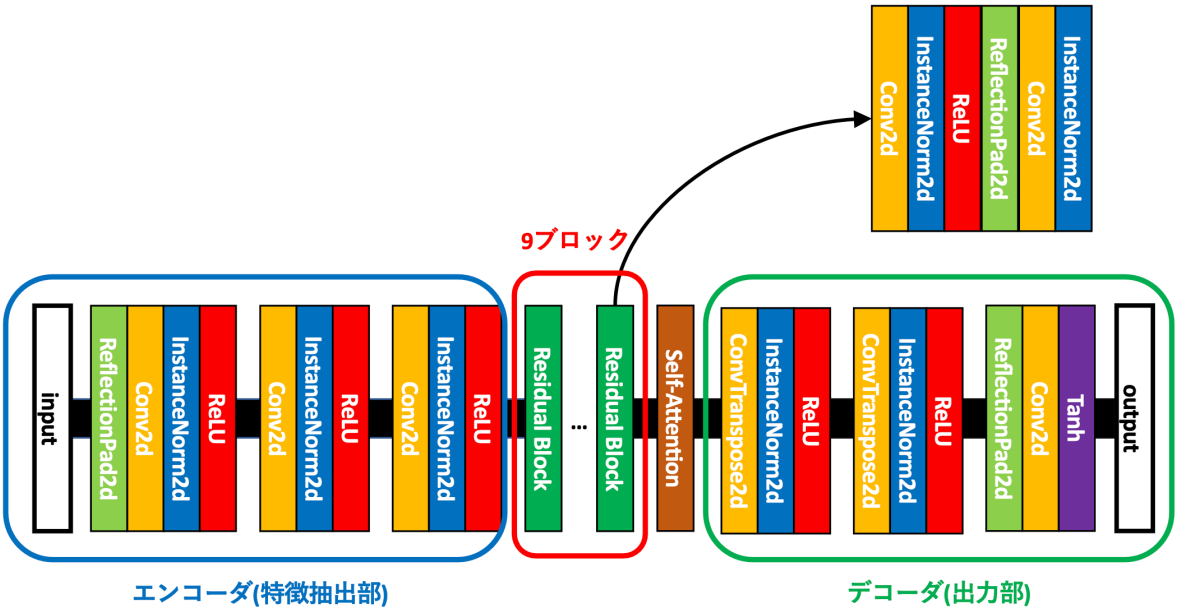


図 4.5: 生成モデルの構造

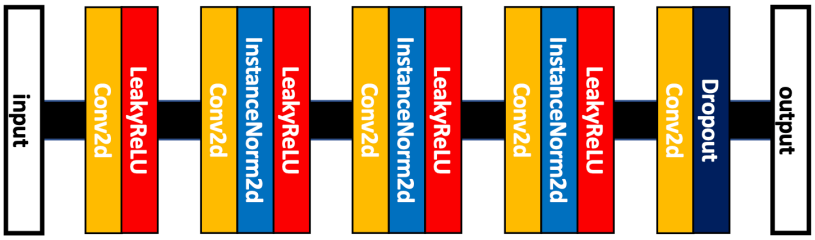


図 4.6: 識別モデルの構造

4.3　実験手順

はじめに, 用意した塵肺・心肥大データセットを用いて分類モデルをそれぞれ学習させる. その後, 学習済み分類モデルに対して画像を入力して得られる勾配を用いて敵対的サンプルを作成する. この時敵対的サンプル作成に使用するデータは作成したデータセットのテストデータとする. 使用する敵対的攻撃手法は PGD を採用しており, 付与する摂動の係数は 0.01, 反復回数は 5 回に設定した. 最終的に 5 回に渡って生成された摂動を足し合わせたものを説明性手法としての指標とする. 実際に PGD を適応する際の流れを図 4.7 に示す. 次に, 得られた敵対的サンプルを解析して, 分類に寄与する領域や領域内のパターン

が得られたかどうかを評価する．さらに，既存の説明性手法である Grad-CAM や Guided Backpropagation，CycleGAN と敵対的サンプルを比較して注目部位の比較や結果の違いなどを評価する．それらの結果から今回の提案手法を有効性を評価する．また，各モデルを学習させるにあたり，ハイパーパラメータは表 4.4 のように設定した．

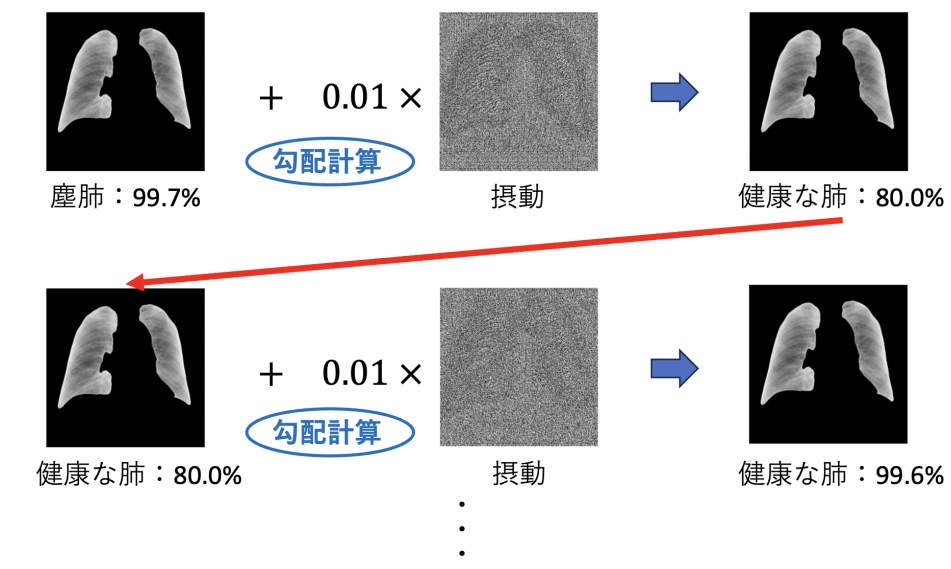


図 4.7: PGD 適応の流れ

この図は，PGD を塵肺患者の胸部 X 線画像に適用した例を示している．もともと 99.7% の精度で塵肺と判定されていた画像に摂動を加えることで，予測が変化し，80.0%の精度で健康な肺と判定されるようになった．さらに，勾配を計算して摂動を加えることで健康な肺であるという予測の精度がさらに向上した．

表 4.4: 各モデルのハイパーパラメータ

		Epoch	Batch size	Learning Rate	Optimizer
分類モデル	塵肺	100	16	1×10^{-4}	Adam
	心肥大	30	16	1×10^{-4}	Adam
CycleGAN モデル	塵肺	700	1	2×10^{-4}	Adam
	心肥大	300	1	2×10^{-4}	Adam

第 5 章

実験結果

5.1　分類モデルの学習結果

ここでは塵肺データセットと心肥大データセットを用いて分類モデルをそれぞれ学習した際の結果を述べる. 塵肺データセットと心肥大データセットを使用した際の Accuracy, Loss を表 5.1 と表 5.2 に, 混同行列を図 5.1 と図 5.2 に示す. また, 学習中の Accuracy, Loss の推移を図 5.3 と図 5.4 に示す.

表 5.1: 分類モデルの評価: 塵肺データセット

	Training	**Validation**	**Test**
Accuracy	100.0%	95.12%	97.14%
Loss	0.0078	0.2029	0.0750

表 5.2: 分類モデルの評価: 心肥大データセット

	Training	**Validation**	**Test**
Accuracy	99.42%	96.15%	97.22%
Loss	0.0184	0.0891	0.1149

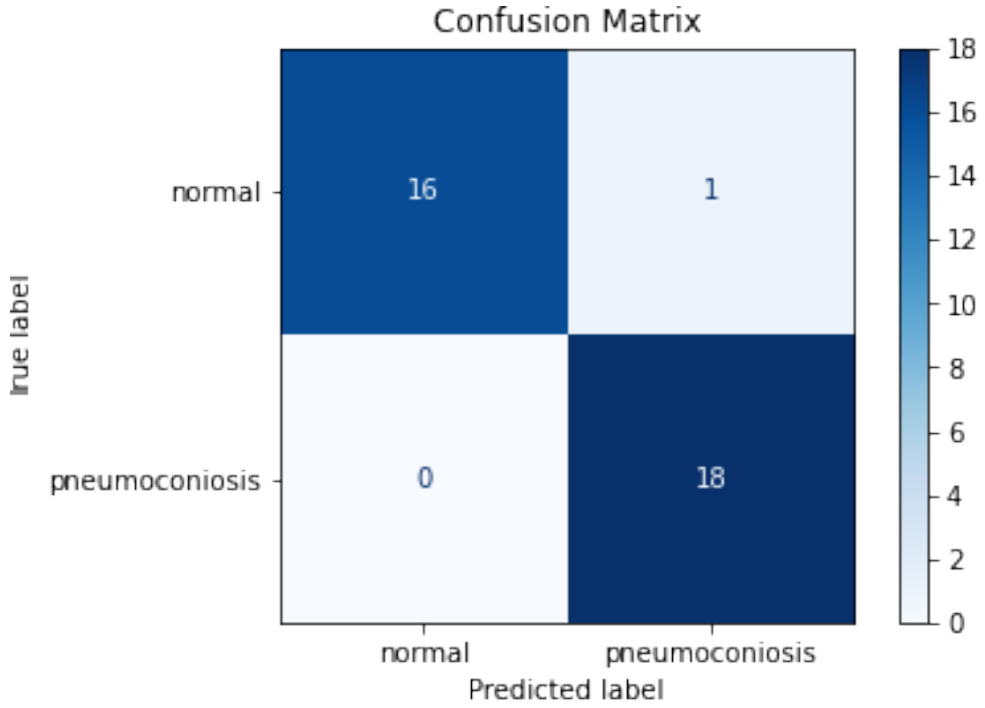


図 5.1: 混同行列: 塵肺データセットの場合

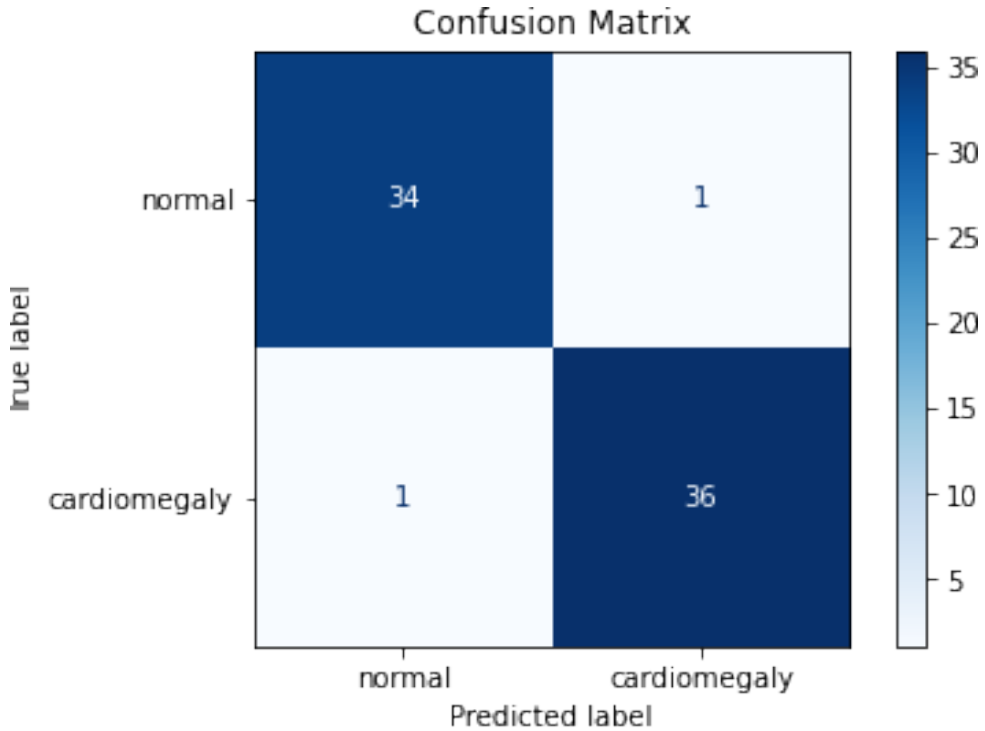


図 5.2: 混同行列: 心肥大データセットの場合

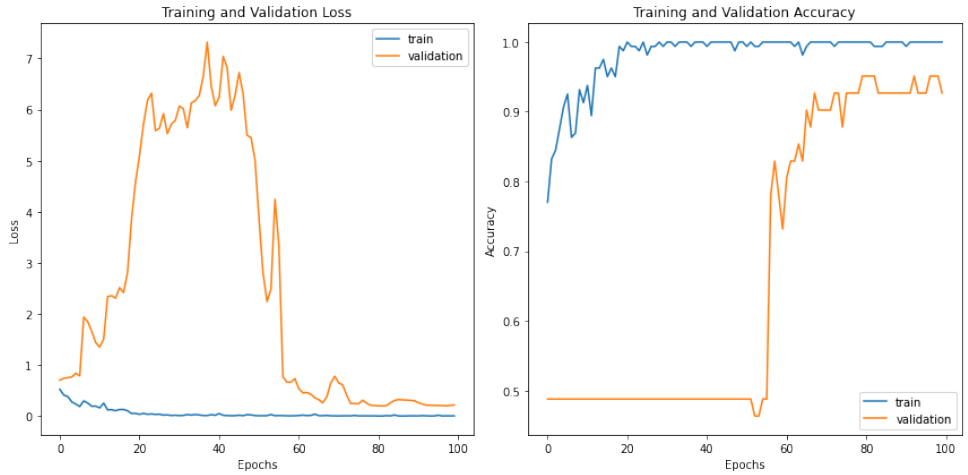


図 5.3: 分類モデルの学習の推移: 塵肺データセット

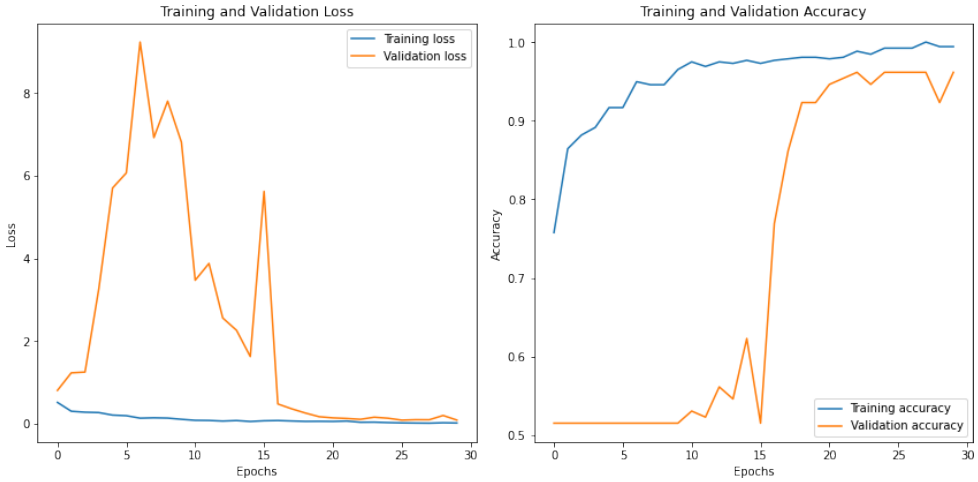


図 5.4: 分類モデルの学習の推移: 心肥大データセット

5.2　CycleGAN の学習結果

CycleGAN の学習結果を示す.

5.2.1　塵肺データセット

塵肺データセットに対して CycleGAN を適用した場合, 「検出なし」→「塵肺」の変換では, 肺野領域全体の画素値を増加させる傾向が見られた. 一方で「塵肺」→「検出なし」の変換では, 先程とは逆で肺野領域全体の画素値を減少させるような変化が見られた. 「検出なし」→「塵肺」の変換結果を図 5.5, 「塵肺」→「検出なし」の変換結果を図 5.6 にまと

める．また, これらの変換後の画像を本研究で使用した分類モデルで分類した結果, 精度は 42.85%となった．分類結果の混同行列を図 5.7 に示す．結果から「検出なし」画像→「塵肺」の変換は上手くいっていると見做すことができるが, 「塵肺」→「検出なし」の変換は不安定であることがわかる．

5.2.2　心肥大データセット

心肥大データセットに対して CycleGAN を適用した場合, 「検出なし」→「心肥大」の変換では, 主に心臓側面を追加する傾向が見られた．一方で「心肥大」→「検出なし」の変換では, 先程とは逆で心臓側面を削るような変化が見られた．「検出なし」→「心肥大」の変換結果を図 5.8, 「心肥大」→「検出なし」の変換結果を図 5.9 にまとめる．また, これらの変換後の画像を本研究で使用した分類モデルで分類した結果, 精度は 77.02%となった．分類結果の混同行列を図 5.10 に示す．結果から「検出なし」→「心肥大」の変換は上手くいっていると見做すことができるが, 「心肥大」→「検出なし」の変換は不安定であることがわかる．

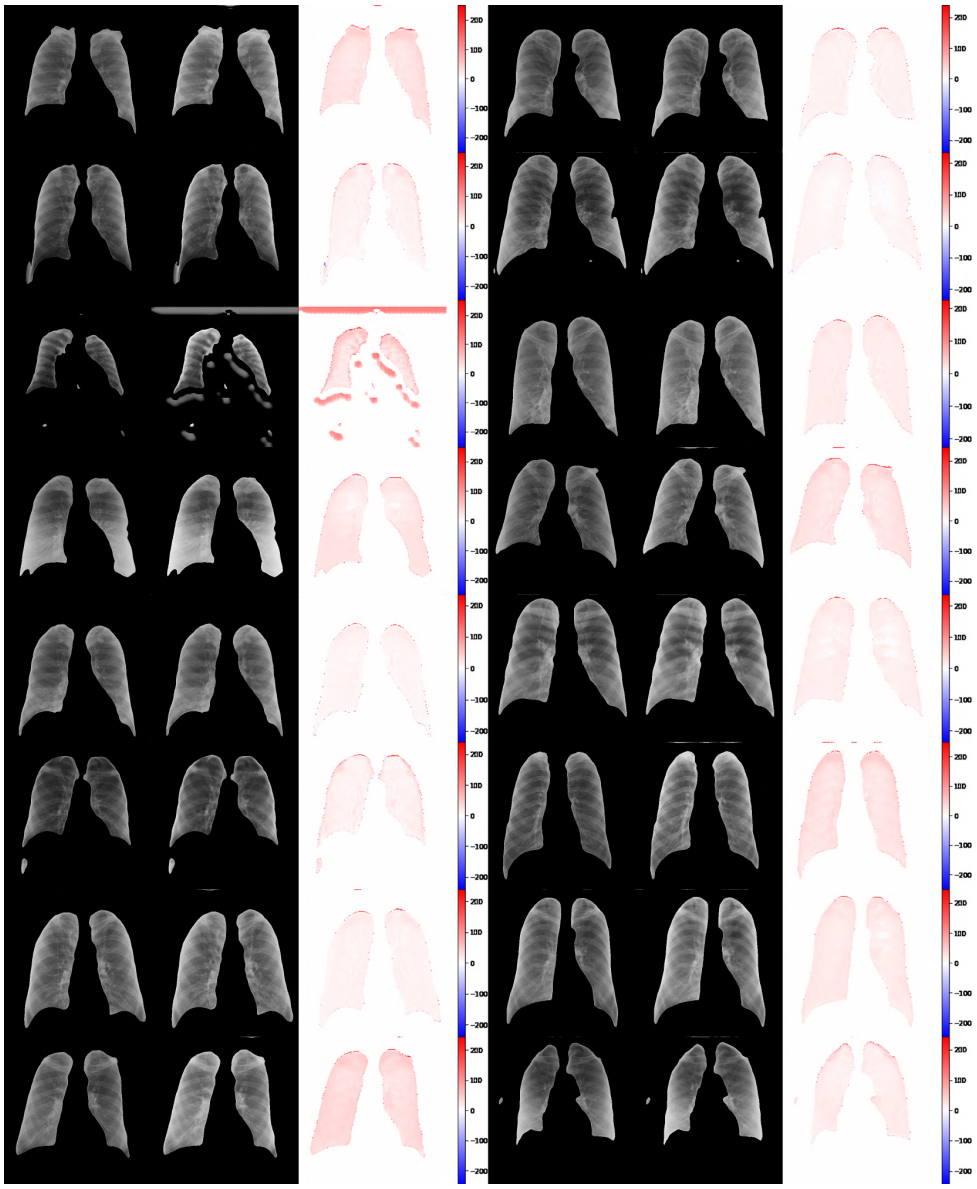


図 5.5: CycleGAN: 「検出なし」→「塵肺」の変換結果

CycleGAN を用いて「検出なし」画像を「塵肺」画像に変換した結果. (1) 変換前, (2) 変換後, (3) 変換前後の差分 となっている. 変換後の画像は変換前の画像と比較して肺野領域全体の画素値を増加させる傾向があることが分かる.

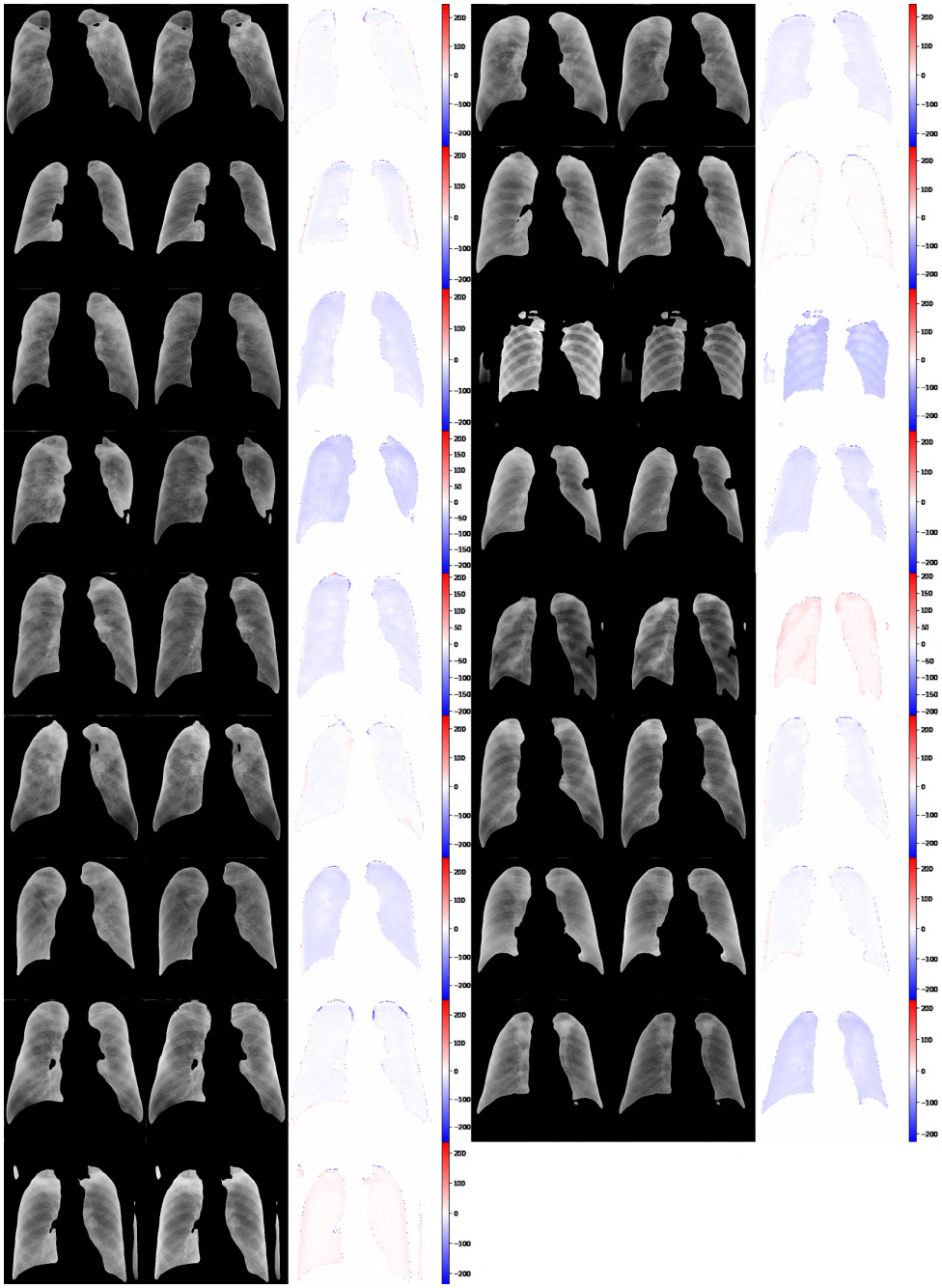


図 5.6: CycleGAN: 「塵肺」→「検出なし」の変換結果

CycleGAN を用いて「塵肺」画像を「検出なし」画像に変換した結果．(1) 変換前，(2) 変換後，(3) 変換前後の差分 となっている．変換後の画像は変換前の画像と比較して肺野領域全体の画素値を減少させる傾向があることが分かる．

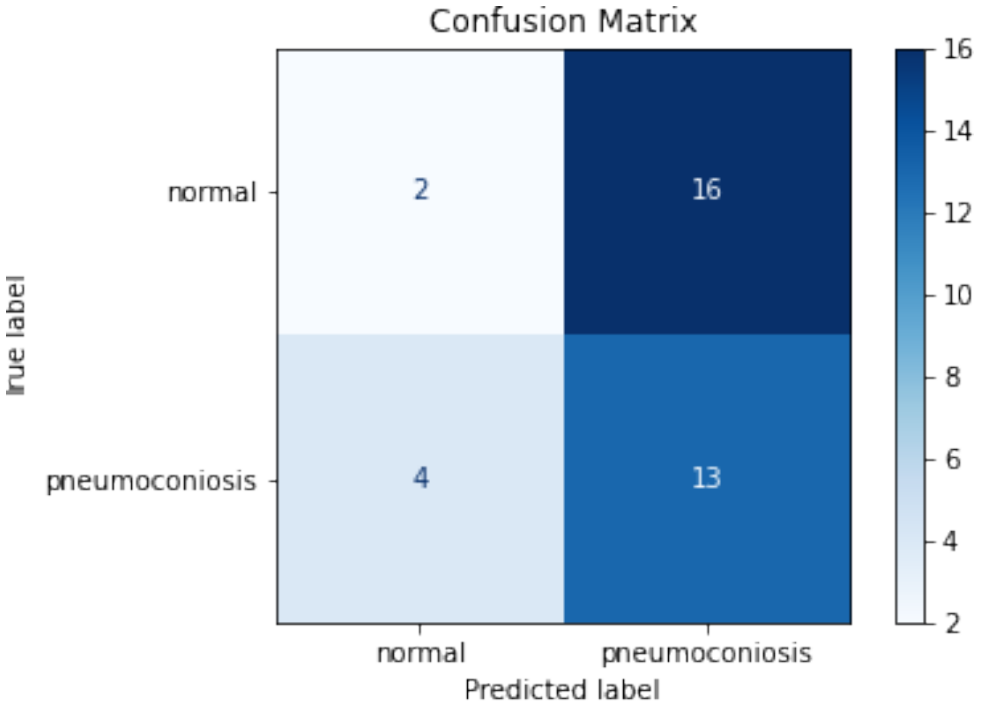


図 5.7: **塵肺データセット** CycleGAN による変換後の画像を分類モデルで予測: 混同行列

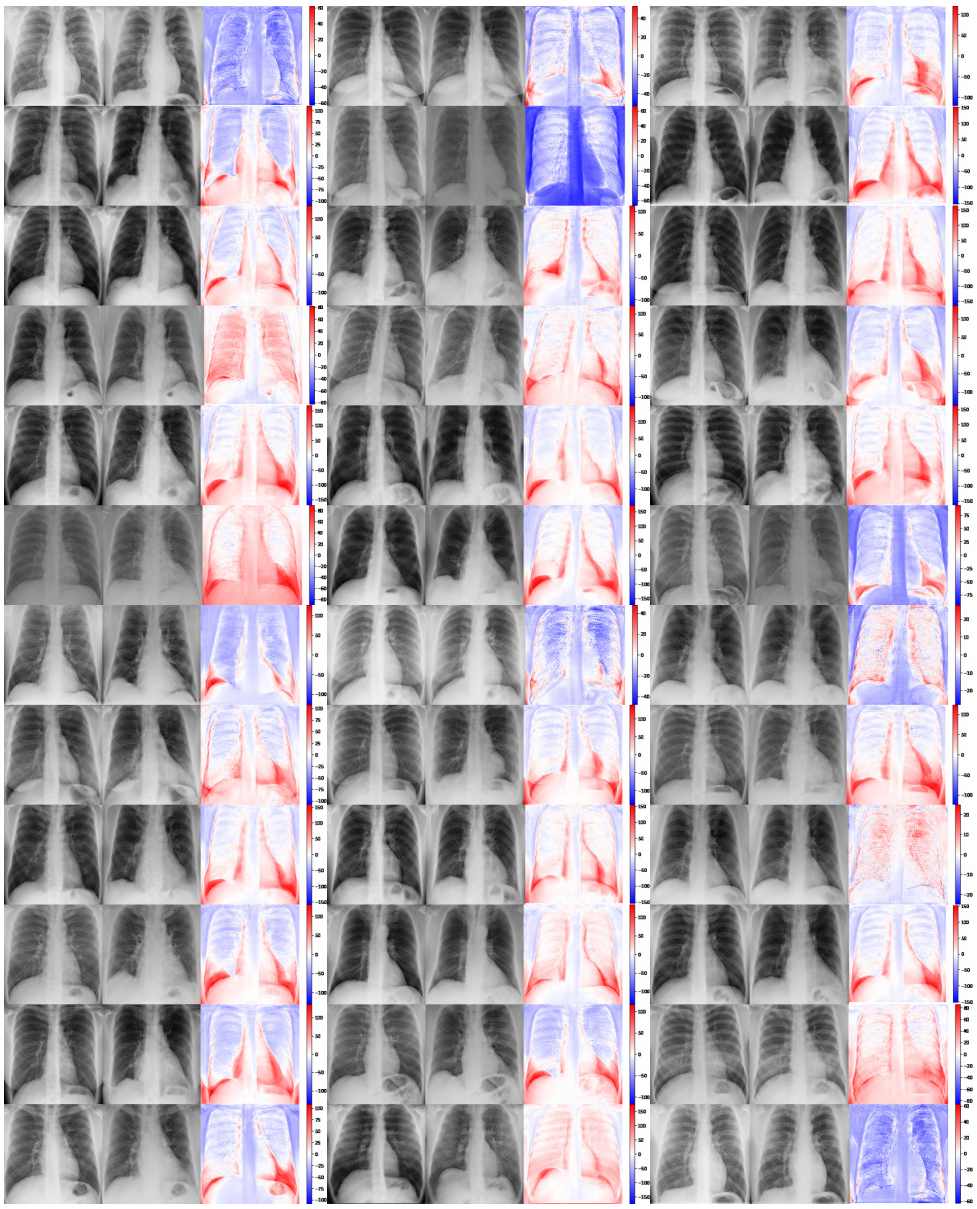
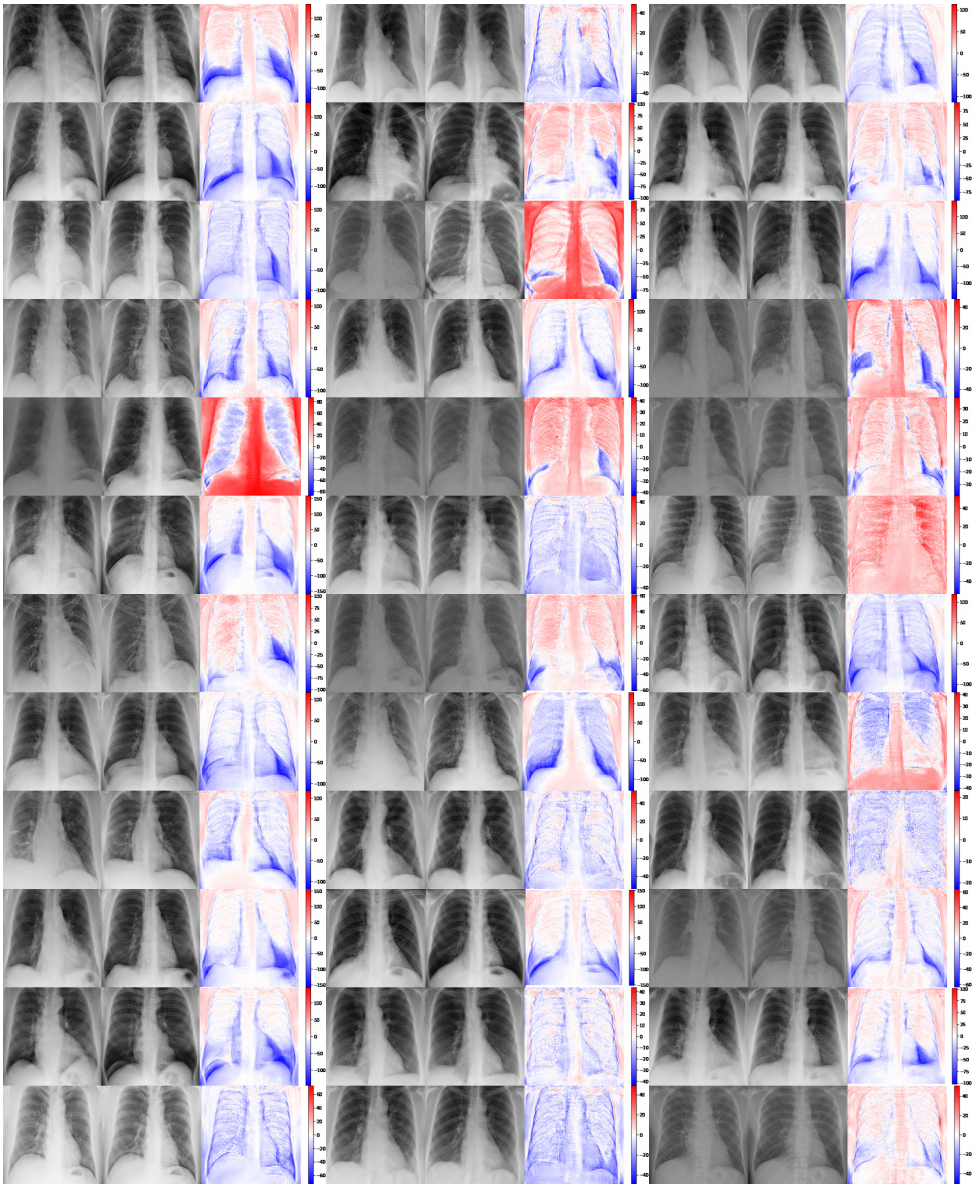


図 5.8: CycleGAN: 「検出なし」→「心肥大」の変換結果

CycleGAN を用いて「検出なし」画像を「心肥大」画像に変換した結果．(1) 変換前, (2) 変換後, (3) 変換前後の差分 となっている．変換後の画像は変換前の画像と比較して心臓の側面部分を追加する傾向があることが分かる．



(1)　(2)　(3)

図 5.9: CycleGAN: 「心肥大」→「検出なし」の変換結果

CycleGAN を用いて「心肥大」画像を「検出なし」画像に変換した結果. (1) 変換前, (2) 変換後, (3) 変換前後の差分 となっている. 変換後の画像は変換前の画像と比較して心臓の側面部分を削る傾向があることが分かる.

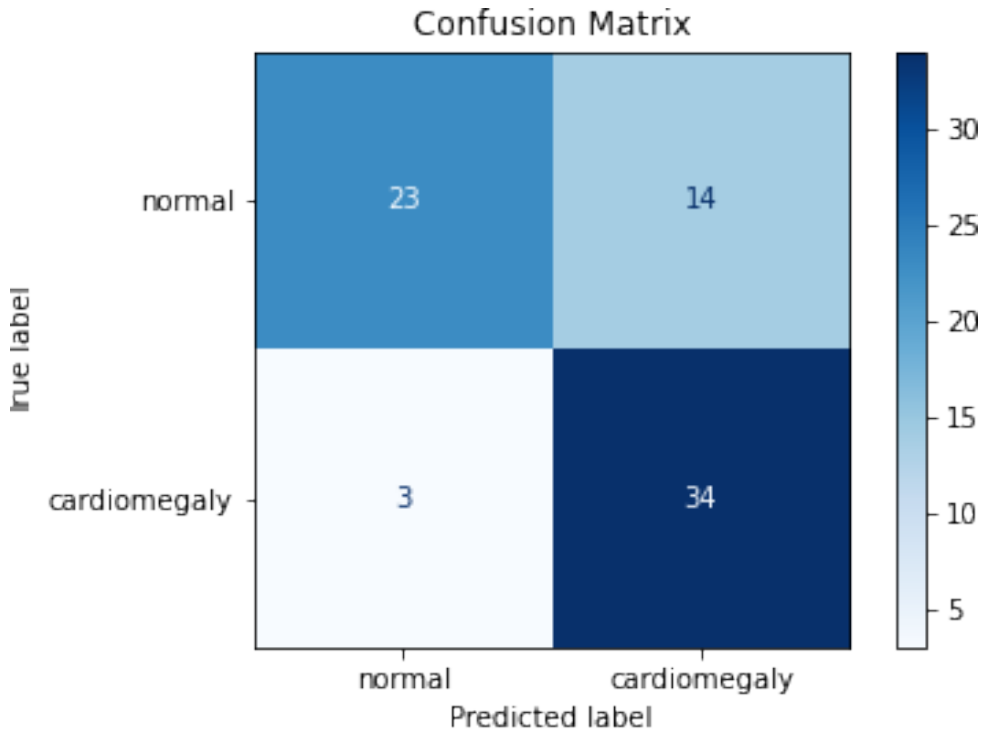


図 5.10: **心肥大データセット** CycleGAN による変換後の画像を分類モデルで予測: 混同行列

5.3　敵対的サンプルの解析

ここでは実際に獲得できた敵対的サンプルが分類モデルの識別にもたらす影響について述べる.

5.3.1　塵肺データセット

図 5.11 は, 塵肺画像から生成された敵対的サンプルの適用例を示している. この摂動は, 5 回繰り返し生成された摂動を重ね合わせたものであり, 注目度を示すカラーマップとして可視化されている. 注目度の高い特徴ほど変化量が大きくなることが確認できる. 本例では, 「塵肺」から「検出なし」へと判定が変化する過程であり, 主に肺野領域の境界とその内側に変化が集中していることがわかる.

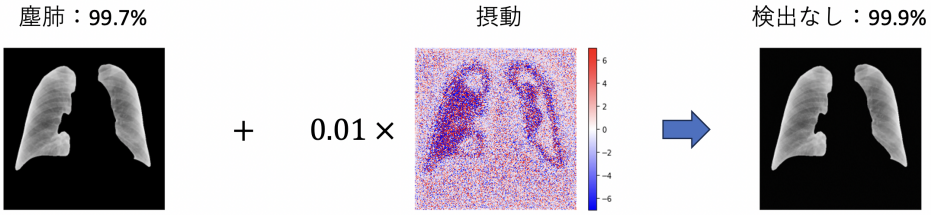


図 5.11: 敵対的サンプルの適用例

ここで, 敵対的サンプルの特徴を視覚的に強調するため, 注目度が最も高い特徴のみを抽出した. 敵対的サンプルの摂動は勾配計算に基づき, 値が 1 または -1 の二値で構成されているため, 5 回の繰り返しにより得られる値の範囲は -5 から 5 となる. ここでは, 注目度が最も高い特徴として値が 5 または -5 の領域を抽出しており, 図 5.12 にその結果を示す. 図 5.12 で示された抽出前の摂動と, 抽出後の摂動を元の塵肺画像に適用した結果, 前者は 99.7%, 後者は 99.2% の精度で「検出なし」と判定された. この結果は他の画像データ群においても一貫しており, 特徴抽出後でも元の摂動と同等のパフォーマンスを維持していることが示された. そのため, 説明性手法の指標にはこの注目特徴を抽出して視覚性をあげた摂動とそれを元画像に加えた敵対的サンプルを使用するものとする.

また, 敵対的サンプルの特徴の性質を解析するために, 生成した摂動に対して 5 × 5mean

図 5.12: 塵肺データセット 敵対的サンプル: 注目度の高い特徴の抽出

フィルタ, 3 × 3median フィルタ, gaussian フィルタを適用した．フィルタを適用した例を「検出なし」，「塵肺」のそれぞれで解説する．図 5.13 は「検出なし」画像の摂動に対するフィルタの適用例である．mean フィルタや gaussian フィルタを適用した結果をみると, 肺野領域の境界の画素値を増加させるとともに, 肺野領域の境界の内側については画素値を減少させていることが確認できる．このことから，「検出なし」画像の肺野領域の境界と内側の差のコントラストを上げて強調することで, 判定を「塵肺」へと反転させていることが読み取れる．

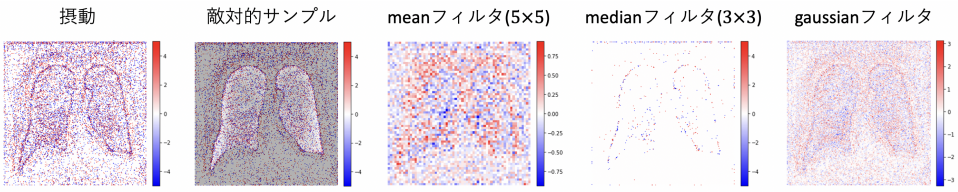


図 5.13: 塵肺データセット 敵対的サンプルへのフィルタ適用: 「検出なし」

図 5.14 は「塵肺」画像の摂動に対するフィルタの適用例である．mean フィルタや gaussian フィルタを適用した結果をみると, 肺野領域の境界の画素値を減少させるとともに, 肺野領域の境界の内側については画素値を増加させていることが確認できる．このことから，「検出なし」の敵対的サンプルとは逆に，「塵肺」の敵対的サンプルは肺野領域の境界の輝度を下げるとともに内側の輝度は上げるという形で肺野領域の境界と内側の差のコントラストを下げて抑制することで, 判定を「検出なし」へと反転させていることが読み取れる．

「検出なし」画像の摂動にフィルタを適用した結果をまとめたものを図 5.15 に，「塵肺」

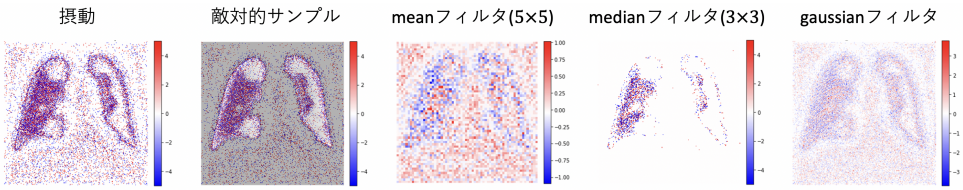


図 5.14: 塵肺データセット 敵対的サンプルへのフィルタ適用: 「塵肺」

画像の摂動に適用した結果をまとめたものを図 5.16 に示す. 「検出なし」画像から生成された摂動では, 共通して肺野領域の境界の輝度を上昇させるとともに肺野領域の内側の輝度を低下させる傾向が見られた. 一方で, 「塵肺」画像から生成された摂動は, 肺野領域の境界の輝度を低下させるとともに肺野領域の内側の輝度を上昇させる傾向が見られた. このことから, 塵肺データから生成される敵対的サンプルは「塵肺」と「検出なし」の判定を反転させる際に, 肺野領域の境界のコントラストを変化させていることが分かる.

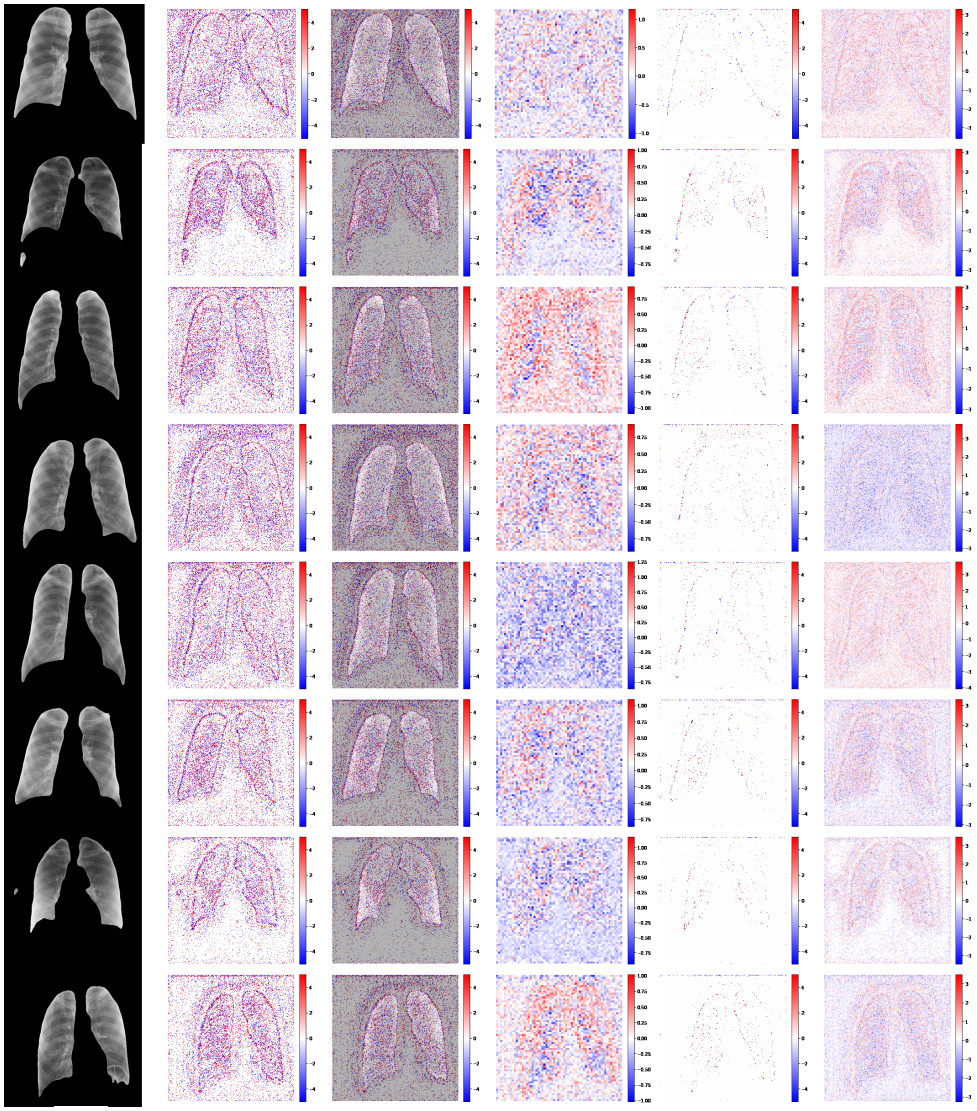


図 5.15: 塵肺データセット 敵対的サンプルの解析: 「検出なし」画像への適用

塵肺データセットにおける「検出なし」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

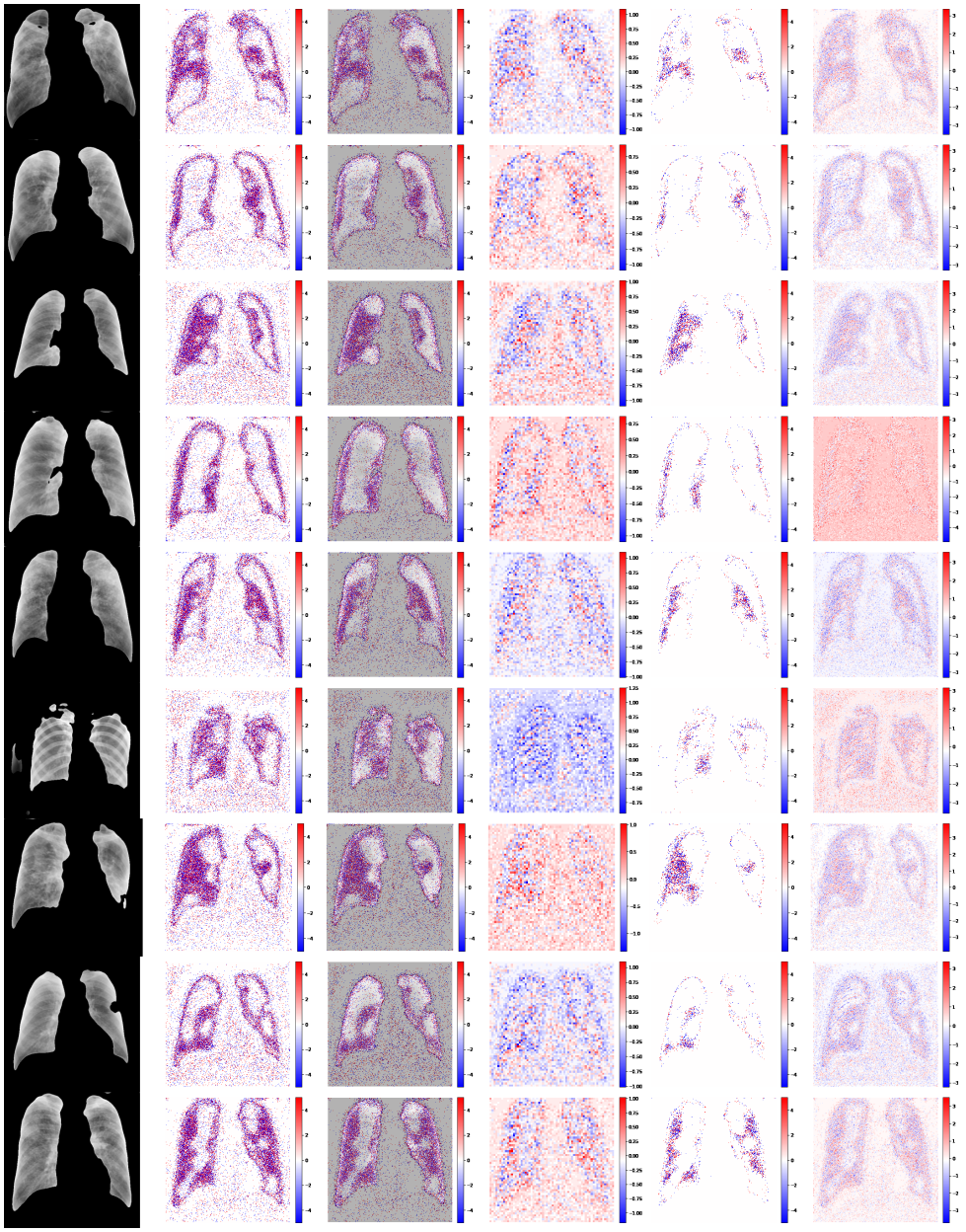


図 5.16: 塵肺データセット 敵対的サンプルの解析: 「塵肺」画像への適用

塵肺データセットにおける「塵肺」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

5.3.2　心肥大データセット

心肥大データから作成された敵対的サンプルを解析する．図 5.17 は「検出なし」画像の摂動に対するフィルタの適用例である．作成された敵対的サンプルを見ると, 心臓の側面の画素値を増加させていることが分かる．mean フィルタや median フィルタ, gaussian フィルタを適用した結果を見ても, 心臓の側面の外側をなぞるようにして画素値を増加させていることが見て取れる．また, 心臓の側面に沿うように画素値が増加している一方で, さらにその外側では増加部分に沿うように画素値が減少している部分があることが確認できる．これは心臓側面の画素値を増加させるとともに, その周りの画素値を低下させることでギャップを作る, つまりはコントラストを強調することで CNN に心臓部分が拡大したように判断させているのではないかと考えられる．

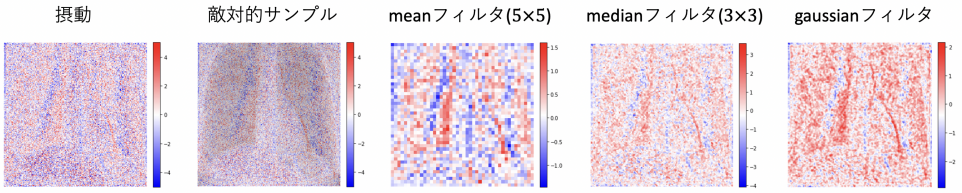


図 5.17: 心肥大データセット 敵対的サンプルへのフィルタ適用: 「検出なし」

図 5.17 は「心肥大」画像の摂動に対するフィルタの適用例である．作成された敵対的サンプルを見ると, 心臓の側面の画素値を減少させていることが分かる．mean フィルタや median フィルタ, gaussian フィルタを適用した結果を見ても, 心臓の側面をなぞるようにして画素値を減少させていることが見て取れる．また, 心臓の側面をなぞるように画素値が減少している一方で, さらにその外側では減少部分に沿うように画素値が増加している部分があることが確認できる．これは心臓側面の画素値を減少させるとともに, その周りの画素値を増加させることで, 心臓以外の暗い部分と心臓の側面の明暗の差を縮める, つまりはコントラストを抑制することで, CNN に心臓が縮小したと判断させているのではないかと考えられる．

「検出なし」画像の摂動にフィルタを適用した結果をまとめたものを図 5.19 に, 「心肥大」画像の摂動に適用した結果をまとめたものを図 5.20 に示す．「検出なし」画像から生成さ

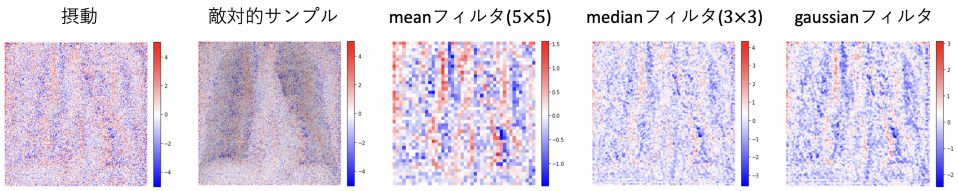


図 5.18: 心肥大データセット 敵対的サンプルへのフィルタ適用: 「心肥大」

れた摂動では, 共通して心臓側面の画素値を増加させるとともにそれに沿うように増加部分の外側の画素値が低下する傾向が見られた. 一方で, 「心肥大」画像から生成された摂動は, 共通して心臓側面の画素値を減少させるとともにそれに沿うように増加部分の外側の画素値が増加する傾向が見られた. このことから, 心肥大データから生成される敵対的サンプルは「心肥大」と「検出なし」の判定を反転させる際に, 心臓側面のコントラストを強調もしくは抑制し, 心臓のサイズを誤認させることで CNN の判定を反転させていることが読み取れる.

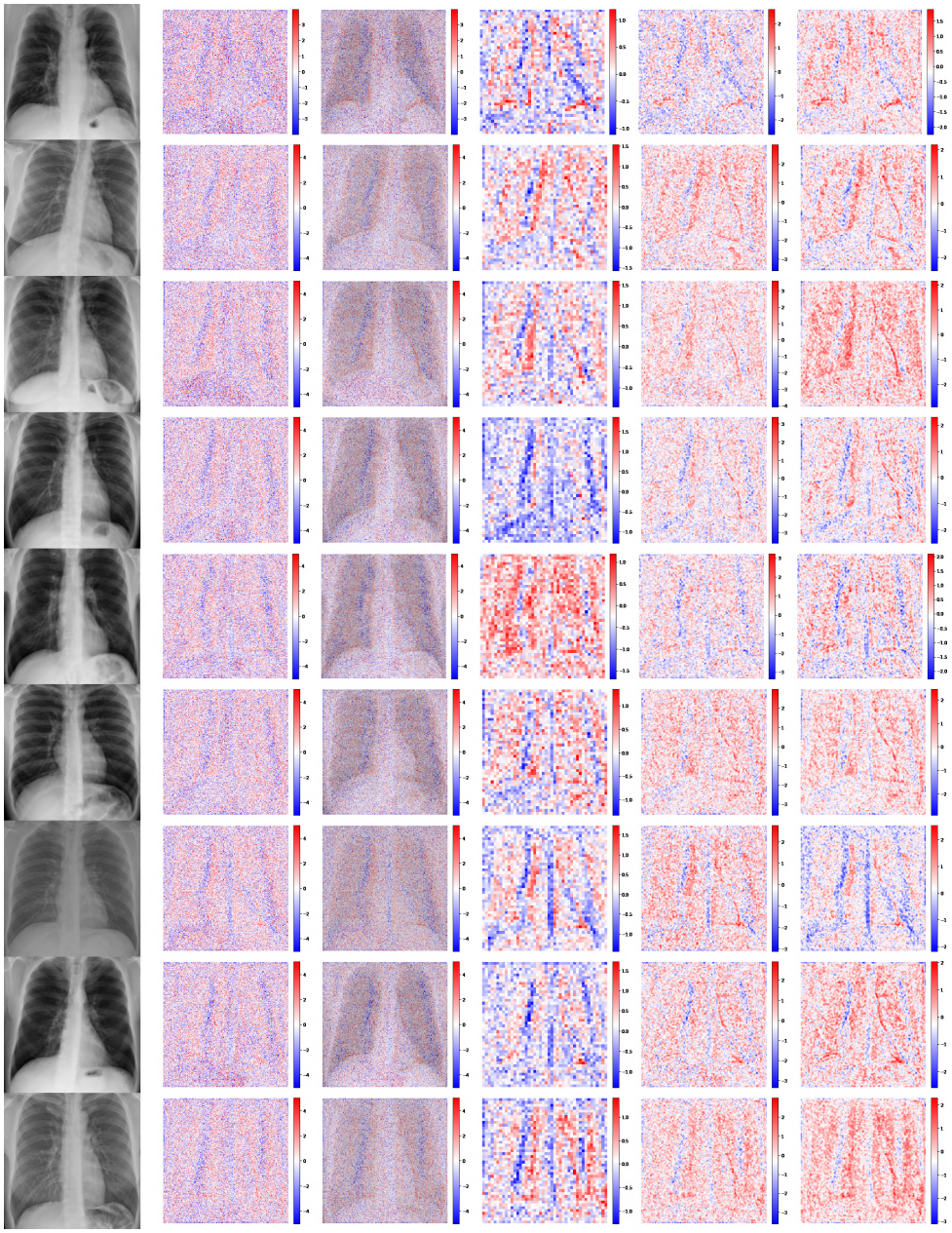


図 5.19: 心肥大データセット 敵対的サンプルの解析: 「検出なし」画像への適用

心肥大データセットにおける「検出なし」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

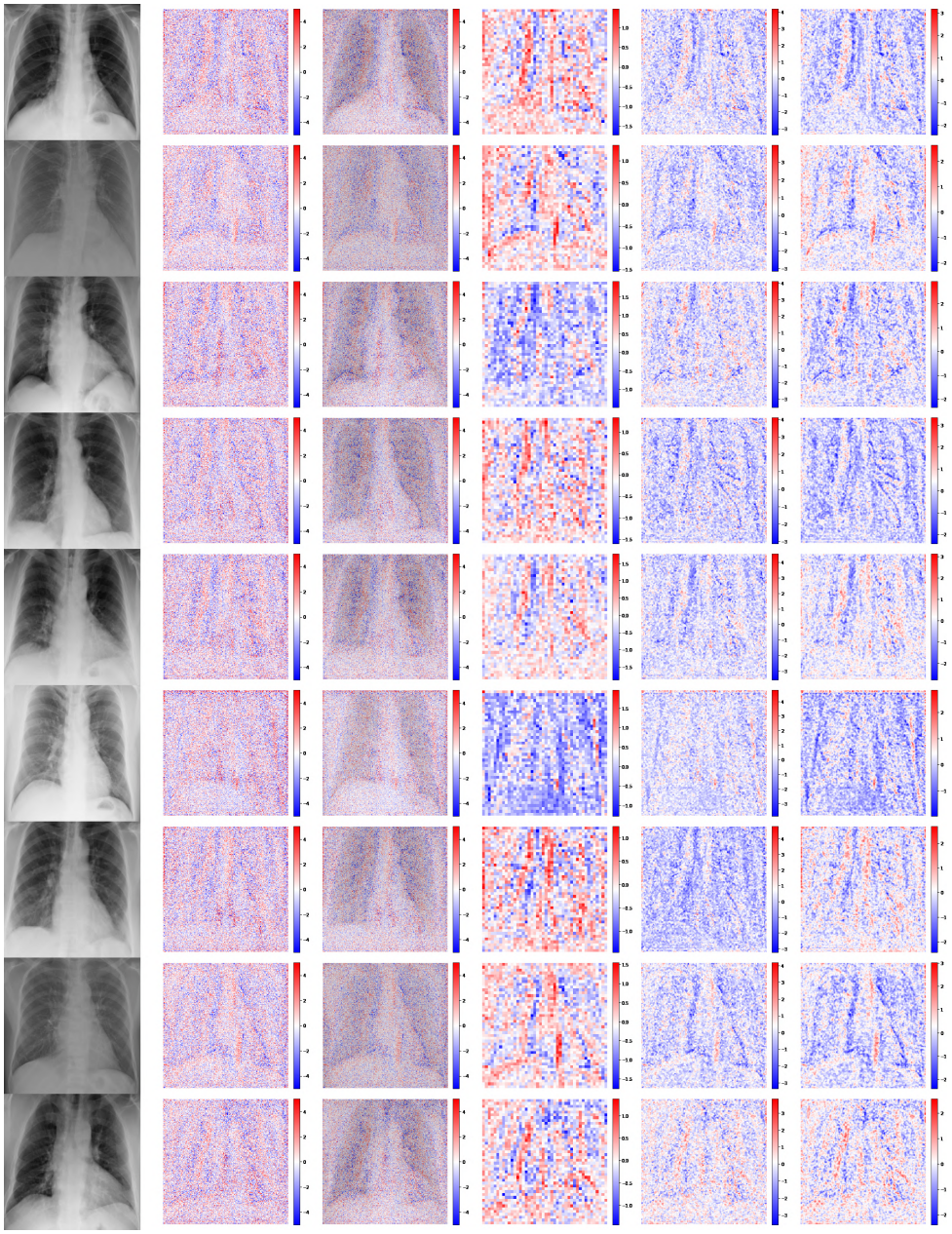


図 5.20: 心肥大データセット 敵対的サンプルの解析: 「心肥大」画像への適用

心肥大データセットにおける「心肥大」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

5.4　他の説明性手法との比較

ここでは敵対的サンプルと Grad-CAM, Guided Backpropagation, CycleGAN と比較した結果を示す. 敵対的サンプルは, 敵対的サンプルそのものと, mean, median, gaussian の三つのフィルタを適用した画像の中で最も特徴の可視性が高いものを比較対象とする. 塵肺データセットでは mean フィルタ, 心肥大データセットでは gaussian フィルタを選出した.

塵肺データセットでの比較を図 5.21, 図 5.22 に示す. 塵肺データセットにおける比較では, 敵対的サンプル, Grad-CAM, Guided Backpropagation の三つ手法の間で注目領域は大まかに一致した. 一方, CycleGAN の比較では, 敵対的サンプルは肺野領域の境界に強く注目しているのに対して CycleGAN では肺野領域全体に注目しているという結果となった.

心肥大データセットでの比較を図 5.23, 図 5.24 に示す. 心肥大データセットにおける比較では, 敵対的サンプル, Grad-CAM, Guided Backpropagation, CycleGAN の全ての手法間で注目領域は大まかに一致した.

また, 全体の傾向として, Grad-CAM, Guided Backpropagation は注目領域と領域内の注目度合いの強弱を示すに留まったが, 敵対的サンプルと CycleGAN ではそれに加えて注目領域内で画素値が変化した.

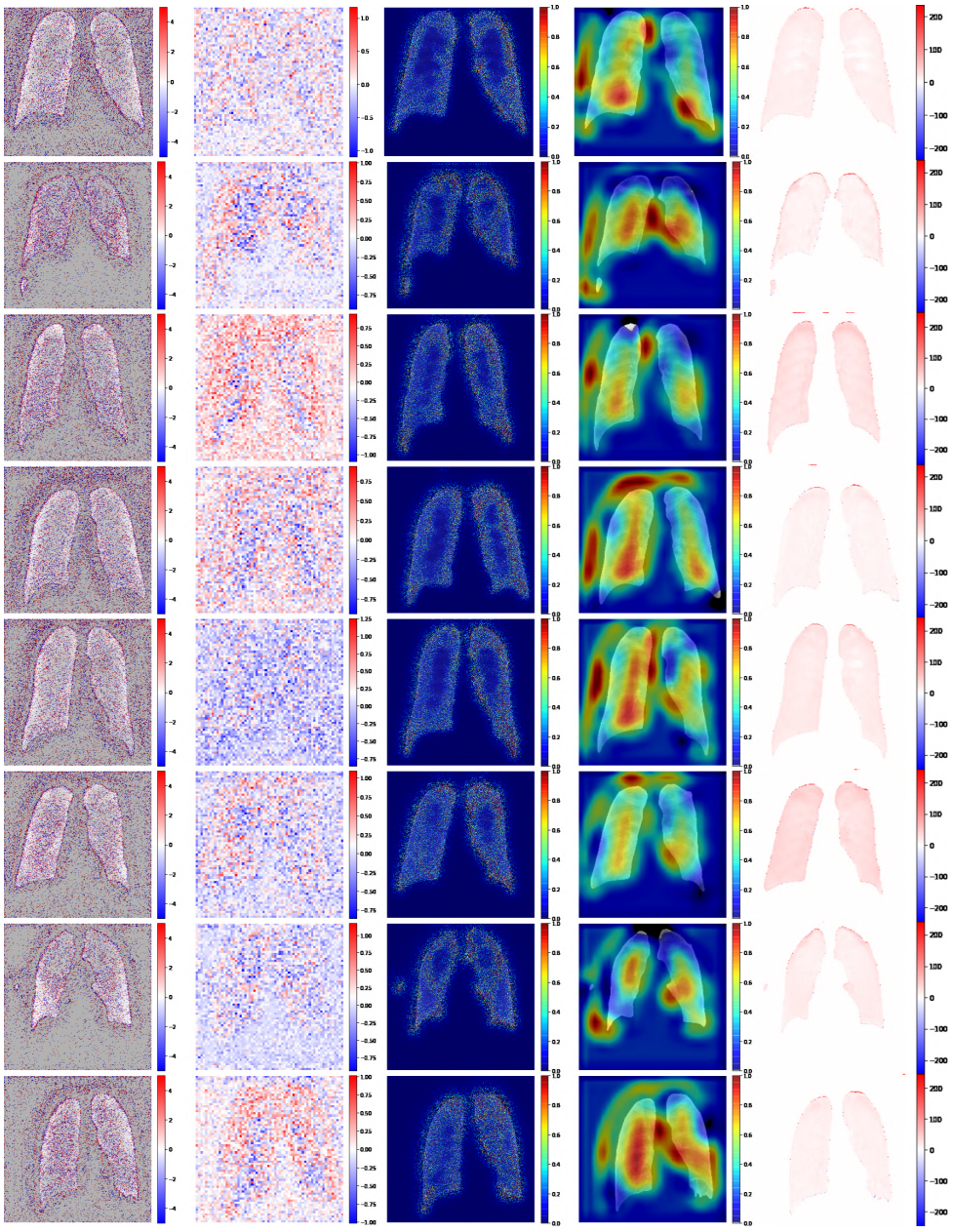


図 5.21: 塵肺データセット 敵対的サンプルと他の説明性手法との比較: 検出なし

(1) 敵対的サンプル, (2) 敵対的サンプル: mean フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

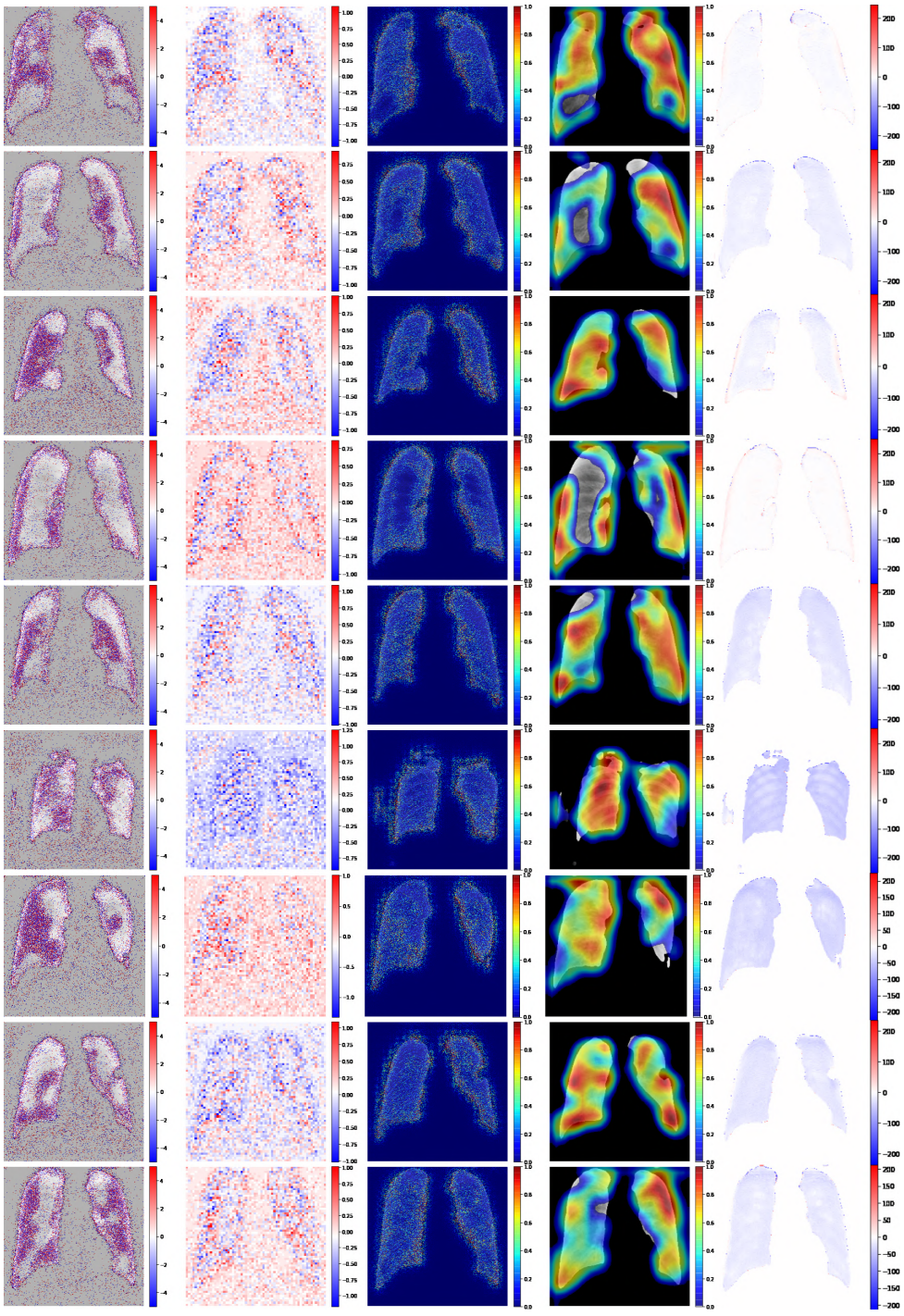


図 5.22: 塵肺データセット 敵対的サンプルと他の説明性手法との比較: 塵肺

(1) 敵対的サンプル, (2) 敵対的サンプル: mean フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

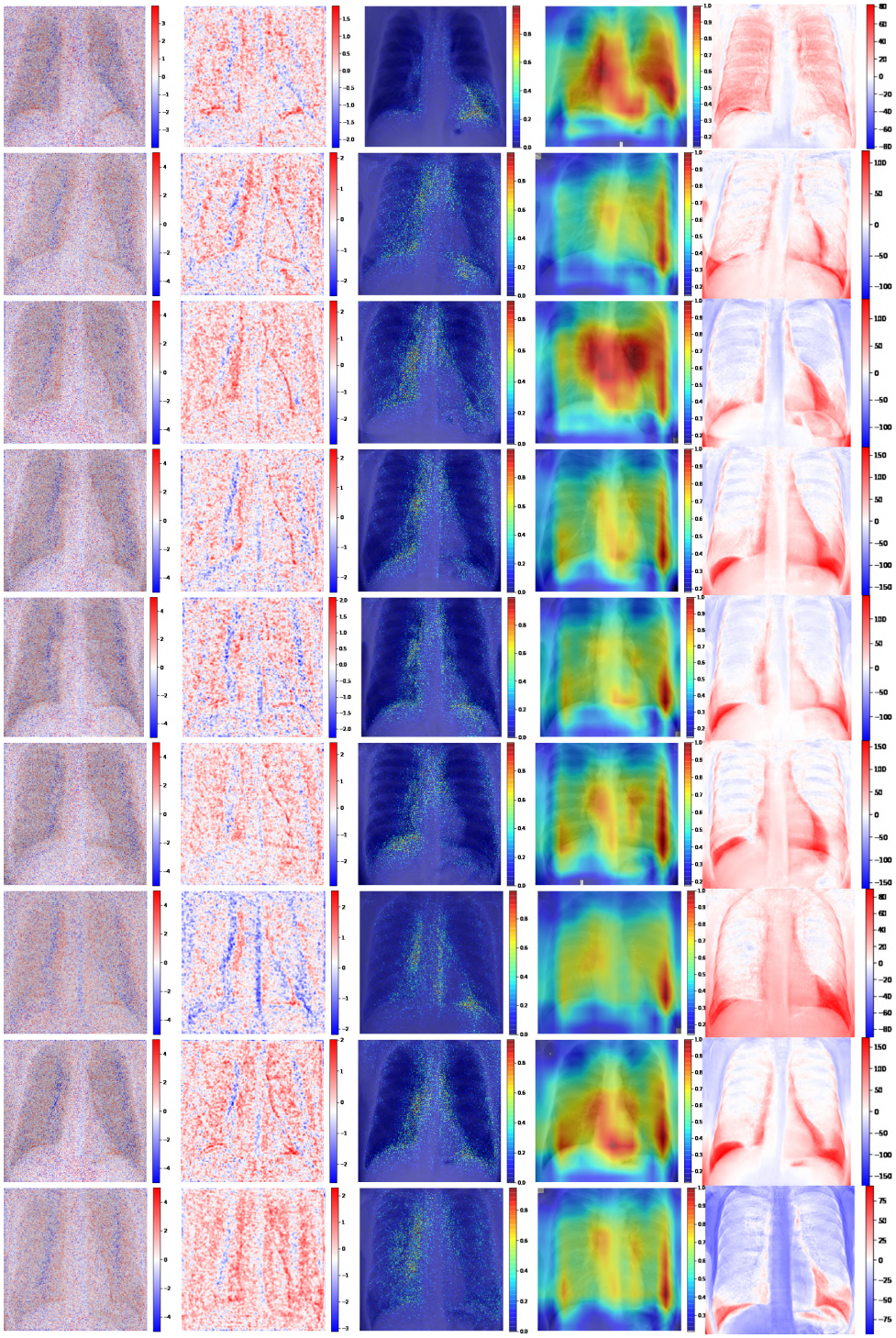


図 5.23: 心肥大データセット 敵対的サンプルと他の説明性手法との比較: 検出なし

(1) 敵対的サンプル, (2) 敵対的サンプル: gaussian フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

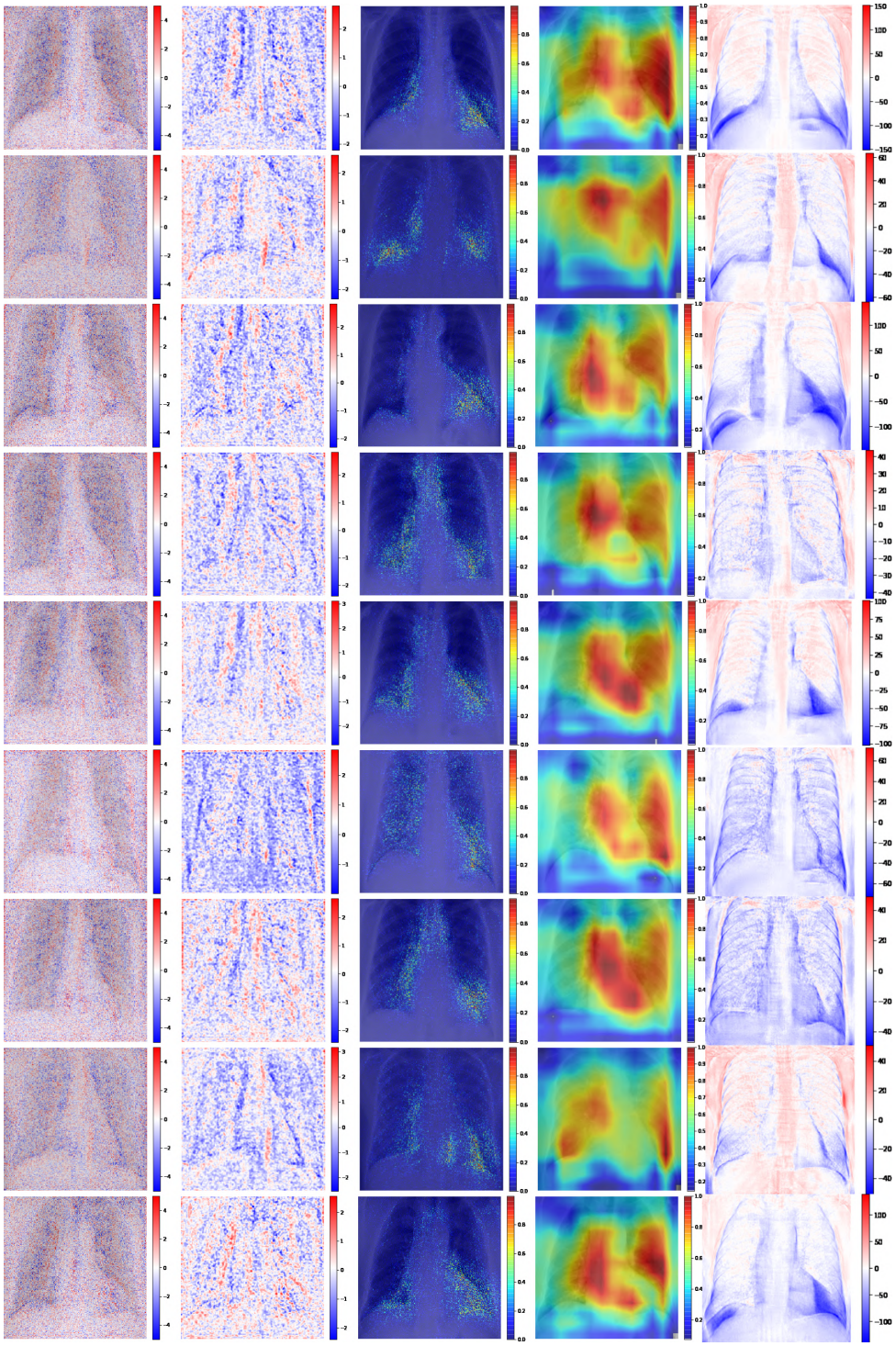


図 5.24: 心肥大データセット 敵対的サンプルと他の説明性手法との比較: 心肥大

(1) 敵対的サンプル, (2) 敵対的サンプル: gaussian フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

第 6 章

追加実験

塵肺データセットに説明可能な手法を適用した結果, CNN が肺野領域の境界のコントラストを分類基準の一つとして使用していることが示唆された. これに対し, この結果は肺の繊維化によるすりガラス陰影の発生という塵肺症の特徴とは一致せず, CNN が肺全体形状や大きさを基に「塵肺」または「検出なし」を判定している可能性が考えられる. そこで, 追加実験として, 塵肺分類モデルに使用した塵肺データセットの train データ 160 枚について, 「塵肺」と「検出なし」の各群における肺の大きさを統計的に比較し, 有意な差があるかどうかを検証する.

6.1 追加実験内容

塵肺データセットの train データ 160 枚を使用して, 肺野領域の面積を算出する. 算出した面積の平均と中央値を比較するとともに, t 検定を適用して「塵肺」と「検出なし」の肺の大きさに有意な差が存在するか検証する.

6.1.1 データセット

塵肺データセットの train データ 160 枚を使用する. データの分布を表 6.1 に示す.

表 6.1: 追加実験データセット

塵肺	検出なし	画像サイズ
80	80	512×512

6.2　追加実験結果

肺野領域の面積の平均と中央値を算出した結果を表に示す．「塵肺」画像の面積の平均は 83302.54, 中央値は 82880.50 となった．一方で，「検出なし」画像の面積の平均は 76907.95, 中央値は 77367.00 となった．

表 6.2: 肺野領域の面積の平均・中央値

	平均	中央値
塵肺	83302.54	82880.50
検出なし	76907.95	77367.00

また, t 検定を行った結果を示す．p 値が約 0.0026 であり, 通常の有意水準である 0.05 よりも小さいため, 帰無仮説（塵肺と検出なしで肺の大きさに差がない）を棄却することができる．この結果から，「塵肺」の肺が「検出なし」の肺に比べて有意に大きいことを示している．肺野領域の面積のヒストグラムと箱ひげ図を図 6.1 に示す．

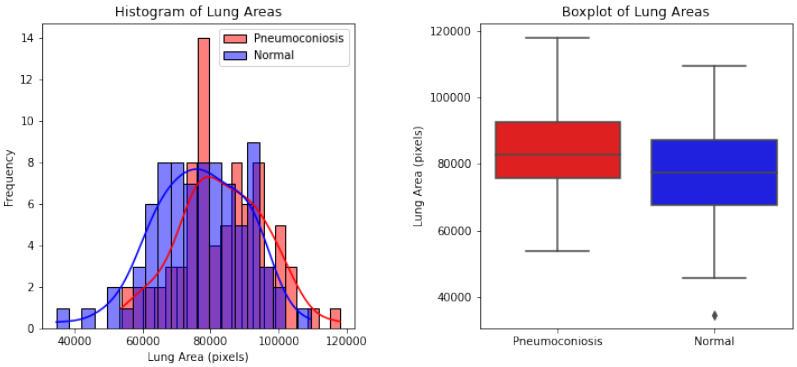


図 6.1: t 検定: ヒストグラム, 箱ひげ図

第 7 章

考察

結果から, 敵対的サンプルによる注目領域と Grad-CAM, Guided Backpropagation による注目領域は一致しており, 特に肺野境界や心臓側面といった部位が共通して強調されるという点で一貫性が認められた. このことから, 敵対的サンプルによって可視化された特徴は分類モデルが「検出なし」と「症状あり」を区別する際に重要視している領域を可視化できていると言える. 敵対的サンプルの解析結果から, 塵肺データセットでは肺野境界の画素値を減少させるような摂動を加えると「検出なし」, 画素値を増加させるような摂動を加えると「塵肺」と誤分類される傾向があることから, 肺野領域の境界のコントラストが分類基準の一つとなっていると考えられる. この結果から, CNN は肺の大きさで「検出なし」か「塵肺」かを分類している可能性が示唆された. これに対し, この結果は肺の繊維化によるすりガラス陰影の発生という塵肺症の特徴とは一致せず, CNN が肺全体形状や大きさを基に「塵肺」または「検出なし」を判定している可能性が考えられる. そこで両者の肺の大きさを統計的に比較したところ, 「塵肺」の肺の方が「検出なし」の肺よりも優位に大きいことが明らかになった. 今回の塵肺データセットの画像は複数の機関から収集されており, 「塵肺」および「検出なし」の画像はそれぞれ特定の機関に偏っていることが確認されている. そのため, 撮影環境や使用機器の違いなどが影響し, 肺領域の大きさに差が生じた可能性が考えられる. 心肥大データセットにおいても心臓側面の画素値を減少させるような摂動を加えると「検出なし」, 増加させるような摂動を加えると「心肥大」と分類される傾向があることから, 心臓の形状や明るさが分類に重要な役割を果たしていると考えられる. このことから, 敵対的サンプルは分類境界の変化に着目することで, 特定の画素値がどのように分類結果に影響を与えるかを定量的に分析可能だと考えられる. 一方, Grad-CAM や Guided

Backpropagation は注目領域を特定することはできるが, 注目領域とその強弱を示すに留まるため, どのような特徴が決定的に分類を変化させるかまでは示すことが難しい.

また, 敵対的サンプルと CycleGAN を比較した場合, どちらも注目領域の特定に加えて, 分類の違いにおける特徴の差を視覚化出来ている. このことから CycleGAN も敵対的サンプルと同様に分類に寄与する特徴を解析できる可能性があると言える. しかし, CycleGAN による変化後の画像を分類した際の精度を見ると, 塵肺データセットでは 42.85%と低く, 心肥大データセットでは 77.02%と比較的高い. これは、塵肺画像の変換がより難しく, CycleGAN が分類に影響を与える特徴を十分に捉えきれていない可能性を示唆している. また, 図 5.7 と図 5.10 の混同行列を見てみると, 塵肺, 心肥大ともに「検出なし」→「症状あり」の変換はうまく判定が変化しておらず不安定だと言える. これらのことから, CycleGAN は変換後の画像が必ずしも変換前の画像の反対クラスに分類されるわけではなく, 特徴の変化と分類の関係が不明瞭になる可能性がある. 一方で, 敵対的サンプルを用いたアプローチは, 変換後の画像が確実に誤分類されるよう設計されており, その際にどの特徴が変化したのかを追跡できる. 以上のことから, 敵対的サンプルを用いたアプローチは, 特定の特徴が分類結果に与える影響をより厳密に検証できるため, CycleGAN よりも高い信頼性を持つ説明手法であると考えられる.

第 8 章

結論

本研究では, CNN の分類過程の説明可能性の向上を目的として, 敵対的サンプルを用いた手法を提案し, 分類に寄与する領域とその領域内のパターンの違いの獲得を目指した. 塵肺データセット, 心肥大データセットに対して実験を行った結果, CNN は塵肺の分類には肺野領域の境界, 心肥大の分類には心臓部分の側面に注目している可能性が示唆された. また, 既存の手法と比較した結果から, 敵対的サンプルの注目領域は既存の説明性手法の注目領域と大まかに一致しており, 一貫性が認められた. さらには敵対的サンプルを用いたアプローチは, 変換された画像の判定が変化した上でその差を見ていることから, 敵対的サンプルから得られる特徴が分類に寄与していることを示す根拠の信頼性も高いことが確認された. 以上のことから, 敵対的サンプルを用いた手法は説明性の観点からみて有効だと判断する.

また, CNN が塵肺を判定する際に, 肺野領域の境界, つまりは肺の大きさに注目していることが示唆されましたが, これは塵肺の症状の特徴とは一致していない. そのため, 学習用データに偏りが生じていないかを調査するため, 肺野領域の大きさを統計的に検証した. その結果, 「塵肺」の肺が「検出なし」の肺に比べて有意に大きいことが示唆された. これは, 今回の塵肺データセットが複数の機関から収集されており, 「塵肺」および「検出なし」の画像がそれぞれ特定の機関に偏っていることが原因であると考えらる. したがって, 今後の研究では, 肺野領域の大きさが統一されたデータを使用して実験を行う必要があると考えられる.

総評として, 敵対的サンプルを用いた手法は, Grad-CAM や Guided Backpropagation による注目領域と一致しており, 一貫性が確認された. その上, Grad-CAM や Guided Backpropagation は注目領域を特定することはできるが, 注目領域とその強弱を示すに留ま

るため, どのような特徴が決定的に分類を変化させるかまでは示すことが難しいのに対して, 敵対的サンプルは分類境界の変化に着目することで, 特定の画素値がどのように分類結果に影響を与えるかを定量的に分析可能であると言える. このことから, 分類モデルが「検出なし」と「症状あり」を区別する際に重要視している領域を可視化できていると判断できるため, この手法は分類モデルのブラックボックス性を解決する上で有効な手段であると言える. 一方で, これは全ての説明可能な手法に共通する課題であるが, 実際に医療現場での導入を考える際には, 可視化された領域や特徴が医学的な「病変の特徴」と一致しているかどうかを医師の医学的知見と照らし合わせて検討する必要がある. そのため, 今後の研究においては, 医師の意見を取り入れた上で, さらに詳細な検討を行うことが求められると考えられる.

謝辞

本研究を進める上でご指導をしてくださった吉田真一教授に心から感謝致します．研究の方向性に悩んだ際や結果が出ずに不安になった際に吉田先生が助言をくださったお陰で修士論文の完成まで無事にたどり着くことができました．

また，副査を引き受けてくださった岩田誠教授と中原潔教授に感謝致します．本研究について頂いた忌憚のない意見は今後の研究に誠心誠意活かしていきたいと思います．

ここまで自分の研究を何かと気にかけてご助力して下さった研究室のメンバーにも感謝致します．研究室メンバーの親身な対応のおかげで豊富な知識を身につけながら順調に研究室活動を行うことができました．

最後にこれまでの大学生活を常に支えて続けて下さった両親に感謝致します．

参考文献

[1] Jun Gao, et al. “Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview.” *Mathematical Biosciences and Engineering* 16.6, 6536-6561, 2017.

[2] Bolei Zhou, et al. “Learning deep features for discriminative localization.” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[3] Tsutsui Yasuyuki, et al. “Analysis of Trained Convolutional Neural Network using Generative Adversarial Network.” *International Workshop on Advanced Computational Intelligence and Intelligent Informatics(IWACIII)*, Oct.31-Nov.3, 2021.

[4] Ramprasaath R. Selvaraju, et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.” *Proceedings of the IEEE international conference on computer vision,* 2017

[5] Jost Tobias Springenberg, et al. “Striving for Simplicity: The All Convolutional Net.” *International Conference on Learning Representations (ICLR)*, 2015

[6] Ian Goodfellow, et al. “Generative adversarial nets.” *Advances in Neural Information Processing Systems* 27, 2014.

[7] Jun-Yan Zhu, et al. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.” *The Institute of Electrical and Electronics Engineers(IEEE) International Conference on Computer Vision (ICCV)*, 2017.

[8] Han Zhang, et al. “Self-Attention Generative Adversarial Networks”, PMLR, 2019

[9] Ian J. Goodfellow, et al. “Explaining and harnessing adversarial examples.” *International Conference on Learning Representations (ICLR)*, 2015.

[10] Aleksander Madry, et al. “Towards deep learning models resistant to adversarial attacks.” *International Conference on Learning Representations (ICLR)*, 2018.

[11] Kang-Hee Lee, et al. “A Development and Validation of an AI Model for Cardiomegaly Detection in Chest X-rays.” *Applied Sciences* 14(17):7465, 2024.

[12] Fan Yang, et al. “Pneumoconiosis computer aided diagnosis system based on X-rays and deep learning.” *BMC medical imaging* 21, 1-7, 2021.

[13] Stefan Jaeger, et al. “Automatic tuberculosis screening using chest radiographs.” *IEEE transactions on medical imaging* 33.2, 233-245, 2013.

[14] Sema Candemir, et al. “Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration.” *IEEE transactions on medical imaging* 33.2, 577-590, 2013.

[15] Liuzhuo Zhang, et al. “A deep learning-based model for screening and staging pneumoconiosis.” *Scientific reports*, 11(1):1–7, 2021.

[16] Xiaosong Wang, et al. “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases.” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106, 2017.

付録 A

敵対的サンプルの解析結果

A.1　塵肺データセット

A.1.1　「検出なし」

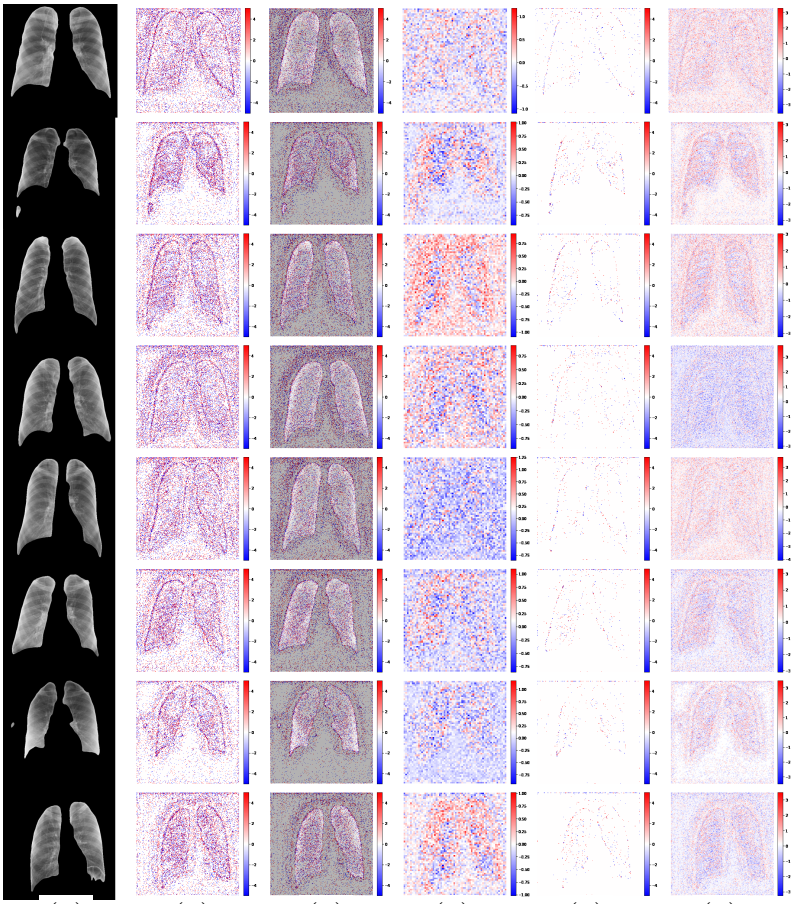


図 A.1: 塵肺データセット 敵対的サンプルの解析: 「検出なし」画像への適用 その 1

塵肺データセットにおける「検出なし」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

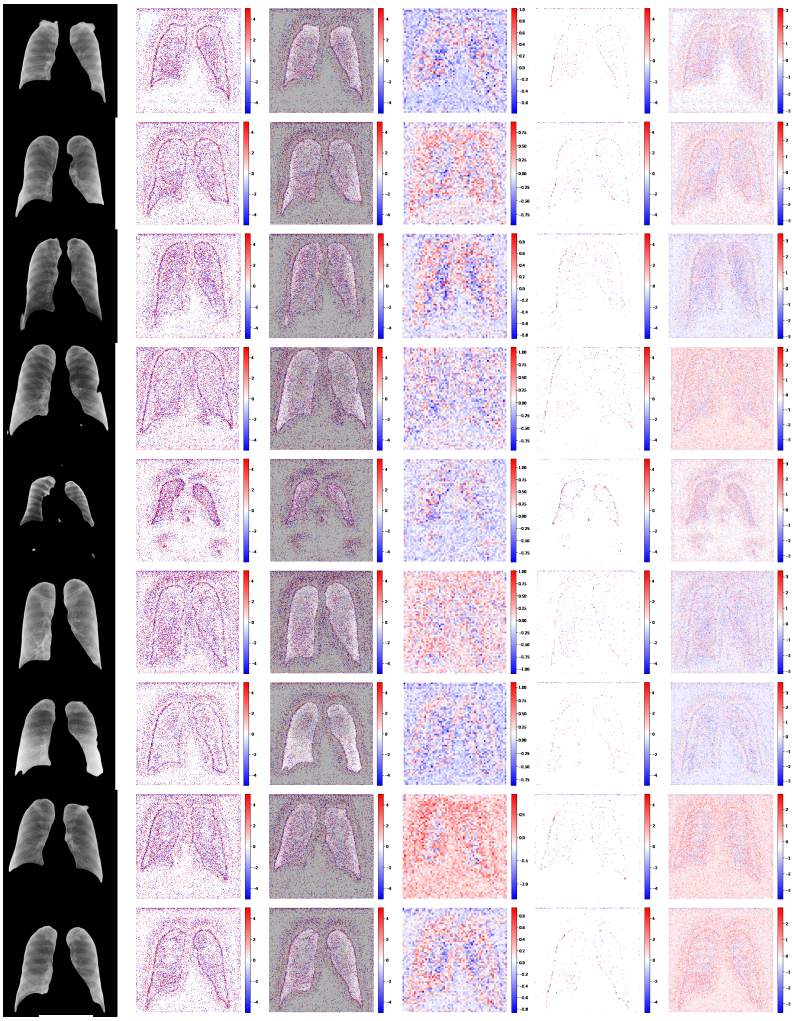


図 A.2: 塵肺データセット 敵対的サンプルの解析: 「検出なし」画像への適用 その 2

塵肺データセットにおける「検出なし」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

A.1.2 「塵肺」

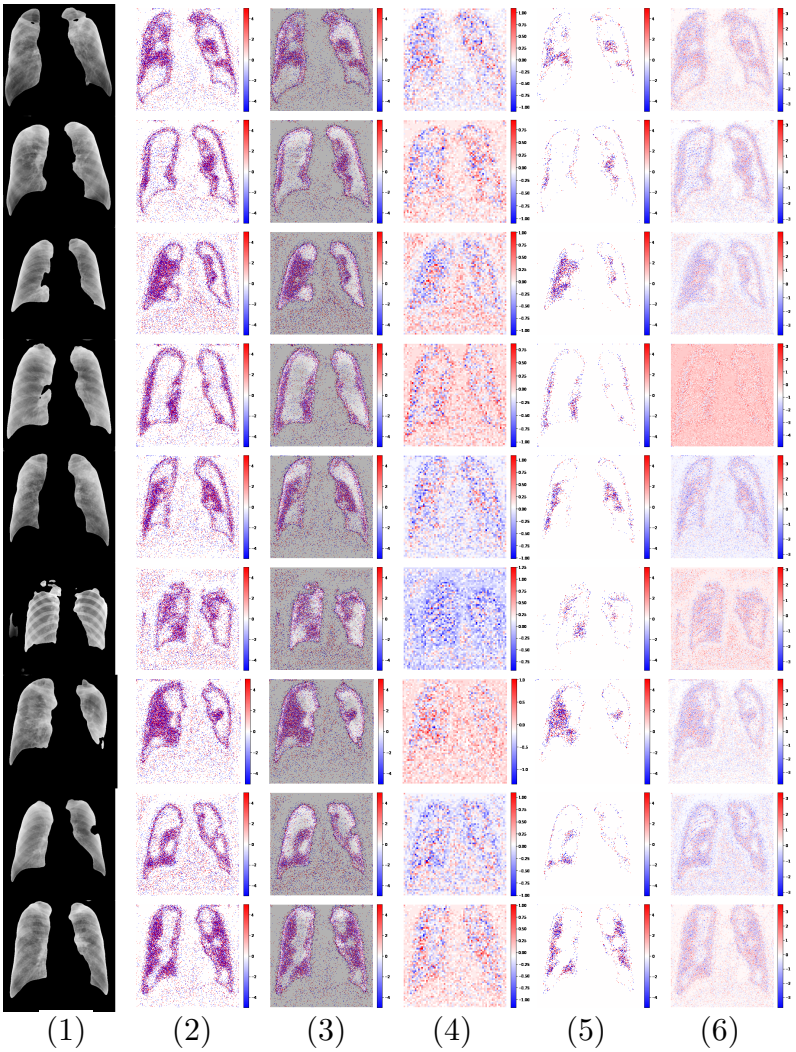


図 A.3: 塵肺データセット 敵対的サンプルの解析: 「塵肺」画像への適用 その 1

塵肺データセットにおける「塵肺」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

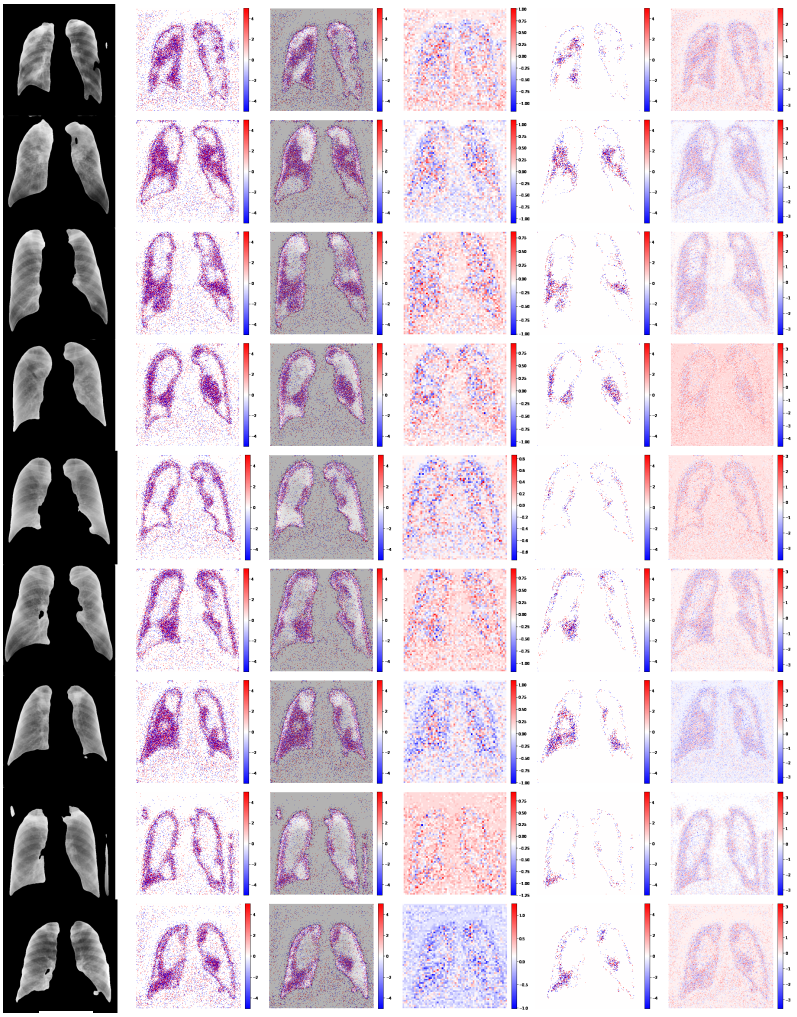


図 A.4: 塵肺データセット 敵対的サンプルの解析: 「塵肺」画像への適用 その 2

塵肺データセットにおける「塵肺」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

A.2 心肥大データセット

A.2.1 「検出なし」

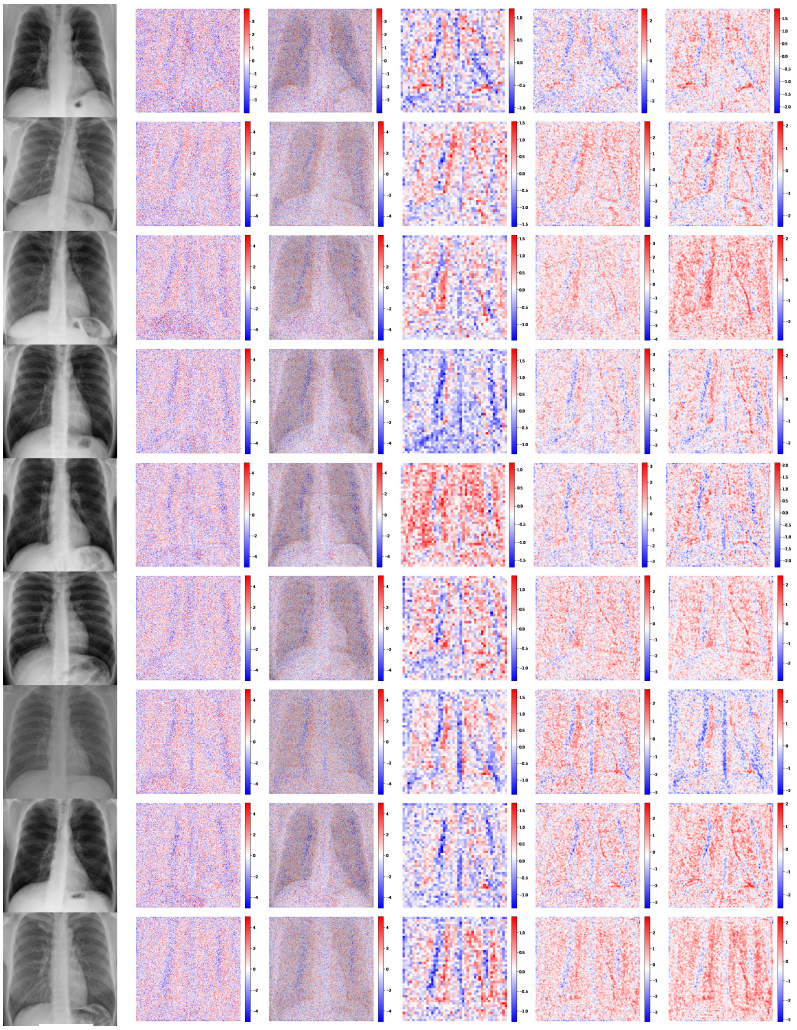


図 A.5: 心肥大データセット 敵対的サンプルの解析: 「検出なし」画像への適用 その 1

心肥大データセットにおける「検出なし」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

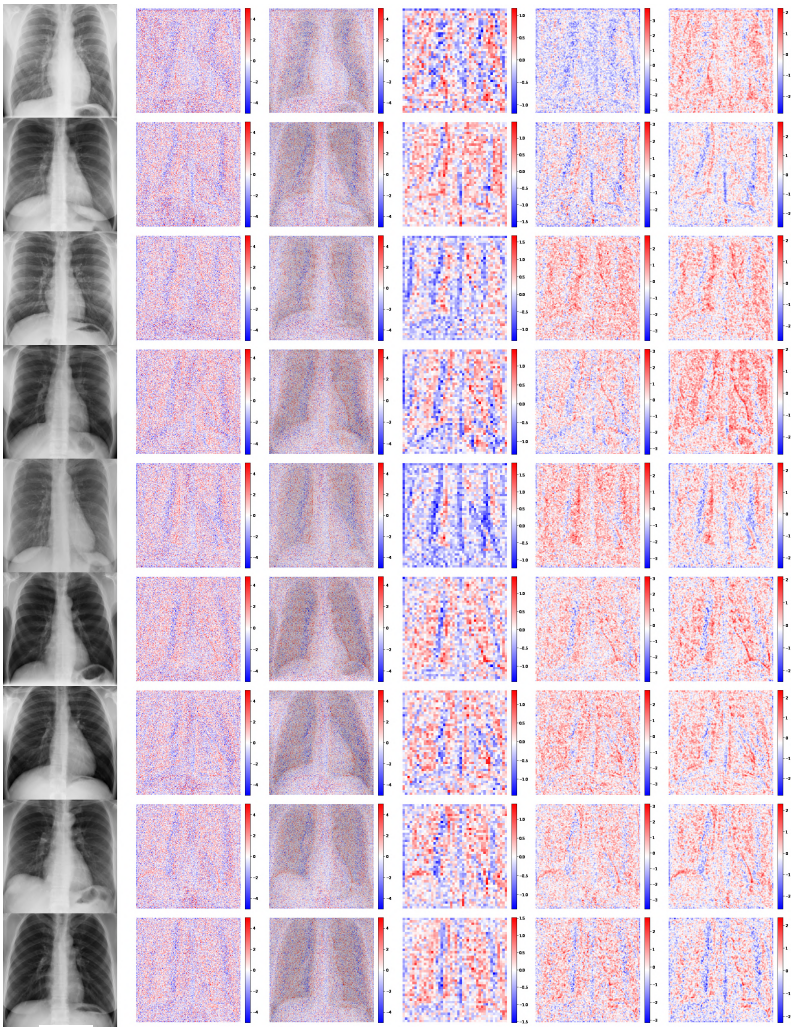


図 A.6: 心肥大データセット 敵対的サンプルの解析: 「検出なし」画像への適用 その 2

心肥大データセットにおける「検出なし」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

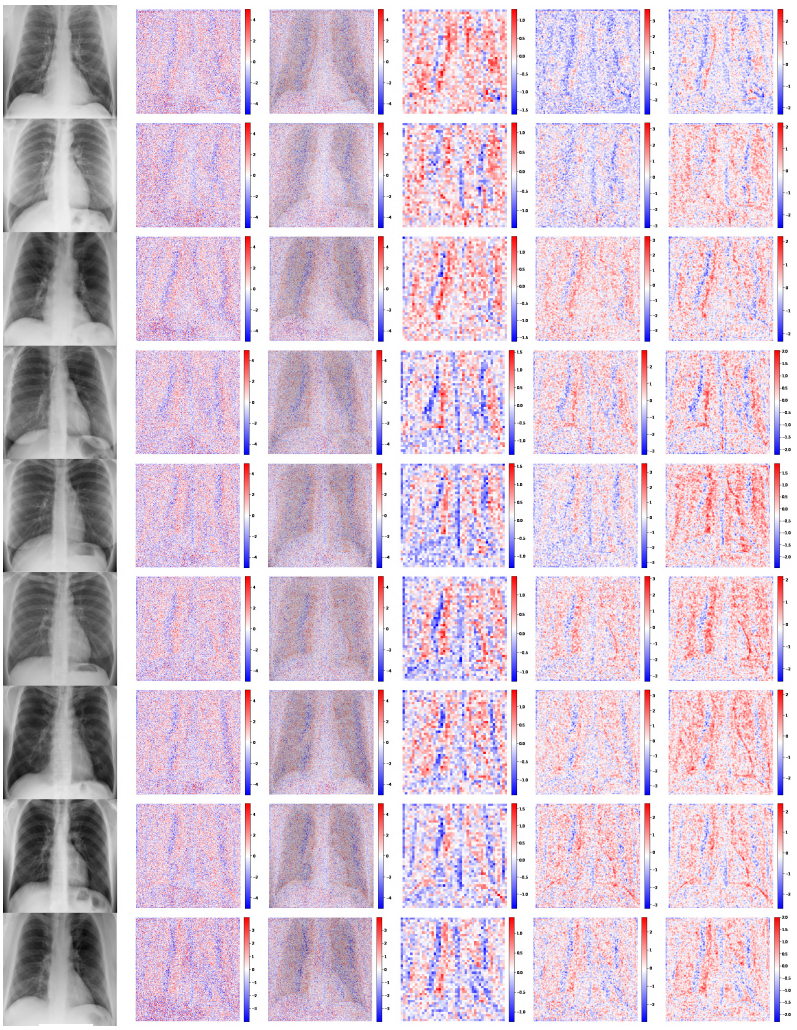


図 A.7: 心肥大データセット 敵対的サンプルの解析: 「検出なし」画像への適用 その 3

心肥大データセットにおける「検出なし」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

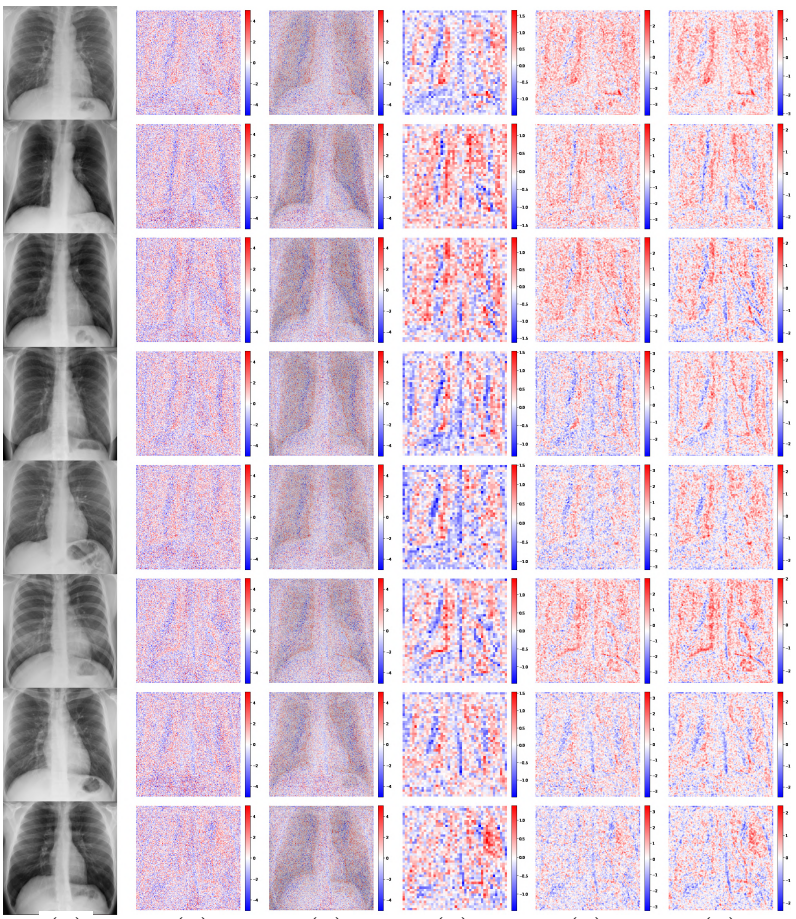


図 A.8: 心肥大データセット 敵対的サンプルの解析: 「検出なし」画像への適用 その 4

心肥大データセットにおける「検出なし」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

A.2.2 「心肥大」

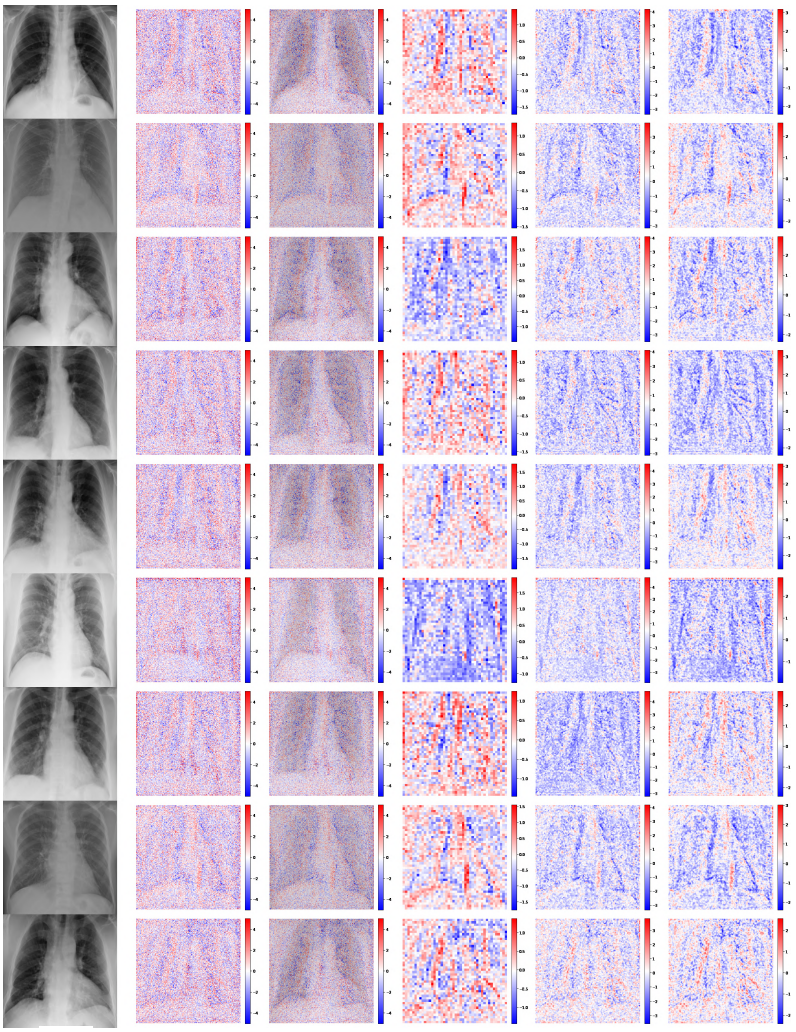


図 A.9: 心肥大データセット 敵対的サンプルの解析: 「心肥大」画像への適用 その 1

心肥大データセットにおける「心肥大」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

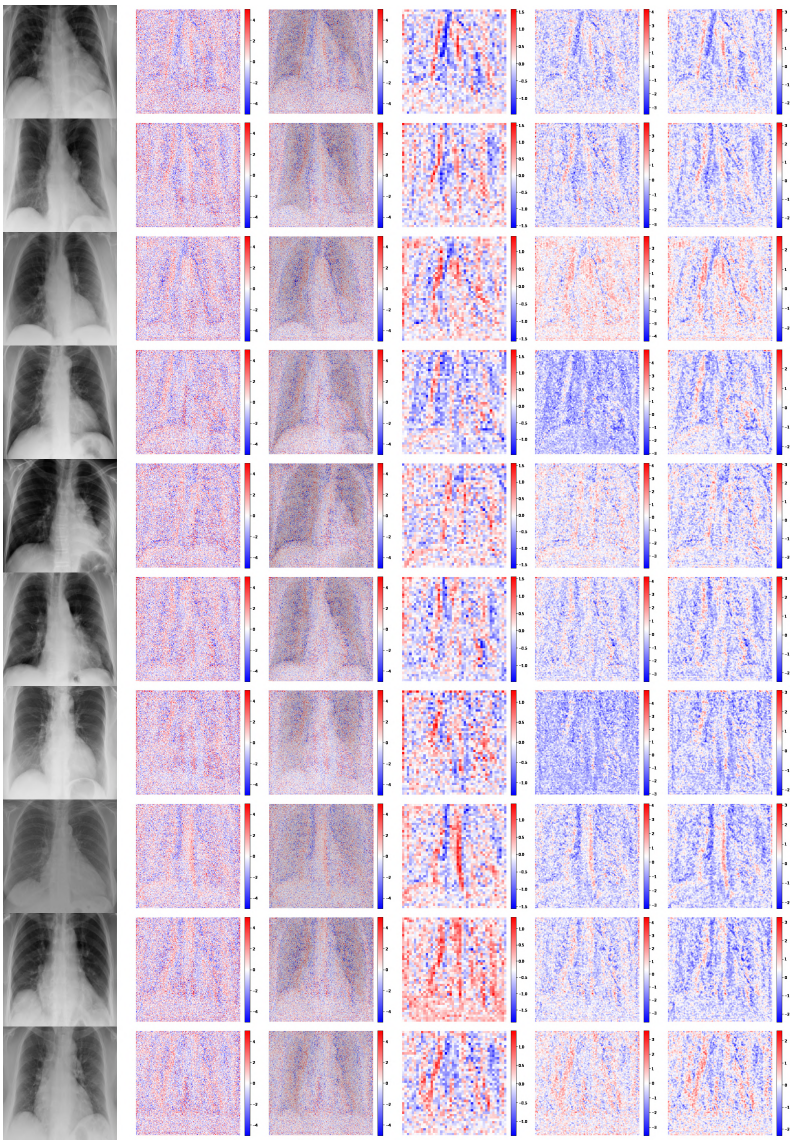


図 A.10: 心肥大データセット 敵対的サンプルの解析: 「心肥大」画像への適用 その 2

心肥大データセットにおける「心肥大」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

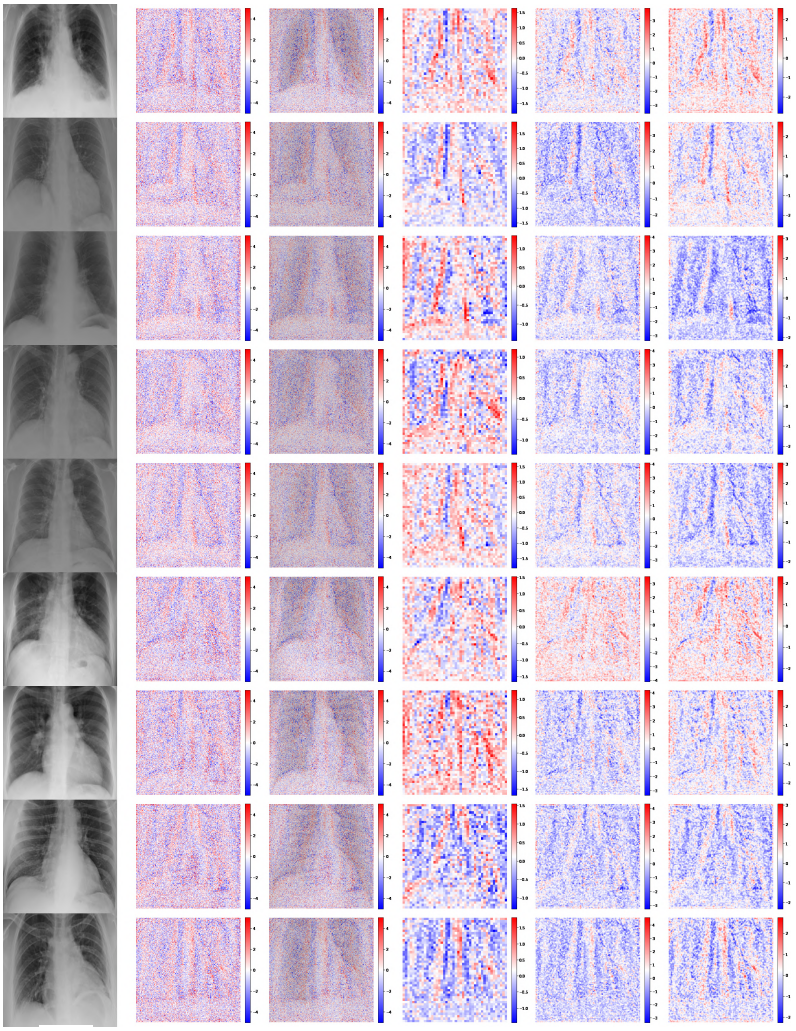


図 A.11: 心肥大データセット 敵対的サンプルの解析: 「心肥大」画像への適用 その 3

心肥大データセットにおける「心肥大」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

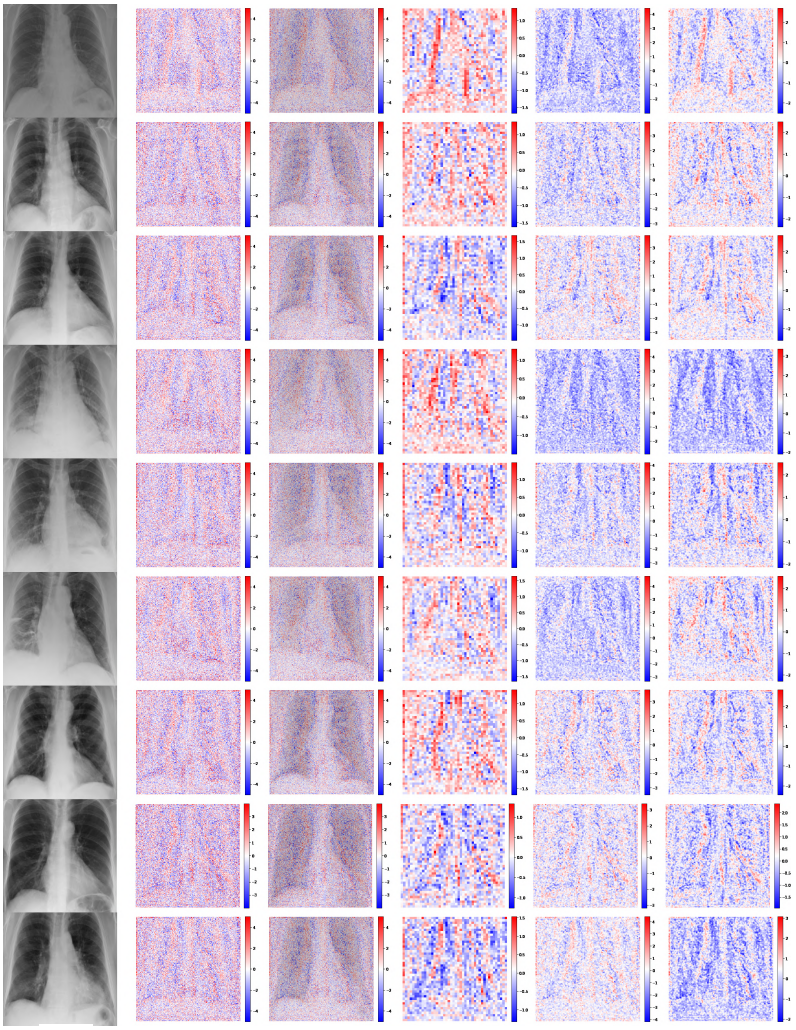


図 A.12: 心肥大データセット 敵対的サンプルの解析: 「心肥大」画像への適用 その 4

心肥大データセットにおける「心肥大」画像の敵対的サンプルの解析結果. (1) 入力画像, (2) 摂動, (3) 敵対的サンプル, (4) 摂動に 5 × 5mean フィルタを適用, (5)3 × 3median フィルタを適用, (6)gaussian フィルタを適用

付録 B

敵対的サンプルと他の手法の比較

B.1　塵肺データセット

B.1.1　「検出なし」

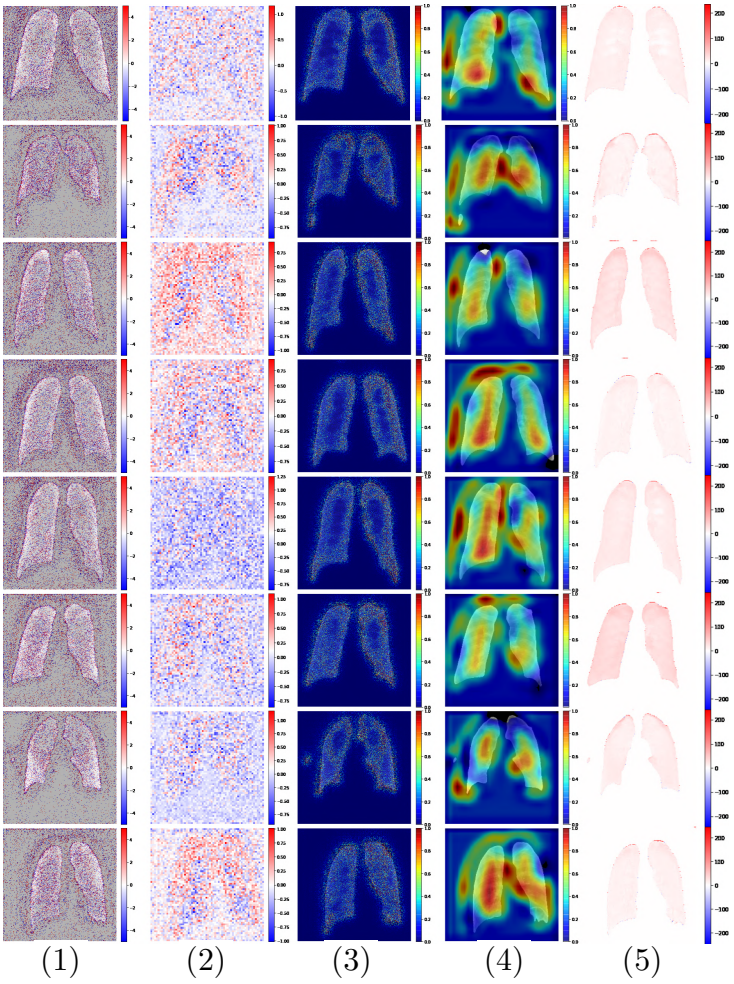


図 B.1: 塵肺データセット 敵対的サンプルと他の説明性手法との比較: 検出なし その 1

(1) 敵対的サンプル, (2) 敵対的サンプル: mean フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

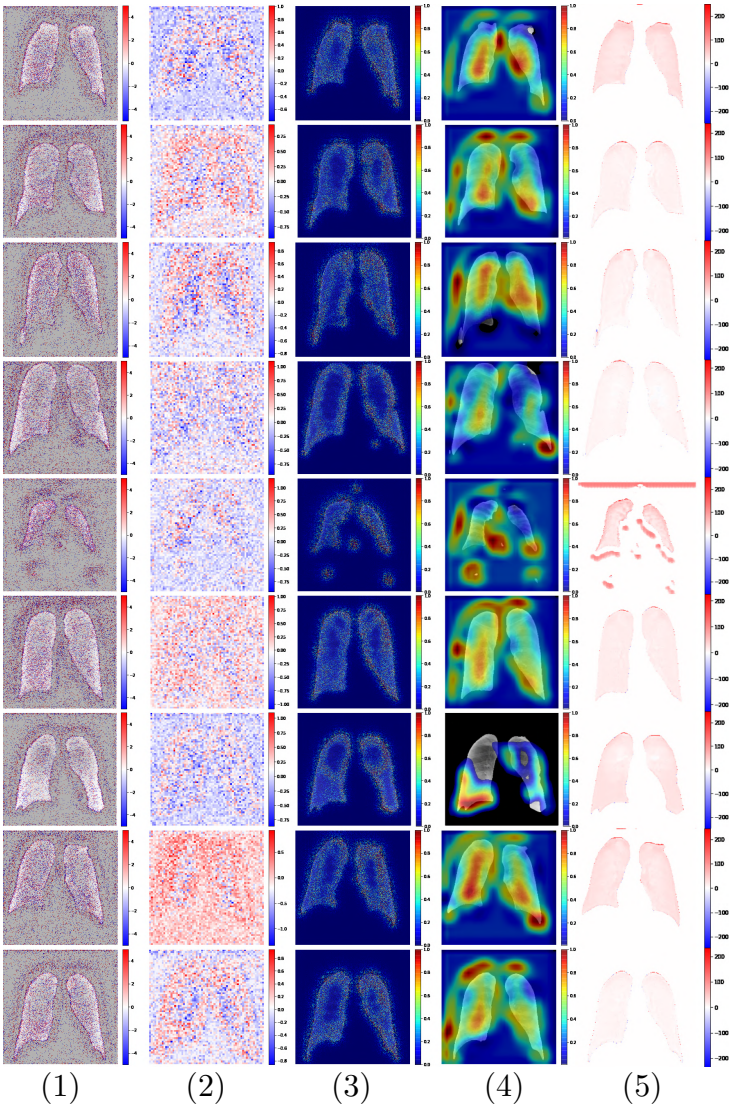


図 B.2: 塵肺データセット 敵対的サンプルと他の説明性手法との比較: 検出なし その 2

(1) 敵対的サンプル, (2) 敵対的サンプル: mean フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

B.1.2 「塵肺」

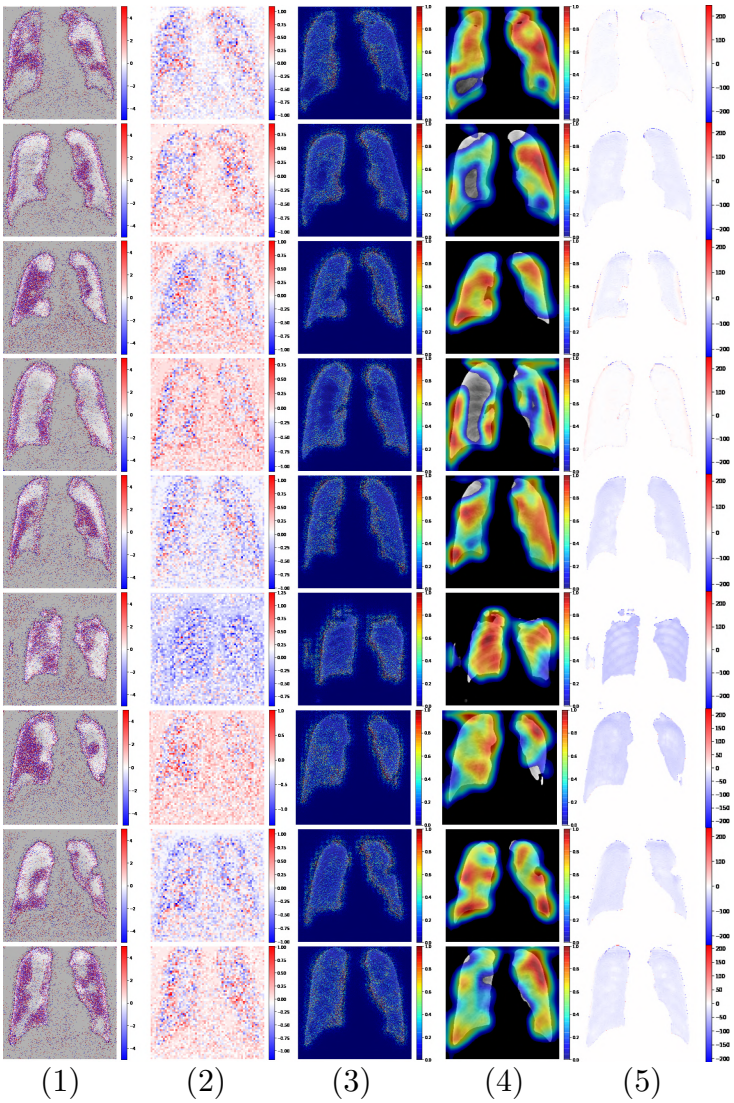


図 B.3: 塵肺データセット 敵対的サンプルと他の説明性手法との比較: 塵肺 その 1

(1) 敵対的サンプル, (2) 敵対的サンプル: mean フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

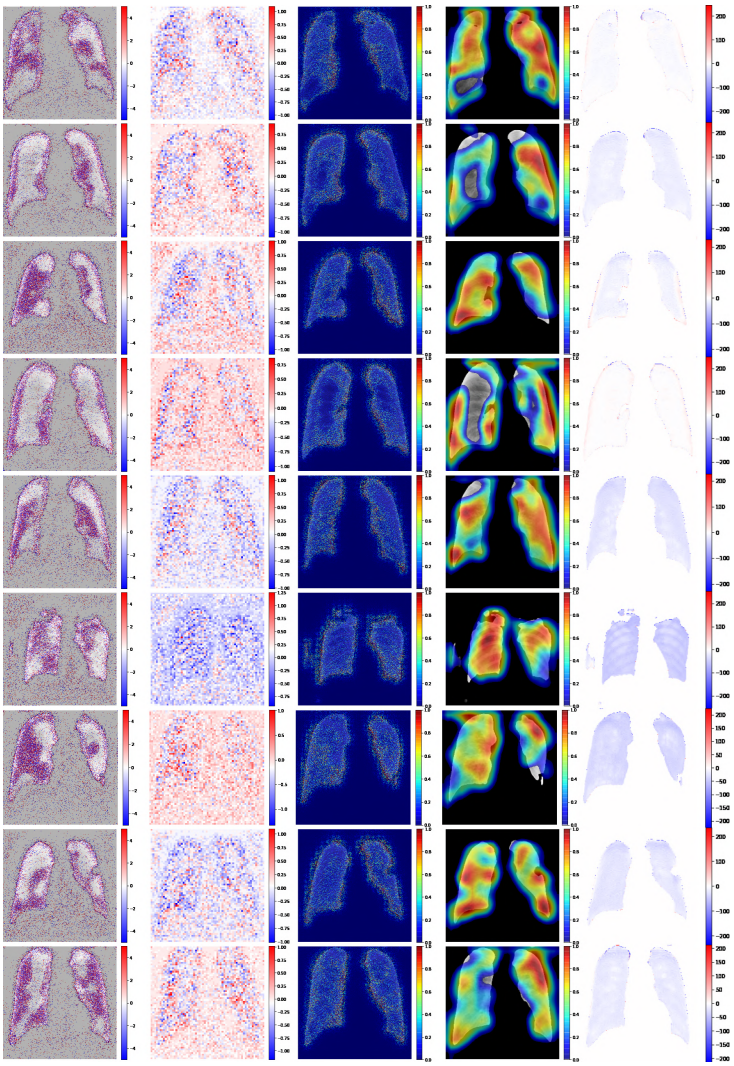


図 B.4: 塵肺データセット 敵対的サンプルと他の説明性手法との比較: 塵肺 その 2

(1) 敵対的サンプル, (2) 敵対的サンプル: mean フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

B.2 心肥大データセット

B.2.1 「検出なし」

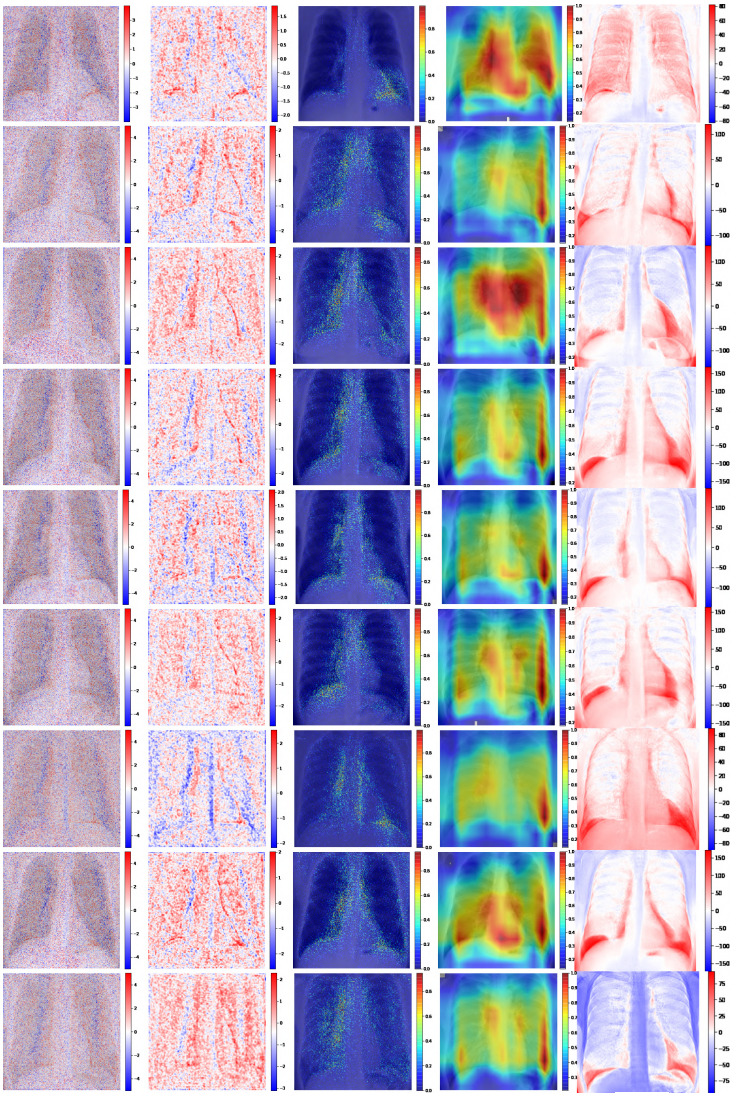


図 B.5: 心肥大データセット 敵対的サンプルと他の説明性手法との比較: 検出なし その 1

(1) 敵対的サンプル, (2) 敵対的サンプル: gaussian フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

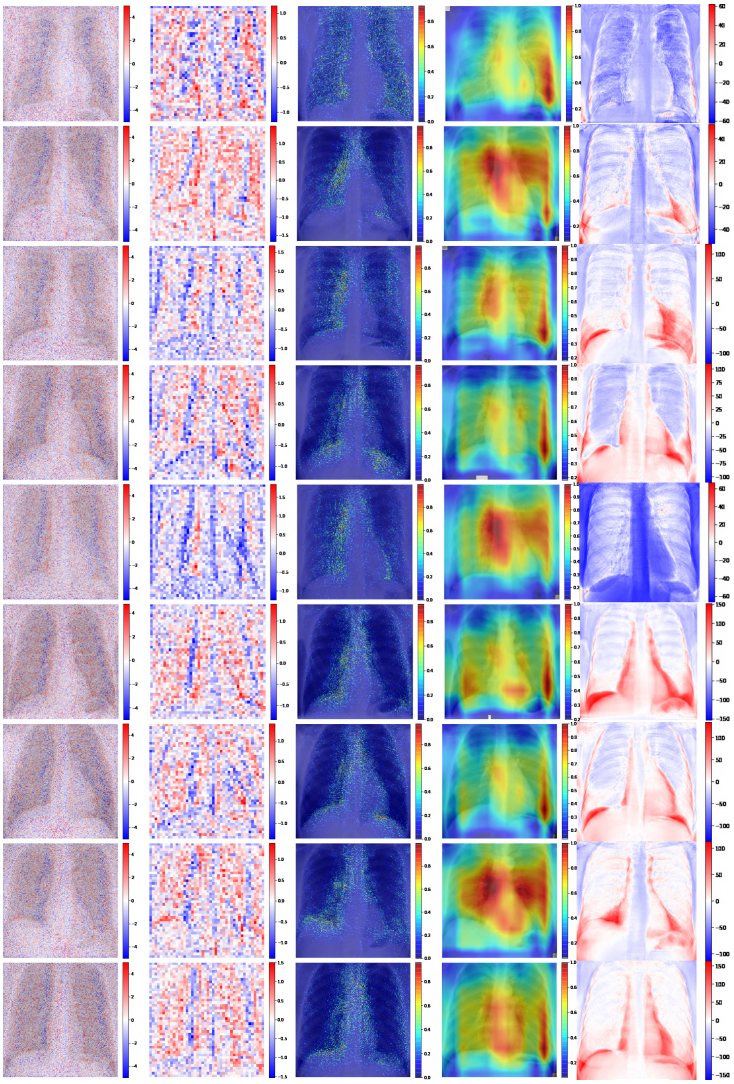


図 B.6: 心肥大データセット 敵対的サンプルと他の説明性手法との比較: 検出なし その 2

(1) 敵対的サンプル, (2) 敵対的サンプル: gaussian フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

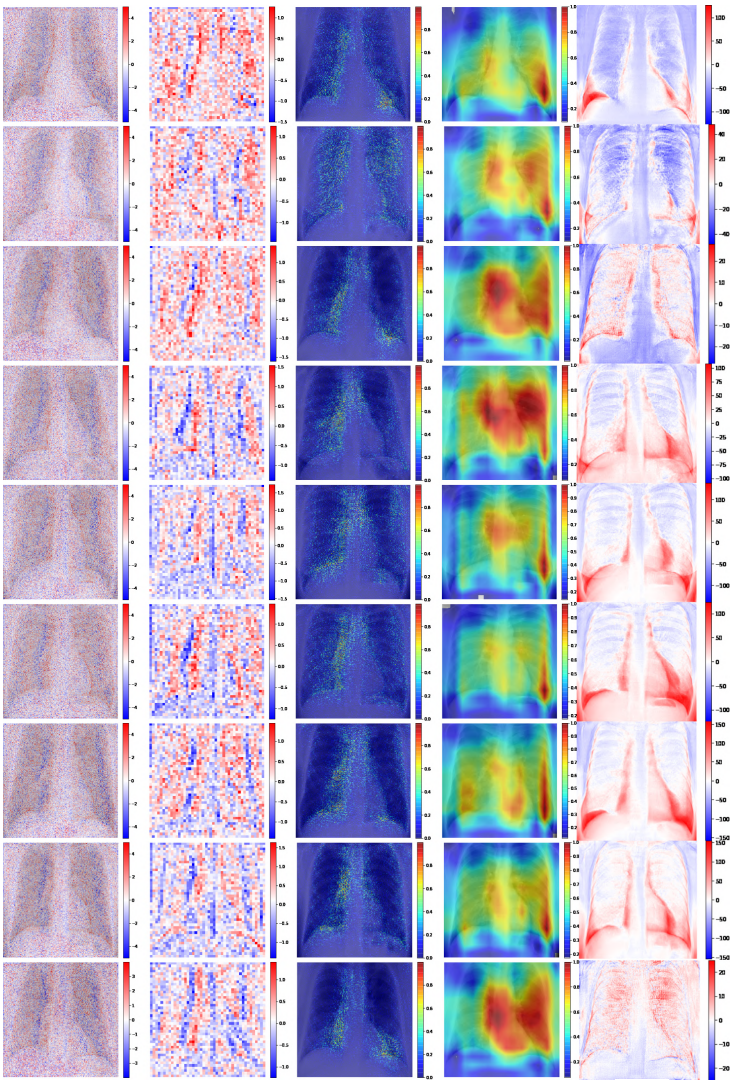


図 B.7: 心肥大データセット 敵対的サンプルと他の説明性手法との比較: 検出なし その 3

(1) 敵対的サンプル, (2) 敵対的サンプル: gaussian フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

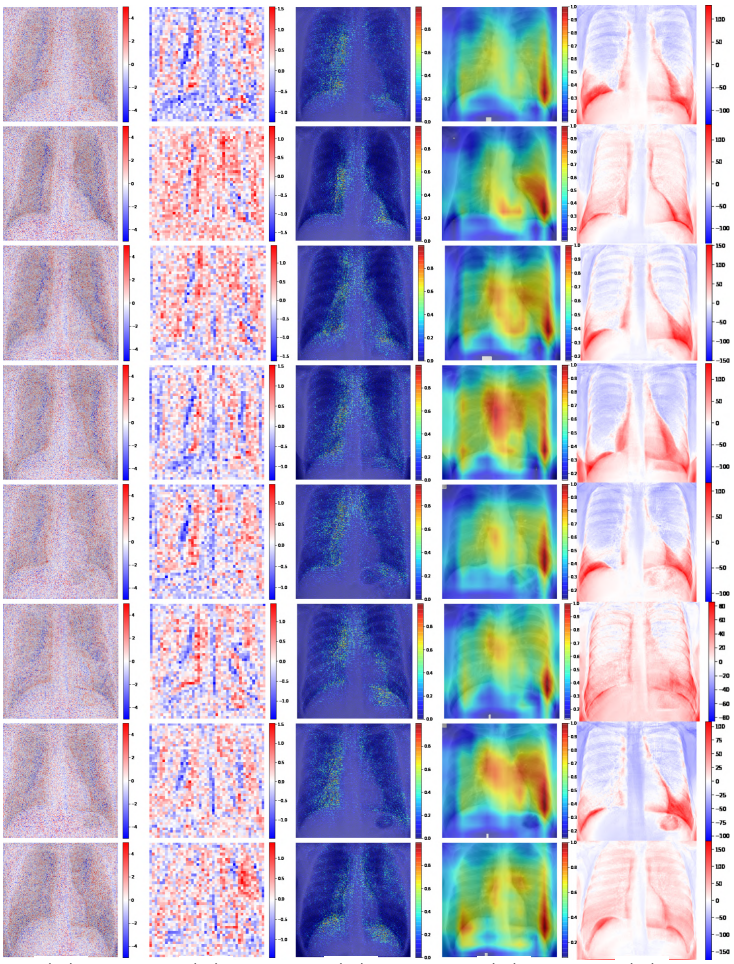


図 B.8: 心肥大データセット 敵対的サンプルと他の説明性手法との比較: 検出なし その 4

(1) 敵対的サンプル, (2) 敵対的サンプル: gaussian フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

B.2.2 「心肥大」

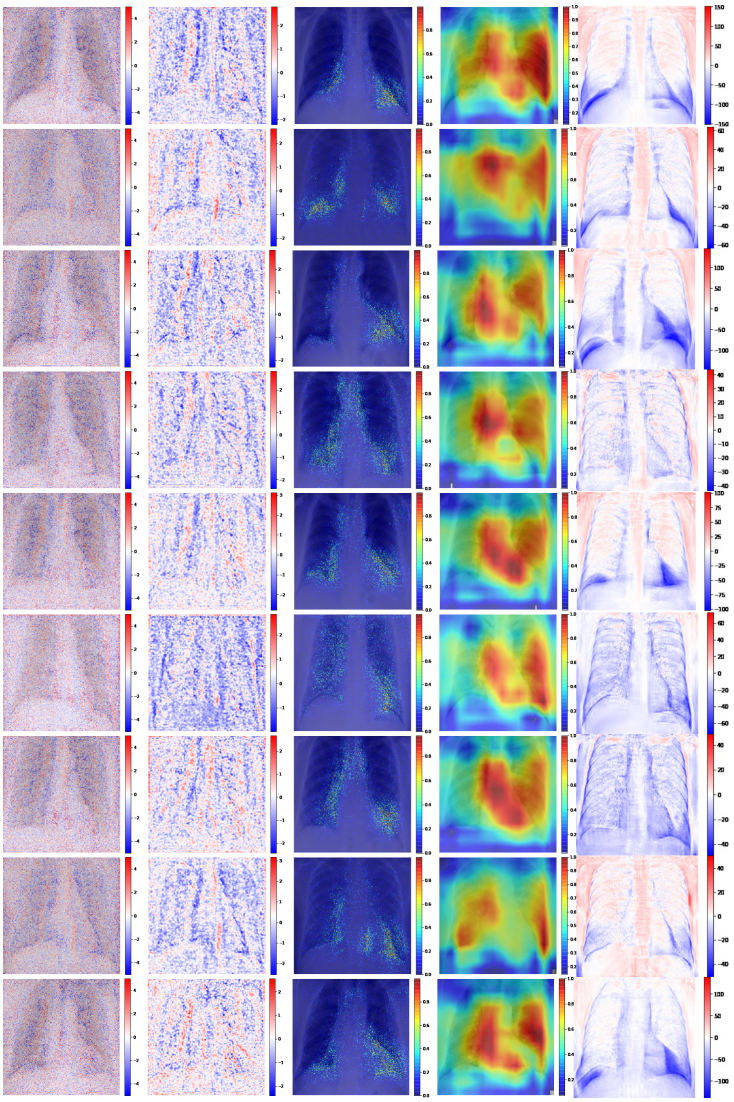


図 B.9: 心肥大データセット 敵対的サンプルと他の説明性手法との比較: 心肥大 その 1

(1) 敵対的サンプル, (2) 敵対的サンプル: gaussian フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

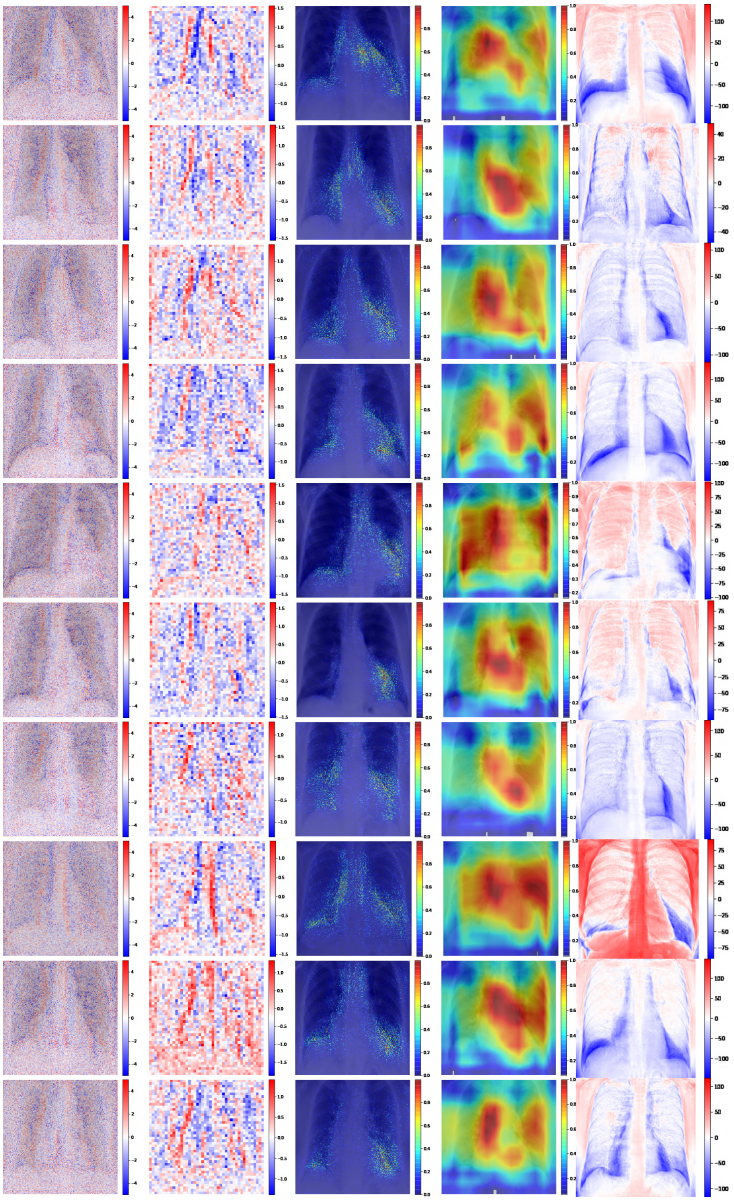


図 B.10: 心肥大データセット 敵対的サンプルと他の説明性手法との比較: 心肥大 その 2

(1) 敵対的サンプル, (2) 敵対的サンプル: gaussian フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

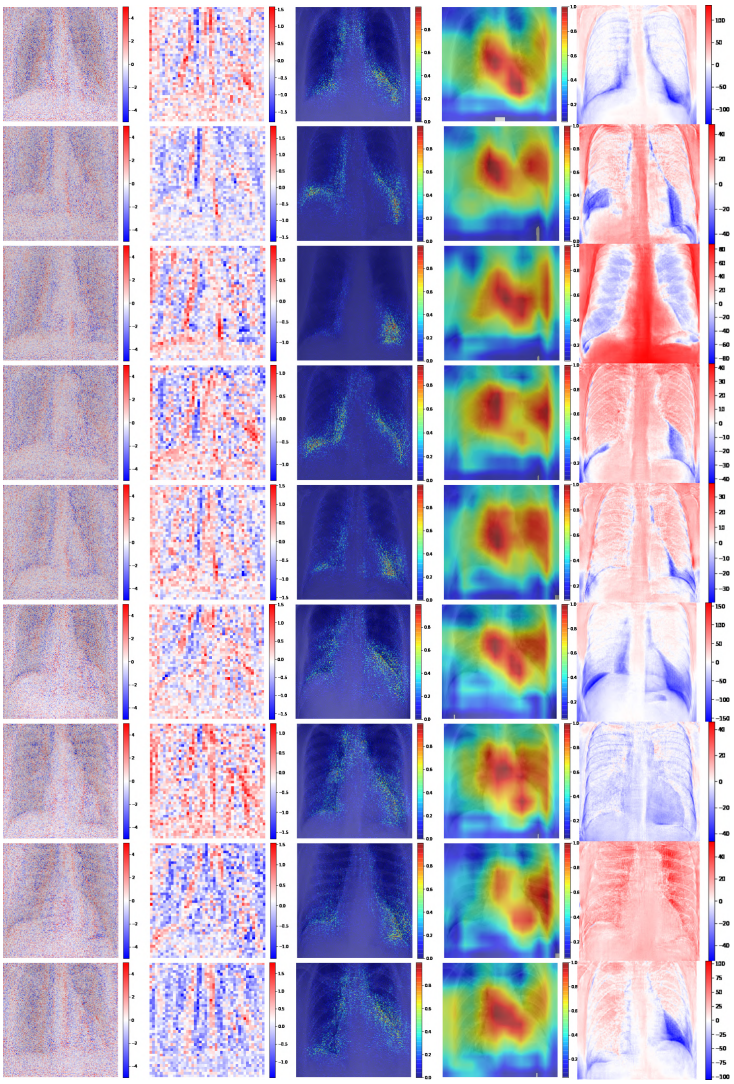


図 B.11: 心肥大データセット 敵対的サンプルと他の説明性手法との比較: 心肥大 その 3

(1) 敵対的サンプル, (2) 敵対的サンプル: gaussian フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN

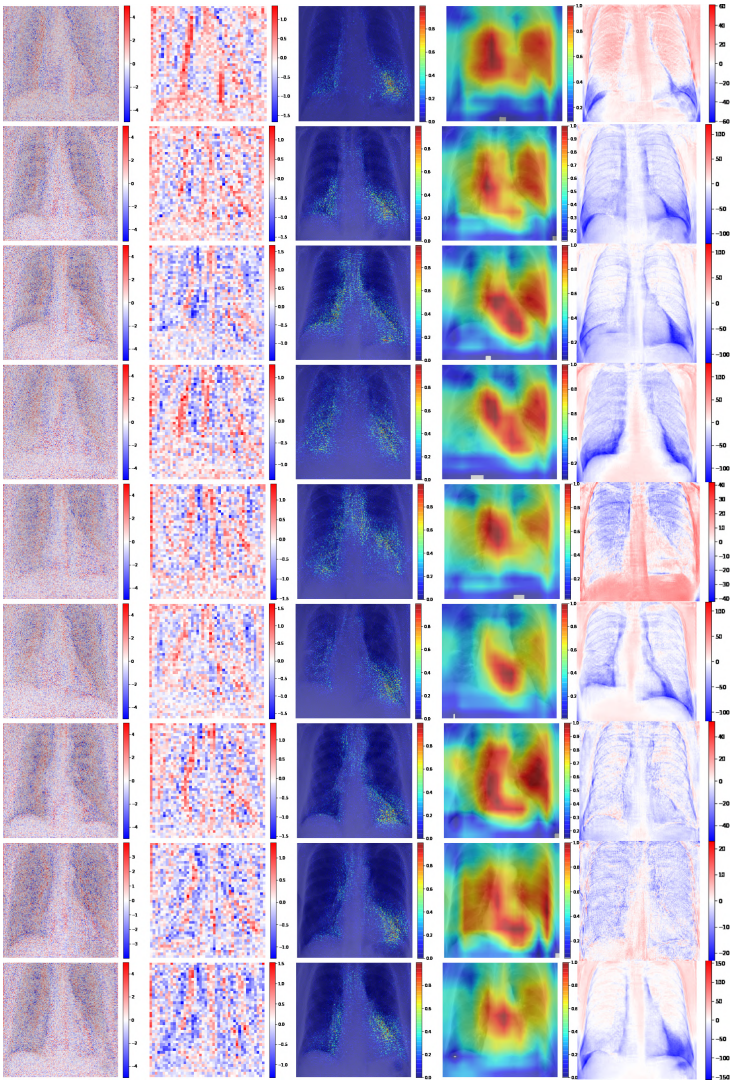


図 B.12: 心肥大データセット 敵対的サンプルと他の説明性手法との比較: 心肥大 その 4

(1) 敵対的サンプル, (2) 敵対的サンプル: gaussian フィルタ, (3)Guided Backpropagation, (4)Grad-CAM, (5)CycleGAN