令和 7 年度

修士学位論文

# 敵対的サンプルを用いた胸部 X 線画像分類モデルの説明性向上手法

## Explainability of Chest X-ray Image Classification Models Using Adversarial Examples

椎葉 啓介

指導教員　吉田 真一

2026 年 2 月 27 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

## 敵対的サンプルを用いた胸部 X 線画像分類モデルの説明性向上手法

椎葉 啓介

機械学習モデルにおいて，予測結果の根拠が不透明であるというブラックボックス性は，特に説明性が求められる医療分野では重要な課題である．またこの課題は，Grad-CAM などの可視化手法では，判断に寄与する特徴そのものを十分に解析できず根本的な解決に至っていない．そこで，CNN に対し CycleGAN や敵対的サンプル (Adversarial Examples) を用い，分類に重要な領域や形状及びパターンの特定で説明性の向上を図った研究が行われている．しかし，複数手法の敵対的サンプルで比較した説明性の検討はされていない．そこで本研究は，7 つの手法について敵対的サンプルを作成し，新たな説明性が得られるか検証する．CNN および ViT を対象とし，敵対的摂動が予測結果に与える影響を比較・分析することで，各モデルが利用している判別根拠を明らかにする．また，特に強い摂動を加えた領域を可視化することで，モデルが学習した特徴の説明性を獲得する．

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

# Abstract

## Explainability of Chest X-ray Image Classification Models Using Adversarial Examples

SHIIBA, Keisuke

In machine learning models, the black-box nature where the basis for prediction results is opaque is a significant challenge, particularly in the medical field where explainability is crucial. Furthermore, visualization techniques like Grad-CAM have not fundamentally solved this problem, as they cannot sufficiently analyze the features contributing to the decision. Consequently, research has been conducted using CycleGAN and adversarial examples on CNNs to improve explainability by identifying regions, shapes, and patterns crucial for classification. However, no studies have compared the explainability achieved using adversarial examples from multiple methods. This research creates adversarial samples for seven methods to verify whether new explainability can be obtained. By targeting CNNs and ViTs and comparing/analyzing the impact of adversarial perturbations on prediction results, it clarifies the discriminative basis each model utilizes. Furthermore, by visualizing regions subjected to particularly strong perturbations, it acquires the explainability of the features learned by the model.

***key words*** Convolutional Neural Network(CNN), Vision Transformer(ViT), Adversarial Example, Explainable Artificial Intelligence(XAI)

# 目次

# 図目次

# 表目次

# 第 1 章

# 序論

深層学習を用いた医用画像解析は急速に発展しており，胸部 X 線画像分類においても高精度な診断支援モデルが数多く提案されている．従来は畳み込みニューラルネットワーク (CNN) が主流であったが，近年では Vision Transformer(ViT)[1] に代表されるトランスフォーマーベースのモデルも導入され，広域的な文脈情報を捉えた高い性能が報告されている．このようにモデル構造が多様化・高度化する一方で，いずれの手法においても，予測結果の根拠が不透明であるという説明性の問題が依然として残されている．特に医療分野では，実用化のため診断結果に対する信頼性や説明可能性が必要不可欠である．この課題に対し，Grad-CAM[2] などの可視化手法を用いて，モデルが注目する画像領域を提示する研究が行われてきた．しかし，これらの手法はモデルの内部表現を間接的に可視化するに留まり，判断に寄与する特徴そのものを十分に解析できない．そこで，中嶋ら [3] は深層学習による医用画像解析の説明性に関し，Grad-CAM よりも説明性のある手法として，CNN に対し CycleGAN[4] や敵対的サンプル (Adversarial Examples)[5] を用い，分類に重要な領域や形状及びパターンの特定で説明性の向上を図っている．敵対的サンプルとは，モデルに誤分類させることを目的として小さな摂動 (ノイズ) を画像などに加えたものである．中嶋らは，敵対的サンプルによって元画像からどのような変更が生じたかを分析することで，モデルの判断根拠を捉えるという観点に基づいていた．しかし，PGD[6] によって生成された敵対的サンプルのみを対象として検討を行っており，その他の敵対的サンプル生成手法については検証されていなかった．敵対的サンプルは，手法ごとに摂動の生成原理や特徴が異なるため，PGD 以外の手法を用いた場合には，モデルが依存する特徴や判断基準を異なる観点から捉えられる可能性がある．しかしながら，PGD 以外の敵対的サンプル生成手法を用い

た，説明性の比較・検討をした研究は十分になされていない．そこで本研究では，FGSM[7]，PGD，DeepFool[8]，FAB[9]，C&W[10]，UAP[11]，OPA[12] を用いて敵対的サンプルを生成し，各手法によって生じる画像変化の違いを分析することで，新たな説明性が得られるかを検証する．

# 第 2 章

# 関連研究

## 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（CNN）は，LeCun ら [13] による LeNet の提案以来，画像認識分野で広く用いられている深層学習モデルである．畳み込み層，活性化関数，プーリング層を階層的に積み重ねることで，画像から局所的なパターンを捉え，層を深めるごとに高次で抽象的な特徴を抽出する．AlexNet[14] や ResNet[15] の登場により，画像分類や物体検出において極めて高い性能を示してきた．

## 2.2 ViT (Vision Transformer)

ViT は，Dosovitskiy ら (2020 年)[1] によって提案された，自然言語処理分野で標準的となった Transformer 構造を画像認識に適用したモデルである．具体的には，入力画像を固定サイズのパッチ (例：16 × 16 ピクセル) に分割し，各パッチを平坦化した後に線形投影を行うことで「パッチ埋め込み (Patch Embeddings)」を生成する．これに，画像の空間的情報を保持するための位置エンコーディング（Position Embedding）と，分類タスク用の特殊な class トークンを付与し，Transformer Encoder へ入力する．ViT の最大の特徴は，CNN が持つ「局所性」や「移動不変性」といった画像特有の帰納バイアス (Inductive Bias) をあえて排除し，自己注意機構 (Self-Attention) によって画像全体の依存関係を初期層から直接学習する点にある．これにより，超大規模データセットで事前学習を行った場合，CNN

より高い性能を示す．しかし，帰納バイアスが欠如しているため，中規模データセットでは十分な精度を得るのが困難という課題があった．

## 2.3 DeiT (Data-efficient Image Transformers)

Touvron ら (2020 年)[16] によって提案された DeiT は，ViT が抱えるデータ効率の課題を解決する手法である．DeiT は，知識蒸留（Knowledge Distillation）を Transformer の学習プロセスに組み込み，中規模データセットを用いた学習であっても高い汎化性能を示す．また，DeiT は従来の class トークンに加え，新たに Distillation トークンが導入されている．このトークンは，CNN などを教師モデルとし，教師モデルが予測したラベルを模倣するように学習される．これにより，Transformer モデルはデータから直接特徴を抽出するだけでなく，教師モデルである CNN が持つ画像認識に適した帰納バイアスを間接的に継承する．その結果，DeiT は ViT と同等の軽量なアーキテクチャであるにも関わらず，従来の CNN に匹敵する学習効率と Transformer 特有の表現力を両立することが可能となった．

## 2.4 敵対的サンプル (Adversarial Example)

敵対的サンプルとは，入力データに対し，人間には知覚困難な微小な摂動を加えることで，深層学習モデルを誤分類させるよう作為的に生成されたデータである．Szegedy ら (2013 年)[5] によってその存在が指摘されて以来，モデルの脆弱性を測る指標となっている．これらは，モデルの勾配情報を利用する White-box 攻撃と，内部情報に依存しない Black-box 攻撃に大別される．

### 2.4.1 FGSM(Fast Gradient Sign Method)

Goodfellow ら (2014 年)[7] によって提案された FGSM は，勾配ベースの攻撃における最も基本的な手法である．モデルの損失関数の勾配を計算し，損失が増加する方向 (勾配の符号方向) へ一定の大きさ $\epsilon$ の摂動を加えることで敵対的サンプルを生成する．計算コストが

非常に低い一方，単一ステップの更新であるため，反復的な手法に比べると攻撃成功率は限定的である．

### 2.4.2 PGD(Projected Gradient Descent)

Madry ら (2017 年)[6] が提案した PGD は，FGSM を多段階の反復処理へと拡張した手法である．本手法では，各ステップにおいて損失関数の勾配方向に微小な摂動を加え，その結果が元の入力データから一定の範囲 ($\epsilon$-ball) を逸脱した場合，許容範囲内へ再度投影 (Projected) する操作を繰り返す．PGD は「局所的な最適解を探索する一階の攻撃 (First-order adversary)」の中で最も強力なものとみなされており，敵対的訓練 (Adversarial Training) における代表的な攻撃モデルとしても広く採用されている．

### 2.4.3 DeepFool

Moosavi-Dezfooli ら (2016 年)[8] によって提案された DeepFool は，決定境界までの最短距離を反復的に探索する手法である．ニューラルネットワークの分類器を各反復において線形近似し，現在の入力点から決定境界への直交射影を計算することで摂動を更新する．これにより，誤分類を引き起こすために必要な「最小の」摂動ノルムを求めることが可能である．そのため，攻撃そのものとしてだけでなく，モデルの堅牢性を定量的に評価する指標としても広く利用されている．

### 2.4.4 FAB(Fast Adaptive Boundary)

Croce と Hein(2020 年)[9] によって提案された FAB は，ハイパーパラメータ調整を不要とした適応的な攻撃手法である．DeepFool と同様に，決定境界までの最小距離を探索する．ただし，各反復において決定境界上の点をより正確に推定し，勾配の射影を用いて摂動を更新する．そのため，学習率などのパラメータ設定に依存せず，かつ反復回数を抑えながら C&W と同等以上の高品質 (最小ノルム) な敵対的サンプルを生成可能で，モデルの評価に

おいて信頼性の高いベンチマークとして機能する.

### 2.4.5 C&W(Carlini & Wagner)

Carlini と Wagner(2017 年)[10] による C&W 攻撃は，敵対的サンプルの生成を制約付き最適化問題として定式化した手法である．独自の目的関数を設計し，画像の画素値の範囲制約を変数変換によって処理することで，特定の距離尺度 ($L_0, L_2, L_\infty$) における摂動を最小化しつつ，極めて高い確率で攻撃を成功させることが可能である．計算コストは高いものの，蒸留 (Distillation) などの防御手法を突破する性能を持つ.

### 2.4.6 UAP(Universal Adversarial Perturbation)

Moosavi-Dezfooli ら (2017 年)[11] によって提案された UAP は，特定の画像に依存せず，あらゆる入力画像に対して一律に加えることで誤分類を誘発する汎用的な摂動である．従来の手法が画像ごとに最適な摂動を計算するのに対し，UAP はデータセット全体の分布に対してモデルの決定境界を越えさせる単一のノイズパターンを反復的に構築する．この摂動は異なるネットワーク間でも転移性が高く，リアルタイムの映像ストリームや物理世界での攻撃においても脅威となることが指摘されている.

### 2.4.7 OPA(One-Pixel Attack)

Su ら (2019 年)[12] によって提案された One-Pixel Attack は，入力画像のうちわずか 1 画素の値を変更するだけでモデルを誤分類させる極端な攻撃手法である．本手法は，モデルの勾配情報を使用しない Black-box 攻撃であり，差分進化 (Differential Evolution) アルゴリズムを用いて，誤分類を引き起こす最適な画素の位置と RGB 値を探索する．ニューラルネットワークが局所的な特徴に過度に依存しているという脆弱性を浮き彫りにした手法であり，非常に限定的な情報修正であってもモデルを欺けることを示した.

# 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) や ViT が分類に際し，どのような特徴を学習しているのかを，敵対的サンプルにより明らかにする手法を提案する．敵対的サンプルは，モデルの勾配を計算することで，入力する画像に対し効率的に誤分類させる．そのため，目視では判別が困難な程度の小さい摂動を加えるだけで，モデルの誤分類を誘発することが可能となる．よって，敵対的サンプルにより生成された摂動は，モデルのクラス分類タスクにおける判断根拠として，目に見えない重要な情報を含んでいると仮定し，分析を行う．

本研究において，FGSM，PGD，DeepFool，FAB，C&W，UAP，OPA の 7 つの攻撃手法を採用した．これらを選定した理由は，計算コスト，攻撃の強力さ，摂動の幾何学的性質，および攻撃の制約条件という異なる観点から評価するためである．まず，勾配ベースの標準的な手法として FGSM と PGD を採用した．FGSM は計算効率に優れたベースラインであり，PGD は反復的な更新と射影操作を組み合わせるた強力な攻撃の一つとして広く認められている．次に，決定境界までの最小距離を推定する手法として DeepFool と FAB を選定した．DeepFool は線形近似に基づき最小摂動を効率的に計算し，FAB はハイパーパラメータへの依存を排除しつつ，DeepFool よりも精密に決定境界を探索する．さらに，最適化ベースの極めて強力な攻撃として C&W を採用した．C&W は計算コストが高い反面，蒸留などの防御手法を無効化するほどの攻撃成功率を誇る．最後に，特殊な攻撃条件下での評価をするため，UAP と OPA を導入した．UAP は特定の画像に依存しない汎用的な摂動を生成し，OPA はわずか 1 画素の変更という極端に制限された情報修正によりモデルの誤分類を誘発する．これら 7 つの手法を組み合わせることで，多角的な視点から評価することが

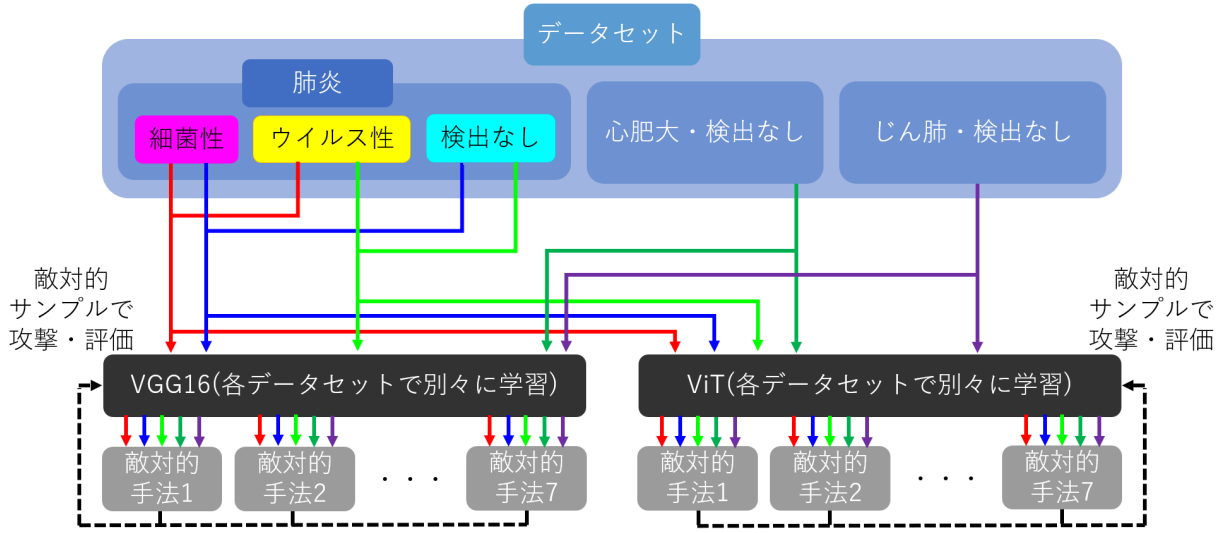



**図 3.1:** 提案手法

できる．そして，各手法で生成した摂動をもとに，モデルがどのような特徴を学習しているのかを明らかにすることで，新しい説明性が得られるかを検証する．また，これら複数種類の敵対的攻撃手法で敵対的サンプルを生成し比較・分析することで，得られた説明性から別の説明性が得られるかを検証する．

本研究の提案手法を図 3.1 に示す．データセットは肺炎，心肥大，じん肺を使用し，さらに肺炎データセットでは，細菌性肺炎，ウイルス性肺炎，検出なしをそれぞれ組み合わせ 3 組のデータセットとした計 5 つのデータセットを使用する．そして，学習モデルである VGG16[17] および ViT で各データセットにおいて学習し，さらに各学習済みモデルに対し，7 つの敵対的手法で攻撃を行い評価をする．

# 第 4 章

# 実験

## 4.1　データセット

### 4.1.1　肺炎

本研究では，Kaggle で公開されている Chest X-Ray Images (Pneumonia) を使用した．本データセットは，肺炎と検出なしの胸部 X 線画像で構成されている．さらに，肺炎には細菌性肺炎とウイルス性肺炎が存在し,「細菌性肺炎（bacteria）」「ウイルス性肺炎（virus）」「検出なし（normal）」の胸部 X 線画像をそれぞれ 1583 枚，1493 枚，1583 枚使用し，合計で 4659 枚を使用した．また，それぞれにおいて訓練データセット (train)，検証データセット (validation)，テストデータセット (test) が 6:2:2 となるようランダムに分割した．ただし，ラベル間でデータ数に差が生じないようにした．具体的には，一方の画像がもう一方の画像よりも多い場合，少ない方の画像と同数になるよう，ランダムに抽出し使用した．表 4.1 に各データセットにおける枚数の詳細を示す．また，図 4.1 に，検出なし，細菌性肺炎，ウイルス性肺炎の胸部 X 線画像の例を示す.

### 4.1.2　心肥大

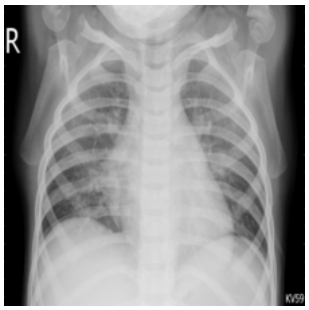
本研究では，Lee らが作成したデータセットを使用した [18]．このデータセットは，Kaggle の National Institutes of Health(NIH) の胸部 X 線データセットから一部の画像を選定している．具体的には，元々のデータセットに含まれていた前後 (AP) ビュー，サポートデバイス，および画像に表示されている人工物 (ペースメーカー，心臓手術の傷跡，脊椎手術の

**表 4.1:** 分割した肺炎データセットの内訳

| データセットの組 | ラベル | 訓練データ | 検証データ | テストデータ | 計 | 合計 |
|---|---|---|---|---|---|---|
| bacteria-normal | bacteria | 949 | 316 | 318 | 1583 | 3166 |
| | normal | 949 | 316 | 318 | 1583 | |
| virus-normal | virus | 895 | 298 | 300 | 1493 | 2986 |
| | normal | 895 | 298 | 300 | 1493 | |
| bacteria-virus | bacteria | 895 | 298 | 300 | 1493 | 2986 |
| | virus | 895 | 298 | 300 | 1493 | |

**(a)** 検出なし

**(b)** 細菌性肺炎



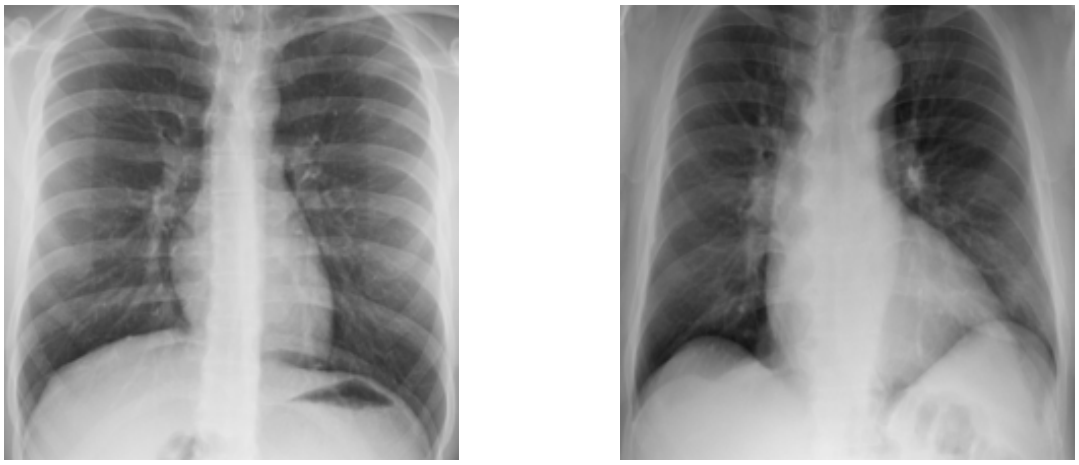
**(c)** ウイルス性肺炎

**図 4.1:** 肺炎データセットの画像の例

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．さらに，選定された胸部 X 線画像に対し，モデルの学習に必要な関連領域のみを含むよう画像を切り抜く処理を施すことで，高品質のデータセットを確保している．その結果，Lee らが作成したデータセットは,「心肥大」368 枚と「検出なし」350 枚の計 718 枚の画像で構成されている．また，このデータセットは，訓練データセットと検証データセットに 7:3 の比率で分割され，検証データセットはさらに検証データセットとテストデータセットに 2:1 の比率で分割されている．表 4.2 に内訳を示す．また，図 4.2 に，検出なし，心肥大の胸部 X 線画像の例を示す.

**表 4.2:** 心肥大データセットの内訳

| ラベル | 訓練データ | 検証データ | テストデータ | 合計 |
|---|---|---|---|---|
| 検出なし | 252 | 63 | 35 | 350 |
| 心肥大 | 264 | 67 | 37 | 368 |
| 合計 | 516 | 130 | 72 | 718 |



**(a)** 検出なし　　**(b)** 心肥大

**図 4.2:** 心肥大データセットの画像の例

### 4.1.3 じん肺

じん肺は，長期間にわたる金属や石炭などの粉じんの吸入により肺組織が線維化し硬くなる疾患である．また，じん肺は胸部 X 線画像の読影により検査可能であり，すりガラス陰影の存在を検出することで診断される．本研究では，National Institute of Occupational Safety and Health(NIOSH)，高知大学医学部 (KM)，National Institutes of Health Clinical Center(NIHCC)[19] で撮影された計 236 枚の胸部 X 線画像を使用した．このデータセットは「じん肺」119 枚と「検出なし」117 枚で構成されている．訓練データセット，検証データセット，テストデータセットに分割し，その内訳を表 4.3 に示す．また，じん肺の胸部 X 線画像における検出において，Zhang らの研究では入力画像の肺野領域を抽出することで分類精度が向上しており, 特に U-Net を使った手法で高い領域抽出精度が報

**表 4.3:** じん肺データセットの内訳

| ラベル | 訓練データ | 検証データ | テストデータ | 合計 |
|---|---|---|---|---|
| 検出なし | 80 | 20 | 17 | 117 |
| じん肺 | 80 | 21 | 18 | 119 |
| 合計 | 160 | 41 | 35 | 236 |

**表 4.4:** U-Net に使用するデータセットの内訳

| 訓練データ | 検証データ | テストデータ | 合計 |
|---|---|---|---|
| 569 | 64 | 71 | 704 |

告されている [20]．そこで，本研究においても，じん肺画像の分類精度向上のため, データセットに対し 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行った.

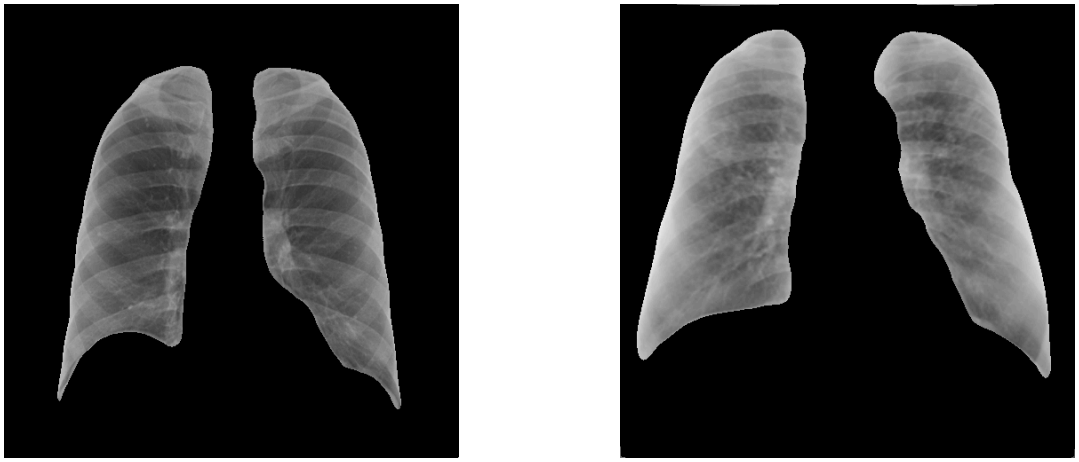
U-Net の学習に使うデータセットとして Montgomery County X-ray Set[21][22] を使用した．このデータセットは，アメリカのメリーランド州モンゴメリー郡の保健福祉省から取得された，800 枚の胸部 X 線画像とそれに対応した 704 枚のマスク画像により構成されている．マスク画像のない胸部 X 線画像は使用せず，704 枚の胸部 X 線画像および対応する 704 枚のマスク画像のみを使用した．データセットを訓練データセット，検証データセット，テストデータセットに分割する．分割した各データセットの画像の枚数を表 4.4 に示す．図 4.3 に元画像，マスク画像，抽出画像の例を示す．また，図 4.4 にじん肺画像と検出なし画像の例を示す.

## 4.2 モデルの学習

すべてのモデルにおいて二値分類を行った.「細菌性肺炎と検出なし」,「ウイルス性肺炎と検出なし」,「細菌性肺炎とウイルス性肺炎」および心肥大，じん肺の 5 つのデータセットでそれぞれ VGG16 と ViT で学習を行った．そして，VGG16 と ViT はどのデータセットで

**(a)** 元画像　　**(b)** マスク画像　　**(c)** 抽出画像

**図 4.3:** U-Net による肺野領域抽出



**(a)** 検出なし　　**(b)** じん肺

**図 4.4:** じん肺データセットの画像の例

あっても訓練データセット，検証データセット，テストデータセットにおいて，全く同じ胸部 X 線画像で構成されているデータセットを使用した．

### 4.2.1 前処理と実験設定

各データセットの画像は，グレースケール画像として読み込み，3 チャンネルに複製しモデルに入力する．また，前処理として 224x224 にリサイズし正規化する．ただし，訓練データセットでは位置依存による過学習を抑制するため，PyTorch の RandomHorizontalFlip により確率 0.5 で水平方向に反転させる．さらに，学習の安定化を図るため，訓練データセットではデータの順番をランダムに入れ替えておく．

VGG16 では，バッチサイズは 32 とし，最適化アルゴリズムは Adam[23] を使用する．学習率は基本 $10^{-4}$ とするが，VGG16 における「細菌性肺炎とウイルス性肺炎」では，精度が伸び悩んだため $10^{-3}$ とした．同様に，エポック数についても基本 30 とするが,「細菌性肺炎とウイルス性肺炎」では 100 とした．また，学習時に学習率減衰を行う．初期の学習率を $\lambda_0$，$t$ 番目の epoch での学習率を $\lambda_t$ として式 (4.1) に示す.

$$\lambda_t = \lambda_0(0.95)^t \tag{4.1}$$

ViT では，バッチサイズは 16 とし，最適化アルゴリズムは AdamW[24] を使用する．学習率は $3 \times 10^{-4}$ とし，重み減衰 (Weight Decay) は $10^{-4}$ とする．また，学習時には PyTorch の CosineAnnealingLR で学習率減衰を行う．学習初期には比較的大きな更新を行い，学習が進むにつれて徐々に微調整へ移行することで，安定した収束を促す.

### 4.2.2 VGG16

本研究で用いた VGG16 は，医用画像分類および説明性解析を目的として複数の点で構造を変更している．本来の VGG は，畳み込み層と ReLU からなるブロックを段階的に重ねた後，Flatten を経て 3 層の大規模な全結合層で分類を行う構造であり，多数のパラメータを有する．一方，本研究で用いた VGG では，畳み込みブロックの構成自体は本来の VGG16 と同様，$3 \times 3$ の畳み込み層と Max Pooling であるが，その内部構造と分類部に変更を加えている.

まず，各畳み込み層の直後に Batch Normalization を導入している．VGG では，本来 Batch Normalization は用いられていないが，学習の安定化および収束の高速化を目的としてこれを追加している．次に分類部において，全結合層ではなく，Global Average Pooling(GAP) を用いて特徴マップを空間的に集約した後，単一の全結合層によって分類を行う構成としている．この変更により，モデルのパラメータ数やモデル自身のサイズを大幅に削減するとともに，過学習の抑制と入力サイズに対する柔軟性を確保している．さらに，本来の VGG で用いられている Dropout は使用せず，Batch Normalization が正則化を行

う．出力層についても，本来の VGG が多クラス分類を前提とした Softmax 出力を用いるのに対し，本研究では 2 クラス分類を想定し，出力次元を 1 とした上で BCEWithLogitsLoss を用いる設計としている．これにより，医用画像における二値判別タスクに適した構成となっている．

### 4.2.3 ViT (Vision Transformer)

モデルとして，DeiT-base patch16 224 を使用する．本モデルは，DeiT-base を基盤とした事前学習モデルを用い，医用画像分類および説明性解析を目的として Fine-tuning を行ったものである．本研究では，医用画像における 2 クラス分類タスクに適用するため，分類ヘッドの出力次元を 2 に変更している．また，ImageNet で事前学習された重みを初期値として用いることで，限られた医用画像データに対しても安定した学習が可能となるよう配慮している．

Fine-tuning では，Transformer 本体のパラメータを固定し，分類ヘッドのみを学習対象とした．この設定により，事前学習によって獲得された特徴表現を保持しつつ，医用画像分類に必要な判別境界のみを調整することが可能となる．加えて，学習パラメータ数や過学習の抑制，学習の安定化に貢献することができる．

## 4.3 敵対的サンプルの生成

本実験では，4 種類の方法で敵対的サンプルを生成した．各条件で学習したモデルそれぞれに FGSM，PGD，DeepFool，FAB，C&W，UAP，OPA を適用した．

FGSM では，最大摂動量 $\epsilon$ を VGG16 では 0.01，ViT では 8/255 とした．PGD では，最大摂動量 $\epsilon$ を 0.03，5 ステップ行い，1 ステップあたりの更新量 $\alpha$ を 0.01 とした．DeepFool では，最大反復回数を 50，決定境界を少しだけ超えるための overshoot を 0.02 とした．FAB では，VGG16 において，最大摂動量 $\epsilon$ を 1.0，100 ステップ行い，1 ステップあたりの更新量 $\alpha$ を 0.1 とし，ViT において，最大摂動量 $\epsilon$ を 8/255，最大 50 ステップとした．

C&W では，摂動の小ささと攻撃成功のトレードオフ係数 c を VGG16 では $10^{-2}$，ViT では $10^{-3}$ とし，ステップ数を VGG16 では 1000，ViT では 500 とし，最適化における Adam の学習率を $10^{-2}$ とした．UAP では，最大摂動量 $\epsilon$ を VGG16 では 0.01，ViT では 0.02 とし，摂動の生成は VGG16 では PGD，ViT では DeepFool を使用した．OPA では，個体数を 30，進化の最大世代数を VGG16 では 100，ViT では 60 とした．

# 第 5 章

# 結果

## 5.1 モデルの学習

表 5.1 に各条件下での学習結果を示す．検出なしと肺炎の胸部 X 線画像を使用したモデルでは，すべてのモデルで検証データとテストデータ共に 90%以上の精度を示し，ViT よりも VGG16 の方がやや精度が高い．また，細菌性肺炎とウイルス性肺炎の胸部 X 線画像を使用したモデルでは，どちらのモデルにおいても検証データとテストデータで 70%から 75%の精度を示し，VGG16 よりも ViT の方がやや精度が高い．心肥大およびじん肺の胸部 X 線画像を使用したモデルでは，VGG16 においては検証データとテストデータ共に 90%以上の精度を示したが，ViT においては検証データでは 90%以上の精度を示したもののテストデータでは 90%を少し下回った．また，VGG16 および ViT のどちらにおいても，最も精度が高いのは「細菌性肺炎と検出なし」のデータセットであった．

## 5.2 敵対的サンプルの生成

生成した敵対的サンプルに対するモデルの予測正解率を，データセットごとに図 5.1，図 5.2，図 5.3，図 5.4，図 5.5 に示す．各図は，元画像つまり攻撃前のテストデータでのモデルの予測正解率と 7 つの敵対的手法による敵対的サンプルでの予測正解率の計 8 つを示している．さらに，モデルアーキテクチャと正解ラベルによる 4 パターンにおける予測正解率を棒グラフで示している．生成した敵対的サンプルに対するモデルの予測正解率であるため，攻撃前の正解率が高い場合は，敵対的サンプルでの正解率が低いと攻撃自体の成功率が高い

**表 5.1:** モデルの学習結果

| モデル | データセット | 検証データ | テストデータ |
|---|---|---|---|
| VGG16 | bacteria-normal | 98.42 | 96.38 |
| VGG16 | virus-normal | 95.47 | 94.50 |
| VGG16 | bacteria-virus | 72.32 | 72.67 |
| VGG16 | cardiomegaly | 93.08 | 95.83 |
| VGG16 | pneumoconiosis | 92.68 | 91.43 |
| ViT | bacteria-normal | 95.89 | 94.18 |
| ViT | virus-normal | 94.13 | 93.00 |
| ViT | bacteria-virus | 75.00 | 74.00 |
| ViT | cardiomegaly | 90.00 | 88.89 |
| ViT | pneumoconiosis | 92.68 | 85.71 |

ことを意味する．

FGSM では，肺炎に関するデータセットで,「細菌性肺炎とウイルス性肺炎の ViT」以外では，ほとんどの画像で攻撃前後で予測ラベルが反転した．また，心肥大データセットやじん肺データセットでは，VGG16 の検出なしで予測正解率が下がりきらなかったものの，それ以外では攻撃後の予測正解率は大きく下がった．PGD では，VGG16 においてすべての画像で攻撃前後で予測ラベルが反転し，ViT においてすべての画像が誤分類された．DeepFool ではどのパターンにおいても，すべての画像で攻撃前後で予測ラベルが反転した．FAB では，VGG16 においてすべての画像で攻撃前後で予測ラベルが反転せず，ViT においては 50%ほどとなった．C&W では，VGG16 ではほとんどの画像で攻撃前後で予測ラベルが反転したが，ViT ではじん肺以外で一方のラベルでの予測正解率が高くもう一方のラベルでの予測正解率は低くなった．UAP での VGG16 においてはどちらか一方のラベルでの予測正解率が 0 になるが，もう一方のラベルでの予測正解率は 100 となった．OPA では，

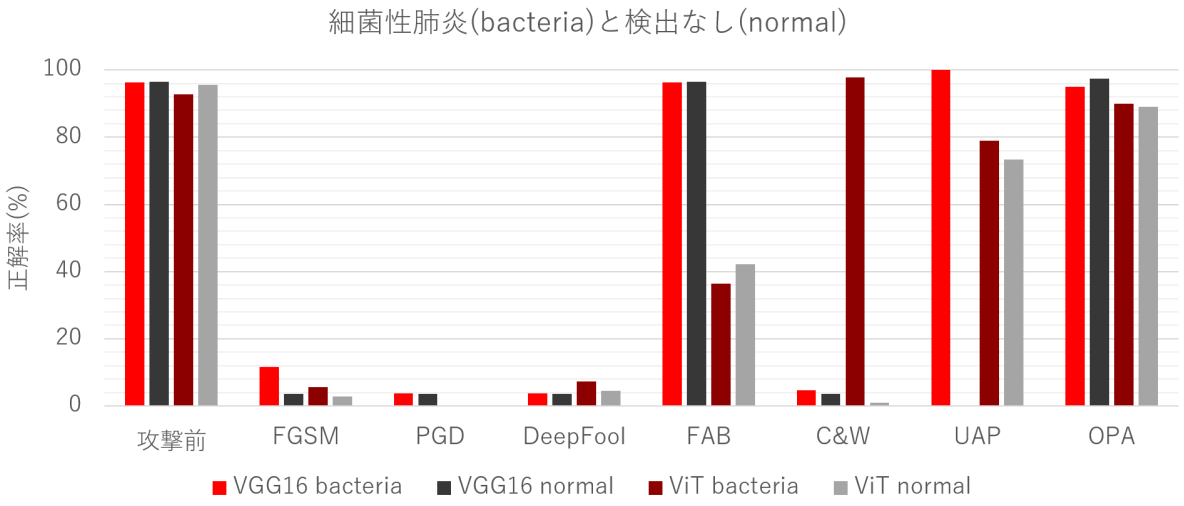



**図 5.1:** 細菌性肺炎と検出なしにおける各敵対的サンプルでの予測正解率

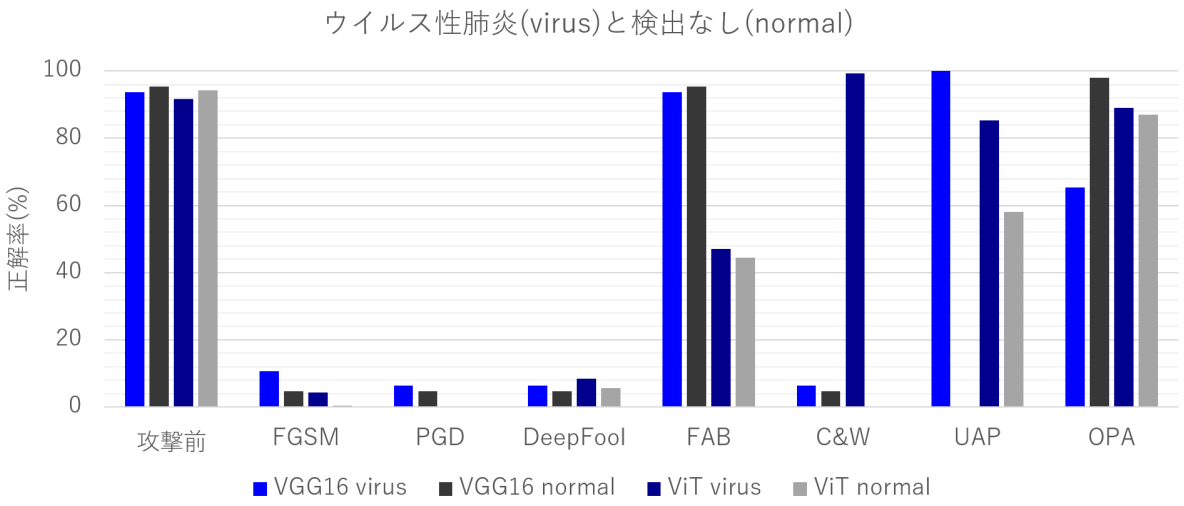



**図 5.2:** ウイルス性肺炎と検出なしにおける各敵対的サンプルでの予測正解率

予測正解率の変化は小さく，VGG16 では一部で上がることもあったが，ViT では攻撃前以下の正解率となった．

## 5.3 摂動を加えた領域の可視化

生成した敵対的サンプルを確認すると，攻撃後の予測正解率は低いものの，加えた摂動が明らかに大きく，元画像から逸脱している場合があった．その例として，PGD における

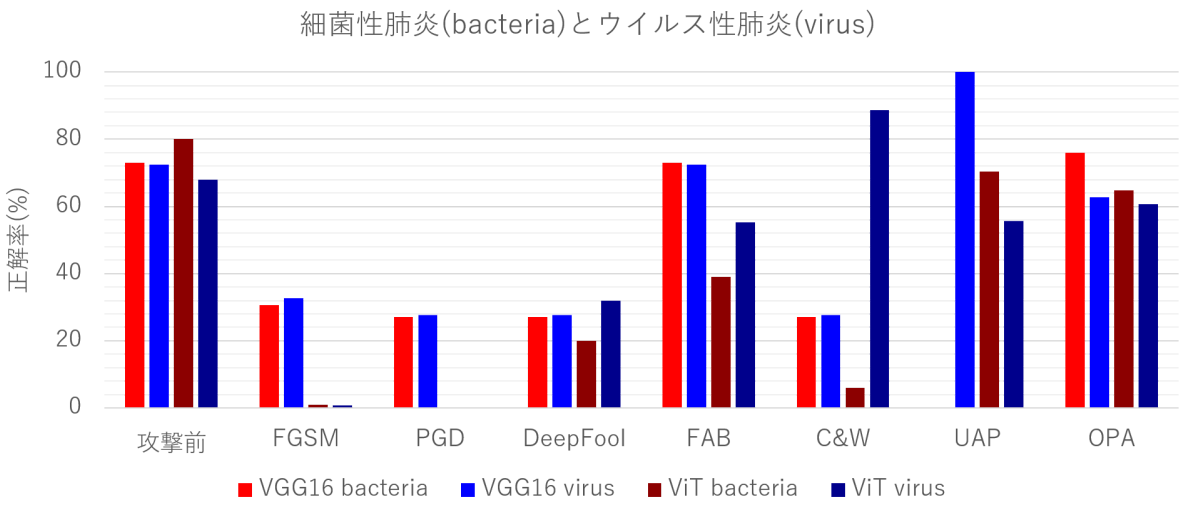



**図 5.3:** 細菌性肺炎とウイルス性肺炎における各敵対的サンプルでの予測正解率

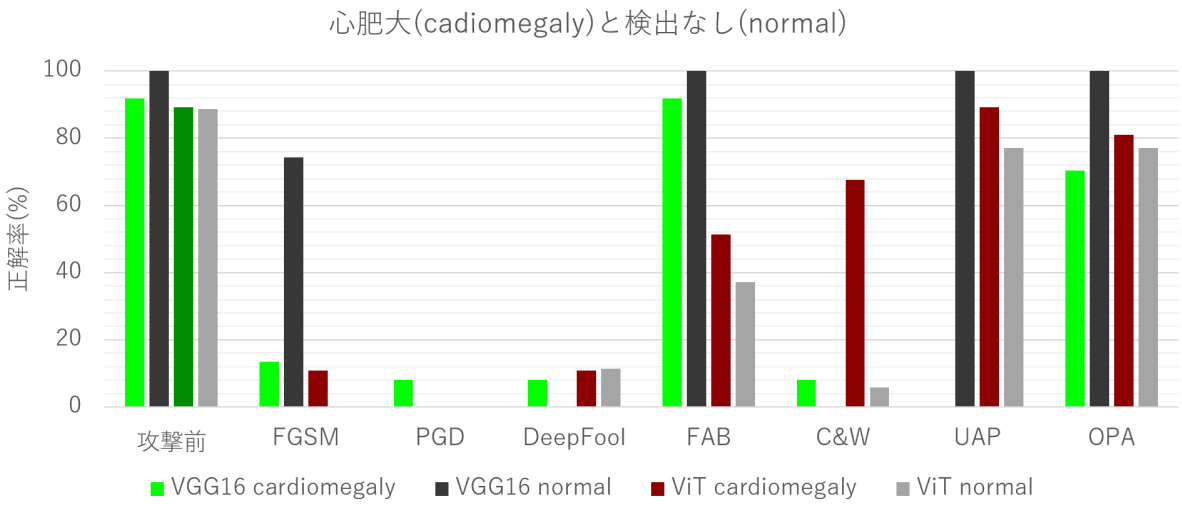



**図 5.4:** 心肥大と検出なしにおける各敵対的サンプルでの予測正解率

ViT での，元画像であるじん肺画像と摂動を加えた画像を図 5.6 に示す．本研究では明らかに摂動が大きすぎる場合を除いて，攻撃後の予測正解率が低いものに対し，摂動を加えた領域の可視化を行う．具体的には，FGSM，VGG16 での PGD，DeepFool において可視化を行う．

図 5.7 に摂動を可視化した例を示す．生成した摂動は，モデルアーキテクチャや敵対的手法による差が大きく，そのままでは定量的な評価を行うのは困難である．そこで，特に強い

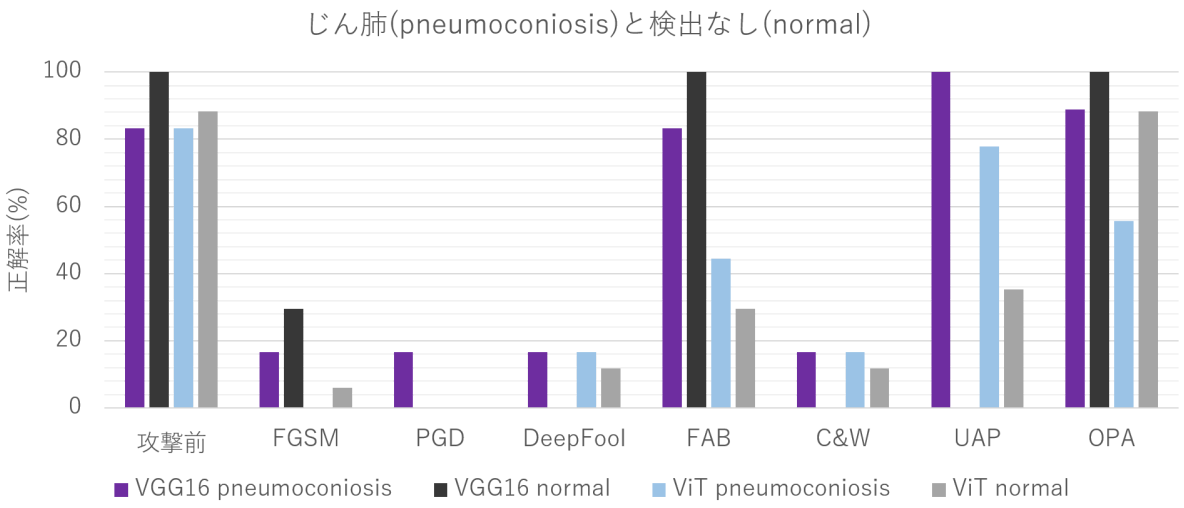



**図 5.5:** じん肺と検出なしにおける各敵対的サンプルでの予測正解率

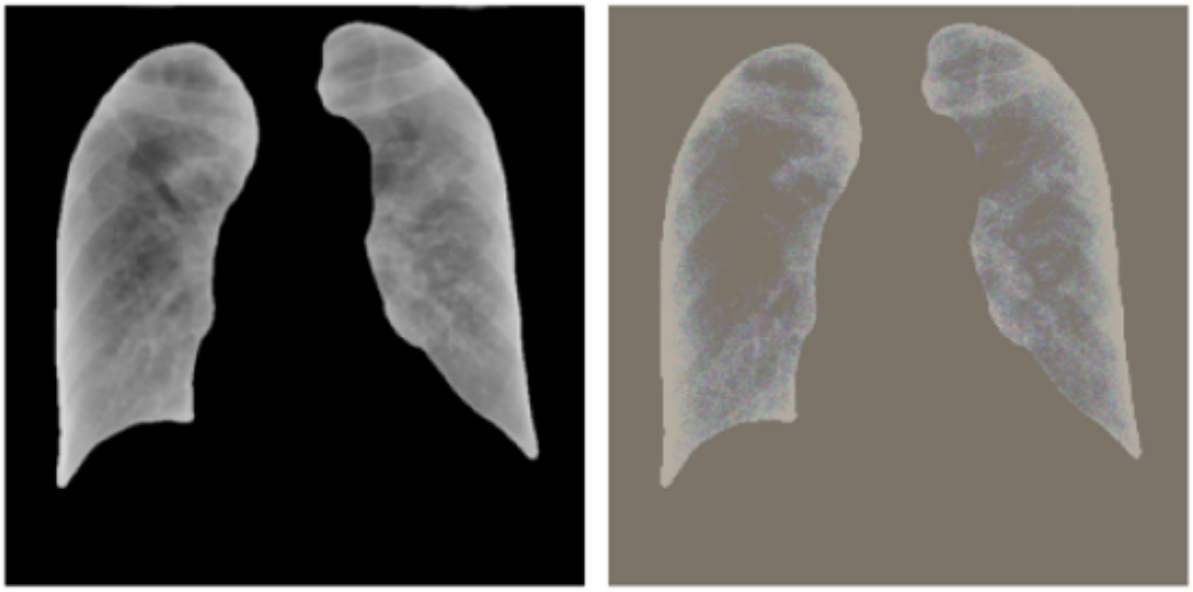


**図 5.6:** PGD における ViT でのじん肺の元画像 (左) と摂動を加えた画像 (右)

摂動を与えた領域に注目する．具体的な流れを，図 5.8 に示す．まず，摂動の絶対値を画素単位で求め，5x5 のガウシアンフィルタで平滑化する．そして，特に強い摂動として上位 15%マスクを生成し，元画像にマスクした摂動を重ねて表示する．

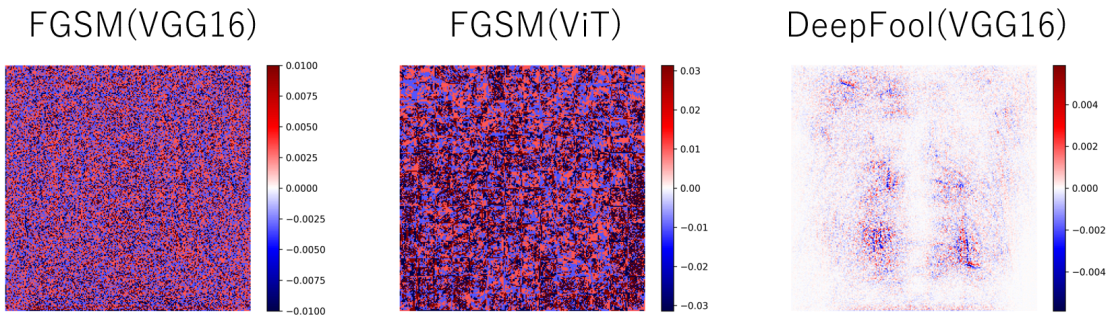



**図 5.7:** 可視化した摂動例

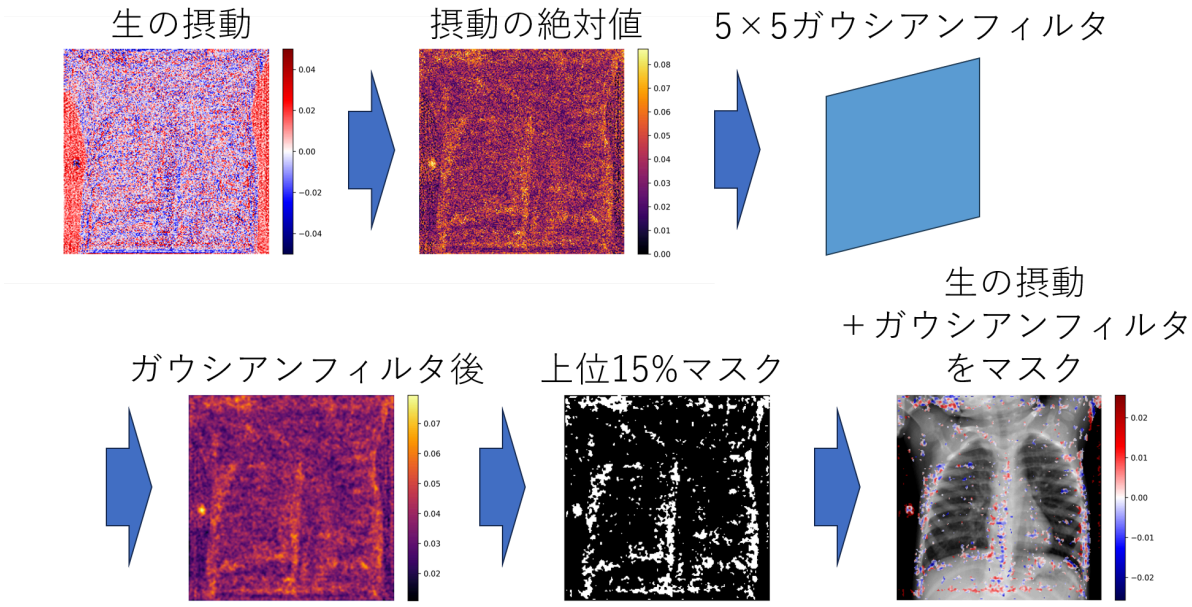



**図 5.8:** 強い摂動を与えた領域の可視化の流れ

### 5.3.1 VGG16 における FGSM

VGG16 での FGSM における可視化の結果を図 5.9 に示す．各画像の左には変換前のラベルである正解ラベルを示している．肺炎に関するデータセットでは，人体と背景の境界，肋骨の周辺，肺野領域の境界といった部分に対し主に強い摂動が加えられていた．心肥大では，肺野領域の中心に集まっており，肺野領域の境界にも強い摂動が加えられていた．じん肺では，肺全体に強い摂動が加えられていた．

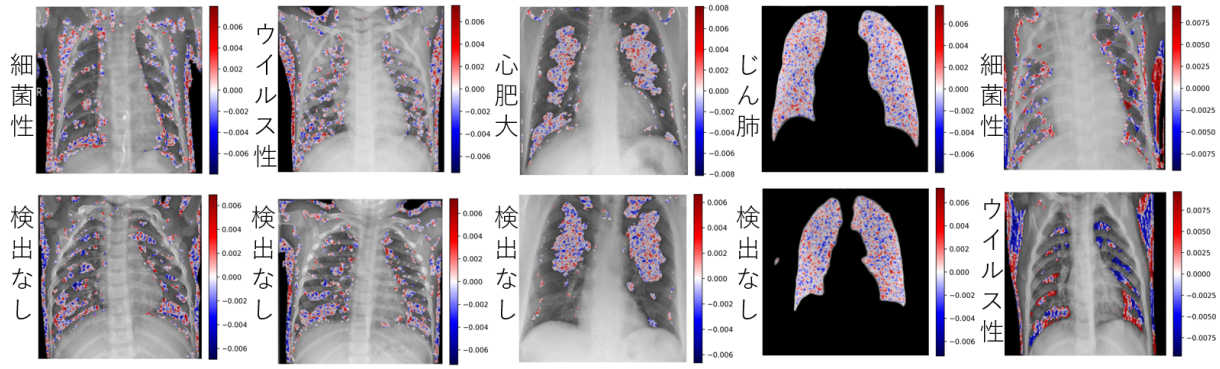



**図 5.9:** VGG16 における FGSM での可視化

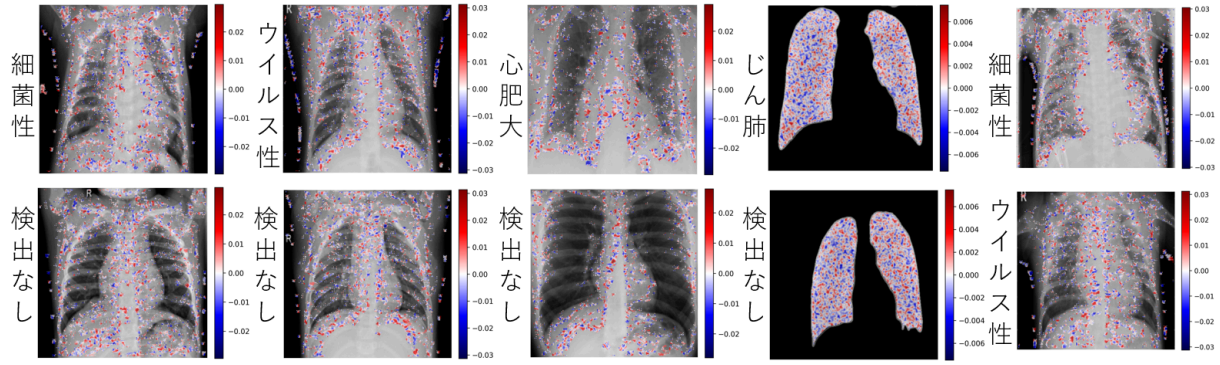



**図 5.10:** ViT における FGSM での可視化

### 5.3.2 ViT における FGSM

ViT での FGSM における可視化の結果を図 5.10 に示す．じん肺以外では，肺野領域内より肺野領域外における境界部分に強い摂動が加えられる傾向があった．じん肺では，肺全体に強い摂動が加えられていた．

### 5.3.3 VGG16 における PGD

VGG16 での PGD における可視化の結果を図 5.11 に示す．肺炎と検出なしでは，人体と背景の境界，骨や内臓の形にそうように強い摂動が加えられていた．心肥大では，心肥大から検出なしへの変換で左右に線を引くように強い摂動が加えられており，検出なしから心肥大への変換で肺野領域内に強い摂動が集中していた．肺炎同士では，肺野領域の境界に強い摂動が加えられていた．じん肺では，肺野領域の境界もしくは中心に強い摂動が集中して

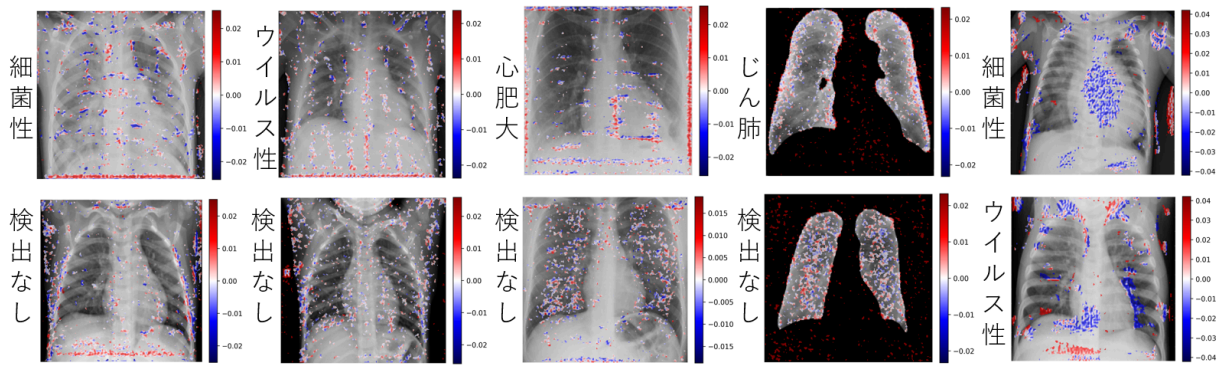



**図 5.11:** 強い摂動を与えた領域の可視化の流れ

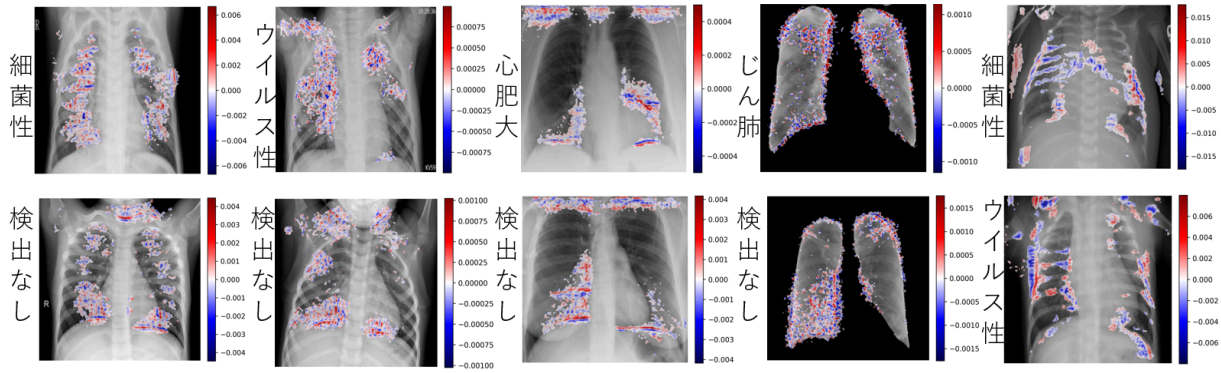



**図 5.12:** 強い摂動を与えた領域の可視化の流れ

いた.

### 5.3.4 VGG における DeepFool

VGG での DeepFool における可視化の結果を図 5.12 に示す．肺炎と検出なしでは，肺野領域内と首やあごに強い摂動が加えられていた．肺炎同士では，肋骨や背骨に強い摂動が加えられていた．心肥大では，肺野領域の上下に分かれて強い摂動が加えられていた．じん肺では，肺野領域の中心と境界どちらにも強い摂動が加えられていた．

### 5.3.5 ViT における DeepFool

ViT での DeepFool では，どのデータセットでも 14x14 のパッチの中心に強い摂動が集まっていた．

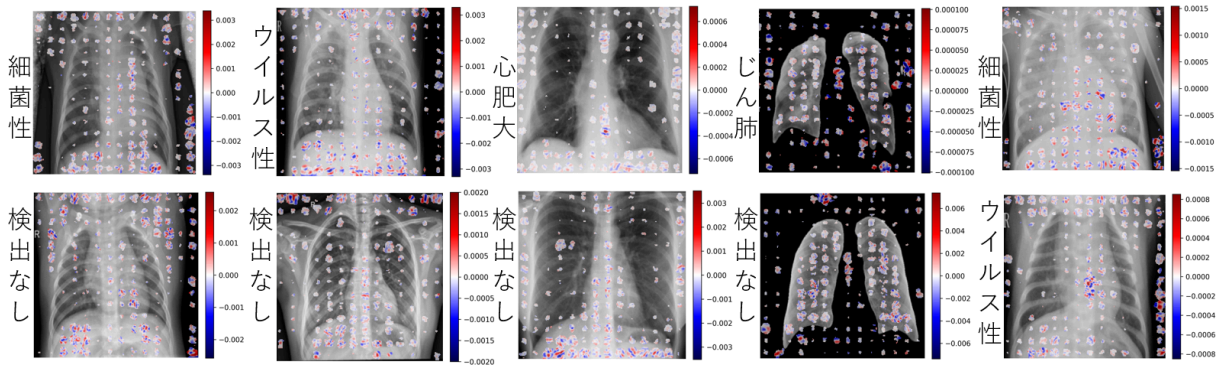



**図 5.13:** ViT における DeepFool での可視化

# 第 6 章

# 考察

## 6.1 モデルの学習

「細菌性肺炎とウイルス性肺炎」の分類タスクでは，他タスクと比較して精度が低い．そのため，細菌性とウイルス性の視覚的な差が非常に小さく，モデルにとって識別が困難なタスクであると考える．VGG16 および ViT において，最も精度が高いデータセットは「細菌性肺炎と検出なし」であったが，このデータセットはすべてのデータセットの中で最も画像枚数が多い．さらに，画像枚数が比較的少ない心肥大やじん肺のデータセットでは，肺炎と検出なしのデータセットより精度が低い傾向にある．そのため，検出なしを含むデータセットでは，構成される画像枚数で精度に差が出た可能性がある．

## 6.2 敵対的サンプルの生成

FGSM において，特に ViT では正解率が 10%以下へと変化した．FGSM では，勾配情報を利用し効率的に決定境界の外側へ誘導するため，モデルが勾配方向に強く依存した特徴を使っていると考える．

PGD では，VGG16 において元画像の予測ラベルと敵対的サンプルの予測ラベルが反転し，ViT において正解率が 0%となった．攻撃対象のクラスを確認すると，VGG16 では攻撃対象のクラスを元画像の予測ラベルとしていたが，ViT では攻撃対象のクラスを正解ラベルにしてしまっていた．そのため，攻撃対象のクラスの違いが原因であると考える．また，このことから結果的に PGD ではすべてのラベルにおいて反転していると解釈可能である．

DeepFool では，どのモデルや画像においても，元画像の予測ラベルと敵対的サンプルの予測ラベルが反転した．VGG16 は，浅い層において一般的なエッジやテクスチャを学習する一方で，深い層においては肺炎に関連する意味的特徴表現に強く依存する．敵対的摂動はこれら深層表現を集中的に変化させており，モデルの判別が高次概念に基づいていると考える．

FAB は多クラス分類において，どのクラスの境界が最も近いかを効率的に見つけ出すアルゴリズムが強みである．そのため，本研究の様な二値分類では DeepFool と似た挙動をとると考えられる．しかし，本実験の VGG16 に対してはどの画像も分類を変えることができなかった．これは，FAB では境界を超えるだけでなく，さらに元画像から離れすぎないよう摂動の調整を行うため，本実験ではこの調整により結果的に境界を越えられなかったのではないかと考える．また FAB において，VGG16 では精度の変化がなかったが，ViT では精度が 50%以下になった．そのため，VGG の勾配が FAB の探索アルゴリズムと相性が悪い可能性がある．

C&W では，VGG16 においてはラベルが反転しているが，ViT ではじん肺以外において片方のラベルの精度が下がりにくかった．そのため，肺野領域をマスクした画像を使うことで改善される可能性がある．

UAP では，VGG16 において 100%か 0%といった極端な結果となった．敵対的サンプルの生成時に使用した画像の順番とラベルを確認すると，一方のラベルが前半に集中し，もう一方のラベルが後半に集中していた．また，正解率が 0%となったラベルが前半に集中していた．そのため，本実験では片方のラベルに対する共通の摂動を生成する結果となった可能性がある．しかし，ラベルごとに UAP で摂動を生成することで，両方のラベルで正解率を大きく下げることができる可能性があり，ラベルごとの共通の特徴を攻撃した摂動を生成できる可能性がある．

OPA では，大きく正解率が下がらなかったものの，正解率に変化があり，1 ピクセルの変更はモデルの予測に多少の影響があると考える．本実験では，画像サイズを 224x224 としていたため，画像サイズや攻撃のピクセル数，ピクセルサイズを変更することで正解率が大

きく変化する可能性がある．しかし，ごく一部のピクセルを変化させて予測ラベルが反転しても，モデルがどんな特徴を学習したのかを客観的に評価するのは困難である．そのため，変更したピクセルが病変部位であるか，周辺であるか，あるいは関係ない場所であるかなどを記録し，ラベルが反転したかどうかの結果と合わせて評価するなど工夫が必要であると考える．

モデルの構造に着目すると，敵対的サンプルでの精度は，C&W 攻撃では VGG16 の方が ViT よりも正解率が低く，逆に FGSM では ViT の方が VGG16 よりも正解率が低かった．ViT は広域的な特徴を捉えるのが得意であるので，入力画像全体に散布される微細なノイズに対し，予測の根拠となる「Attention」機構が乱されやすい性質を持つ可能性があると考える．

## 6.3 摂動を加えた領域の可視化

強い摂動を加えた領域を確認すると，病変部位全体ではなく，人体と背景の境界，骨や内臓の境界である傾向があった．これは，モデルが病変部位における画像の特徴を学習しているのではなく，画像全体におけるコントラストを学習していることを示していると考える．また，ViT における DeepFool では，パッチの中心に強い摂動が集まっており，ViT がパッチごとに処理を行っている部分に攻撃していると考える．強い摂動が多いパッチを確認すると，じん肺では肺野領域に，それ以外では肺野領域外にある傾向があった．そのため，病変部位そのものよりも病変部位と周囲との差を学習している可能性がある．

本実験では，攻撃後の予測正解率が低く，摂動が大きすぎない場合において可視化を行った．しかし，敵対的手法ごとに生成のアルゴリズムが異なることから，各手法により適した可視化手法や評価方法が存在する可能性がある．さらに，今回可視化していないものでもラベルの特徴を攻撃できている可能性がある敵対的サンプルも存在しており，それらを含めて評価をする方法や指標を検討する必要がある．

# 第 7 章

# 結論

本研究では，複数の敵対的サンプルを生成し，生成した摂動をもとにモデルがどのような特徴を学習しているのかを明らかにする手法を提案した．各データセットを学習したモデルに攻撃を行うと，PGD と DeepFool においてすべてのラベルが反転し，FGSM や C&W では一部で攻撃後の正解率が大きく下がった．FAB では ViT でのみ正解率がおよそ半分ほどとなり，OPA では正解率の変化はあったものの大きく下がらなかった．また UAP では，VGG16 で正解率が 100%か 0%となった．そのため，本実験の敵対的サンプルの生成においては，FAB，UAP，OPA ではモデルの説明性に用いるには不十分であった．続いて，正解率が大きく下がった場合の摂動を確認すると，C&W と ViT での PGD では明らかに大きな摂動が加えられており，本来の敵対的サンプルの特徴である目視で判別が困難な摂動から大きく逸脱していた．そのため，それらを除き，特に強い摂動を加えられた領域の可視化を行った．すると，病変部位全体ではなく，人体と背景の境界，骨や内臓の境界である傾向があり，モデルが病変部位における画像の特徴を学習しているのではなく，画像全体におけるコントラストを学習している可能性が示された．しかし，各敵対的手法の特徴を強調するような評価方法を検討し評価をすることで，さらなる説明性を獲得できる可能性がある．

# 謝辞

本研究を進めるにあたり，多大なるご指導とご助言を賜りました吉田真一教授に，心より感謝申し上げます．研究の方向性から細部に至るまで丁寧にご指導いただき，本研究を完成させることができました．

また，副査を引き受けていただいた岩田教授と敷田教授に感謝致します．本研究に対し，貴重なご指摘と有益なご助言を賜りましたこと，深く感謝申し上げます．副査としてのご指導は，本研究を多角的に見直す上で大変有意義なものでした．

また，研究活動を通して貴重なご意見や議論をしてくださった研究室の皆様に深く感謝いたします．日々の議論や助言は，本研究を進める上で大きな支えとなりました．

さらに，研究生活を支えてくれた家族や友人に感謝の意を表します．皆様の支えがなければ，本研究を成し遂げることはできませんでした．

# 参考文献


[1] DosoViTskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[2] 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,* pages 618–626, 2017.

[3] T. Nakajima, et al., "Explainability of CNN Classification Models Using CycleGAN and Their Application to Medical Imaging," in *Computational Intelligence and Industrial Applications,* 2025.

[4] 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.

[5] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

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

[7] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

[8] Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016.

[9] Croce, Francesco, and Matthias Hein. "Minimally distorted adversarial examples with a fast adaptive boundary attack." *International conference on machine learn-*

*ing.* PMLR, 2020.

[10] Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." *2017 ieee symposium on security and privacy (sp).* Ieee, 2017.

[11] Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2017.

[12] Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks." *IEEE Transactions on Evolutionary Computation* 23.5 (2019): 828-841.

[13] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (2002): 2278-2324.

[14] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).

[15] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016.

[16] Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention." *International conference on machine learning.* PMLR, 2021.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556,* 2014.

[18] 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.

[19] 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,* pages 2097–2106, 2017.

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

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

[22] 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.

[23] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations, arXiv:1412.6980*, 2015.

[24] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." *arXiv preprint arXiv:1711.05101* (2017).