博士論文

ディープラーニングを用いた放射線皮膚炎グレード判定システムに関する研究

2022 年 3 月

和田 清隆

目次

第1章　序論

1.1　背景

1.1.1　がん治療における放射線治療の現状

放射線治療は，手術，化学療法（抗がん剤治療）とともに，がんの 3 大治療として確立されている．これら 3 つの治療法は，表 1-1 に示すようにそれぞれ特徴がある．放射線治療の特徴は，手術と異なり患者の負担が小さいことや，治療する臓器の形態や機能が温存されることである．化学療法と異なる点は，病巣に集中した治療をできることである．このようなことから，放射線治療の場合，原則として入院の必要はなく，日常生活をおくりながら仕事や学校を休まなくても，通院で治療を行うことができる．がん治療における患者に負担が少ない治療である．図 1-1 に示すように放射線治療を受けた患者数は，増加傾向にあり 2015 年には，患者数が 271,000 人，そのうち 225,000 人が新規に放射線治療を受けたと推定されている [1)]．

表 1-1 がん治療の特徴

	手術	化学療法	放射線治療
対象	局所	全身	局所
機能･臓器	摘出	温存	温存

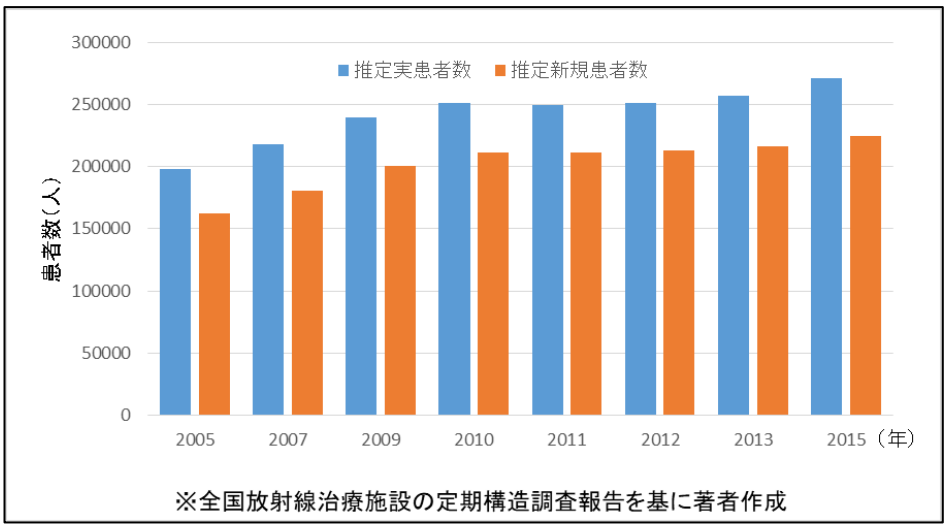


図 1-1 放射線治療を受けた患者数（推定）

1.1.2　放射線治療における有害事象

がん治療における患者に負担が少ない治療である一方で，放射線治療の過程には，放射線性皮膚炎や粘膜炎など有害事象の局所反応がみられ，この局所反応から生じる苦痛は，治療を完遂する患者にとって最も大きな悩みとなる[2)]．表 1-1 に示したように放射線治療は局所治療のため，有害事象は照射された部位に限定される．照射される部位により，口腔粘膜炎，放射線皮膚炎，放射線肺炎などがある．放射線治療を施行したときには，腫瘍周辺の正常組織の障害も何らかの形で生じると考えなければならない．正常組織に対する有害事象の発生については，さまざまな臓器に対して多彩な病態をとる．

図 1-2 に示すように一般に放射線治療は分割照射法により週 3〜5 回，3〜7 週間にわたって施行されることがほとんどであり，各人体組織に生じる有害事象は，この期間中に積み重なるものと考えられる．放射線の生体に対する反応は，物理学的，化学的反応を通じ，やがて生物学的反応として発現される．その発生機序は，経時的に捉えられ有害事象の発生時期により，治療中または 3 カ月以内に出現する急性期有害事象と 3 カ月以降に出現する晩期有害事象がある[2,3)]．

急性期有害事象は，細胞分裂の盛んな細胞の放射線による分裂死が原因である．例えば，皮膚，粘膜，骨髄など細胞の生成・破壊が早い組織では，細胞の補充が間に合わないため，組織全体に機能低下により炎症が起こる．これに対して晩期有害事象は，臓器の耐用線量や実際の照射線量や体積に影響され，いったん発症すると，回復しにくい特徴がある[4)]．

紺屋ら[5)]は，図 1-3 に示すように皮膚線量と放射線皮膚炎との関係について調査し，皮膚線量が 45Gy 以下では皮膚炎はグレード 1〜2 に留まり，治療終了後 2.5 週間で回復に向かったが，45Gy 以上では，全症例においてグレード 3 が出現し，2〜4 週間でピークを迎え，5 週で回復に転じた，さらに，皮膚炎に対する適切な評価と処置が行えなかった症例では，皮膚炎の改善までの期間が延長したことを報告している．

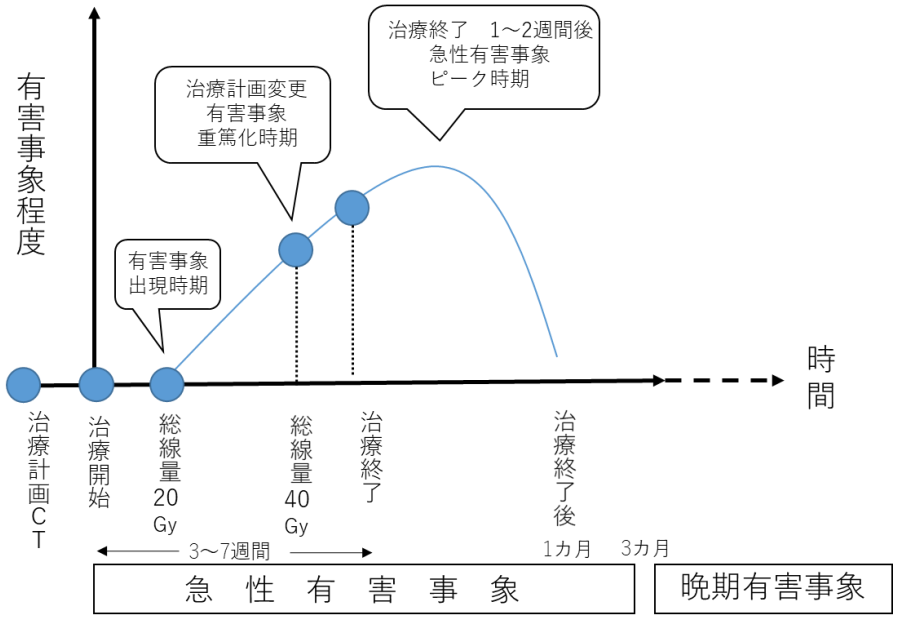


図 1-2 放射線治療の有害事象の出現時期

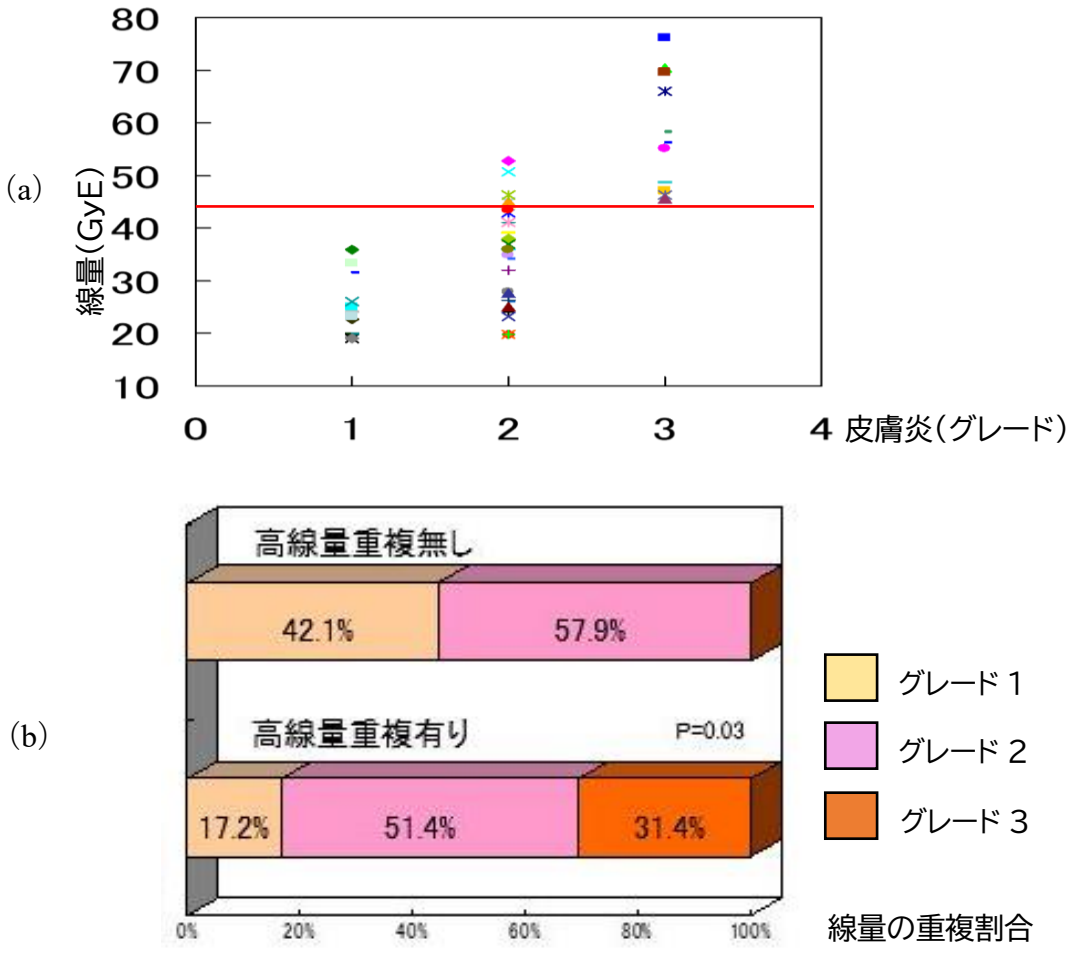


図 1-3　皮膚線量と放射線皮膚炎の関係

(a) 皮膚線量と皮膚炎（グレード）の関係

(b) 高線量領域の重複の有無の比較

1.1.3　放射線皮膚炎の評価とケア

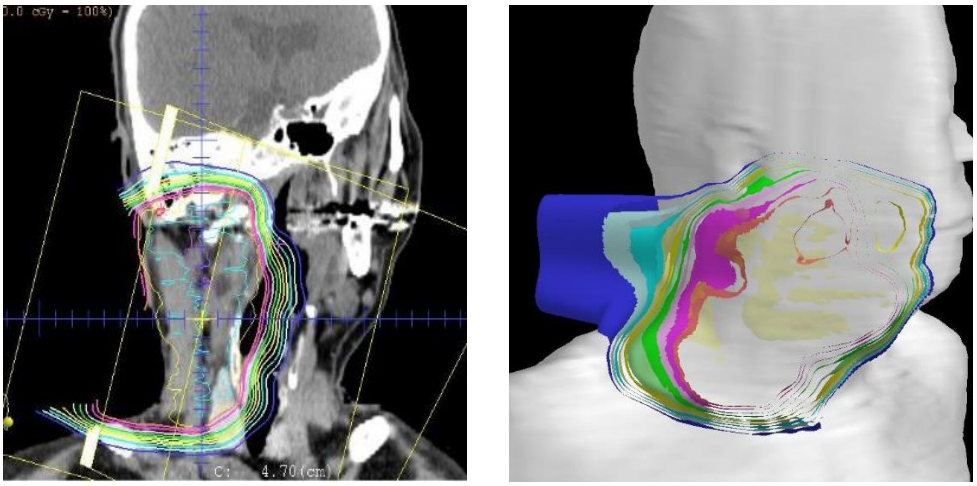
放射線皮膚炎は放射線治療における一般的な有害事象の一つであり，放射線治療を受ける患者の 90%以上の患者がそれを発症すると報告されている [5,7]．その程度は軽度の紅斑から湿性落屑，そして時には潰瘍にまでに及ぶことがある．多くの施設では，放射線皮膚炎の評価を表 1-2 に示す有害事象共通用語規準（Common Terminology Criteria for Adverse Events: CTCAE）[8] を用いてグレードの判定を行っている．CTCAE は放射線皮膚炎を軽度のグレード 1 から有害事象による死亡の 5 まで分類している．放射線皮膚炎は，図 1-4（a）に示すように外部から放射線を腫瘍に対して照射する．図 1-4（b）に示すように皮膚への放射線通過により，基底細胞層が損傷し，基底細胞層における細胞の正常な産生と皮膚表面の細胞の破壊との間に不均衡が生じることにより引き起こされる [9]．放射線治療開始から約 2～3 週間で起こり始め，照射が進むに従って段階的に増悪し，治療終了時から 2～3 週間でピークを迎える．その後，徐々に改善していくことが多い．放射線皮膚炎による症状の程度には個人差があり，放射線総線量，皮膚の強さなどが影響するが，そのグレードに応じた適切な対応が不十分であれば，放射線皮膚炎の増悪や遷延を起こす可能性がある．通常分割の放射線治療でグレード 4 以上の放射線皮膚炎を生じることは極めてまれであるが，化学療法の併用や 1 回線量の高い治療では，重度の放射線皮膚炎を発症することがあるので評価に応じたケアが必要である [9]．図 1-5 に放射線皮膚炎の症例を示す．

放射線性皮膚炎に対するケアとして，グレード 1 は経過観察になることがほとんどであり，グレード 2 は保清・保湿・保護，治療としてステロイド軟膏の使用が広く知られている．通常，グレード 3 やグレード 4 の潰瘍に対する皮膚科に診療を依頼するなど外科的処置を含めた専門の処置となる [10]．放射線皮膚炎を正しく評価し，症状に応じて適切に管理することができれば，一貫性のある質の高いケアが可能となり，放射線皮膚炎の低減と患者の Quality of life（QOL）改善に繋がる．

表 1-2 Common Terminology Criteria for Adverse Event（CTCAE）v4.03

Grade1	Grade2	Grade3	Grade4	Grade5
わずかな紅斑や乾性落屑	中等度から高度の紅斑 まだらな湿性落屑. ただしほとんどが皺や襞に限局している	皺や襞以外の部位の湿性落屑; 軽度の外傷や摩擦により出血する	生命を脅かす; 皮膚全層の壊死や潰瘍; 病変部より自然に出血する	死亡

（a）



（b）

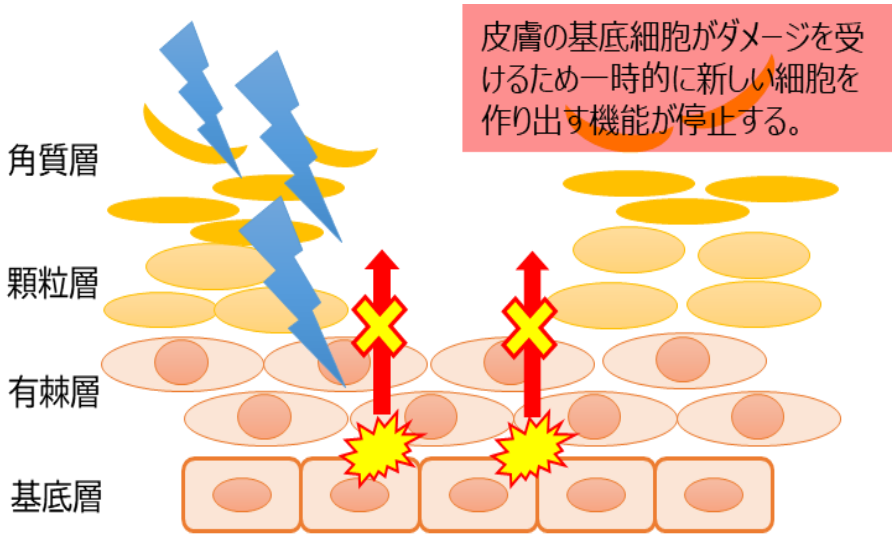


図 1-4 放射線治療線量分布図と放射線皮膚炎の発生機序
（a）放射線治療線量分布図　（b）放射線皮膚炎の発生機序

図 1-5 放射線皮膚炎の症例

1.2 放射線皮膚炎の評価におけるこれまでの研究

CTCAE に基づく放射線皮膚炎のグレード判定をするプロセスは，病理診断などに基づく確定診断と異なり，視覚的評価に基づいている．放射線皮膚炎は，図 1-6 に示す例のように分類問題としては，境界が不明瞭な症例であり不確実性をともなう. 例えば, 白と薄い灰色の境界を示すとする場合, 評価者によっては A, B で異なる結果となる．さらに濃い灰色と黒の境界となると C，E，または E とする場合もある．放射線皮膚炎もこのような皮膚の変色（紅斑など）によるものが多く，厳密な統一の難しさがある．そのため個人の知識や経験などに左右されやすく，評価の結果に個人差があると考えられる．これまで，CTCAE に基づく

評価統一を目的にした取り組み，報告も行われている．

また，近年では Artificial Intelligence（AI）が大きく進歩し，自ら学習して特徴量を作り出す畳み込みニューラルネットワーク（Deep convolutional neural network: DCNN）と呼ばれるディープラーニングが画像識別にも応用されており，医用画像分野でも多くの研究が行われている[11~13]．以下にこれまでの放射線皮膚炎の評価統一への取り組みと医用画像にディープラーニングを用いた研究について述べる．

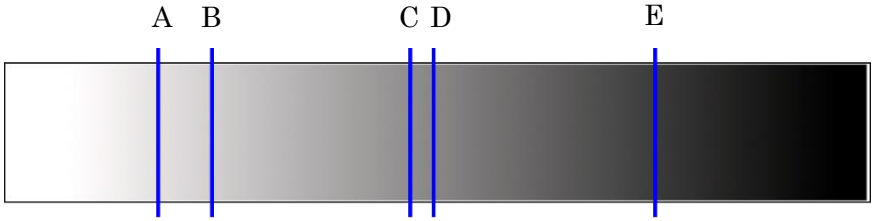


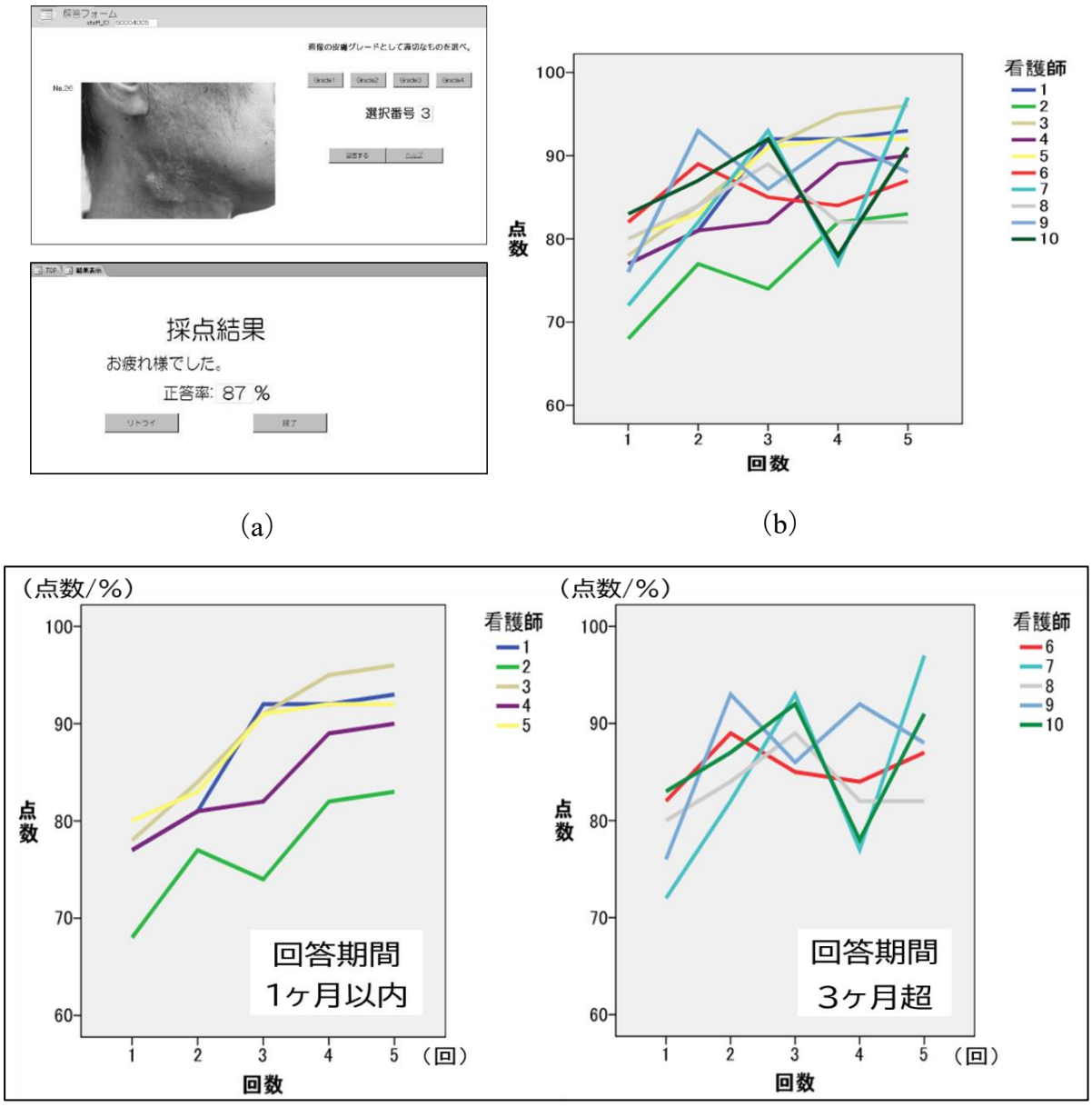
図 1-6 色に対する分類例

1.2.1 学習ソフトを用いた放射線皮膚炎の評価統一への取り組み

大薗ら[14]の施設では，患者の放射線皮膚炎に対する評価を主に看護師が行っているが，看護師間での評価が一致せず，現場で戸惑うことも少なくなかった．放射線皮膚炎の評価基準は存在するが，皮膚炎の評価は個人の知識や経験などに左右され，施設や組織全体にも大きな影響を与えうる．そこで，図 1-7（a）に示すように放射線皮膚炎の評価統一を目的として，放射線腫瘍学会認定放射線腫瘍医 2 名の評価を基準に厳選した放射線皮膚炎画像 100 枚を用いたクイズ形式の学習ソフトを作成した．看護師 10 名を対象に 5 回の学習を行い，正答率の推移を評価した．さらに，看護師を 2 グループに分けて実施間隔の違いによる正答率を評価した．

図 1-7 (b) に示すように学習を重ねる毎に正答率（平均値）は 77.3%から 89.9%に上昇し，学習効果が見られた．学習回数を重ねる毎に正答率は上昇する一方で，実施期間が空くほど正答率は低下し，評価結果に個人差がみられた．また，実施間隔について評価した結果，短期間で連続して行うと正答上昇率が高まるが，3

週間以上の空白で，学習した内容が抜け始めた．これらの結果より，人間の評価には個人差が生じ，より正しく放射線皮膚炎の評価を行うためには継続した学習が必要であると考えられた．



（a）　（b）

（c）

図 1-7 放射線皮膚炎評価の学習結果
（a）クイズ形式学習ソフト　（b）看護師 10 名の学習結果
（c）正答率の推移　（回答期間 1 ヶ月 vs.3 ヶ月）

1.2.2　放射線皮膚炎の視覚的評価基準アトラス作成への取り組み

Zenda ら[15]は，表 1-2 に示した CTCAE は，一文のみで重症度を説明し，個人の経験や知識による主観的な解釈のリスクがあるとして，4 施設からなる研究グループにおいて放射線皮膚炎の評価ツール開発を行ったと報告している．頭頚部がんの放射線治療を対象に視覚的評価による放射線皮膚炎のグレード判定を統一した基準で行う視覚的ガイドツールとしてグレーディングアトラスを作成した．

グレーディングアトラスは，グレードを判定する 4 名の放射線腫瘍医と 2 名の放射線治療に習熟した看護師が CTCAEver.4.03 に従って頭頚部がんの放射線皮膚炎症例の写真をグレード毎に分類し，グレードの判定基準となる写真を選択したテキストである．アトラス作成のために放射線皮膚炎の写真が 1,600 枚集められ，111 枚の代表的な写真を選択した．6 名の評価（グレード付け）が一致したものは，34 枚であり，評価者により過大，過少評価する傾向があったと報告している．また，アトラスに選択された写真はグレードの高い症例は少なく，グレード 3 が 6 枚，グレード 4 は 1 枚であった．そのグレードの判定基準を示すためにはより多くの症例が必要であることを記している．最終的なグレーディングアトラスの作成に 38 枚の写真が選択された．

図 1-8 に作成されたアトラスの一部を示す．Zenda ら[15]は，臨床で役立つためには，写真の品質が課題とされ，臨床研究プロトコルに取り入れるべきであると記している．また，グレードの高い症例が少なく適していない可能性があると報告している．図 1-2 に示したように放射線治療の有害事象は，急性期から晩期有害事象に至る．放射線皮膚炎は，治療中から症状が現れ始め，治療後までその程度が変化するため，適切なケアが必要である．しかし，治療後の放射線皮膚炎画像を収集することは，一般的ではないため写真収集および，品質の課題も生じると考えられた．

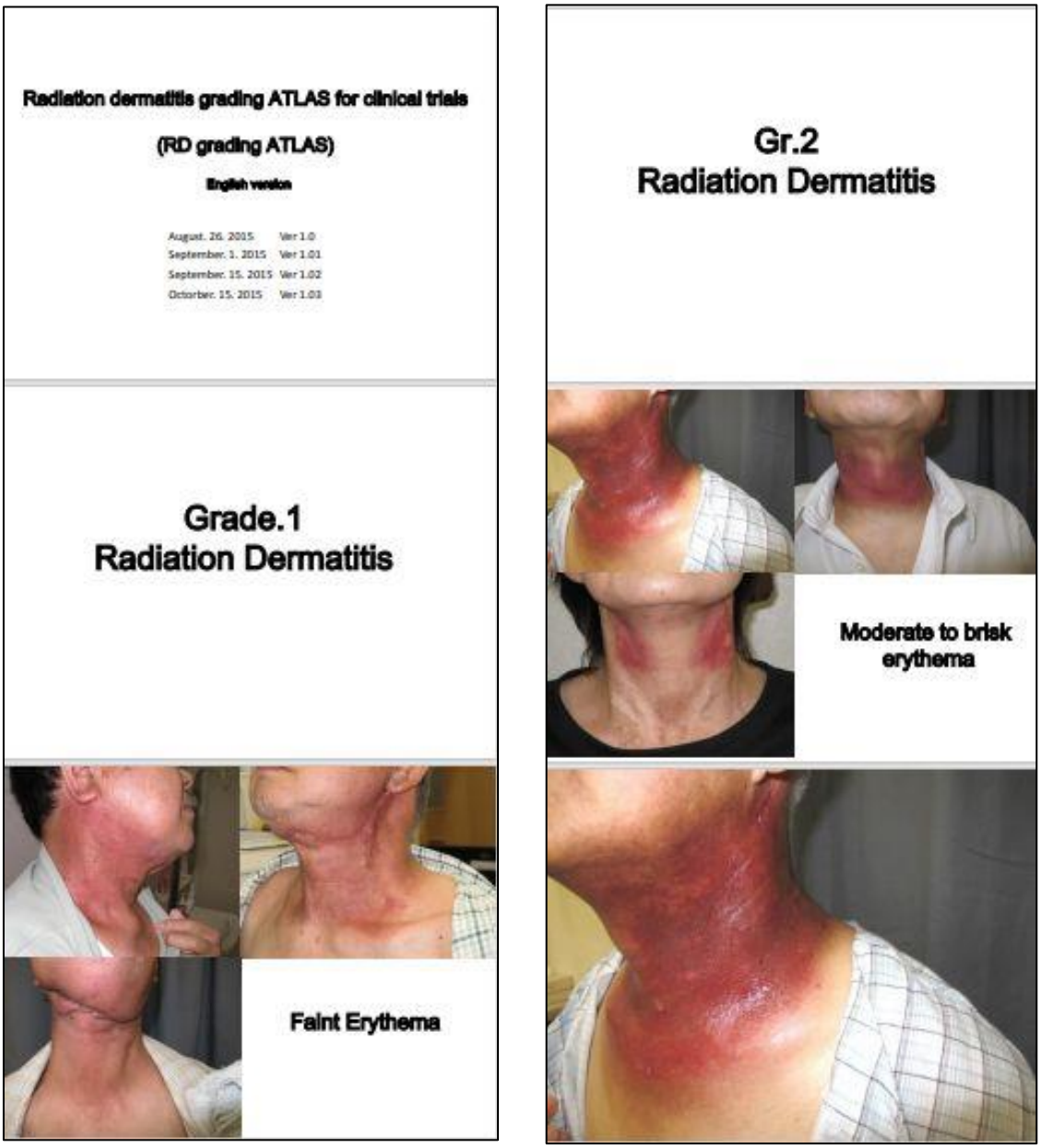


図 1-8 Zenda S, Yosuke Ota H, et al. A prospective picture collection study fora grading atlas of radiation dermatitis for clinical trials in head-and-neck cancer patients. Journal of Radiation Research 2016; 57: 301-306 より抜粋

1.3 本研究の位置付けと概要

本研究の目的は，CTCAEver.4.03 に準じて医師や看護師によってグレード判定された放射線皮膚炎データベースを利用して，DCNN を用いた放射線皮膚炎のグレード判定システムを構築することである．放射線皮膚炎の評価統一への必要性から，前項で述べたように人間のグレード判定の統一を目的に学習ソフト

が作成され，多施設においても評価統一できる視覚的な基準（グレーディングアトラス）が作成されている．人間の学習結果は，正答率（平均値）89.9%であったが，一方で，継続した学習が必要であり，評価結果に個人差が生じた．また，人間が行う評価では，評価経験が浅い評価者による学習不足，または体調不良などによる思考能力低下が評価に影響を与え，異なるグレード判定をする可能性がある．グレーディングアトラス作成による評価統一への効果は，評価されていない．これに対して，機械学習である DCNN は，データベースを基に人間の評価で起きる知識や経験の差（個人差）に依存しない判定が可能である．これは，放射線皮膚炎の管理を行う上で，セカンドオピニオンやサードオピニオンのような判定の補助システムとなることが期待できる．

本研究では，放射線皮膚炎の症例画像が収集されているデータベースを利用して，DCNN の学習画像を選択し DCNN に有効なデータセットを検討し，グレード判定システムの判定精度から有用性を評価する．

第 2 章では，近年，医用画像にも応用されている DCNN を用いた放射線皮膚炎の評価法について述べる．はじめに医用画像における DCNN の研究動向について調査し，DCNN を用いた画像識別の先行研究について述べる．先行研究より，放射線皮膚炎の評価を DCNN で行うための課題を整理する．これより，本研究の以下の課題 1～4 が明らかとなり，課題に対する対処を講じながら第 3 章，第 4 章では，課題に対処するため用いた手法を検証し，放射線皮膚炎のグレード判定システムを作成する．

課題 1．放射線皮膚炎グレード 1～4 における症例画像の収集と画像の品質

課題 2．不均衡なデータ数と少数画像の取り扱い

課題 3．稀な症例の取り扱い（極少数画像の取り扱い）

課題 4．放射線皮膚炎のグレード判定の相違

第 3 章では，課題 1，2，3 を解決するため，ハイブリッド生成法による DCNN を用いたグレード判定システムに関する研究について述べる．はじめに課題 1 を

解決することを目的として，医師や看護師によって収集画像の品質評価と収集画像のグレード判定を行う放射線皮膚炎画像選定プロトコルを定める．

DCNN の作成は，放射線皮膚炎の特徴を学習させる必要がある．しかし，放射線皮膚炎画像は，治療する部位によって発現場所が異なるため症例の多い部位や少ない部位によって，データ数が異なる．また，人間の共通した特徴である目や耳などの特徴抽出を避ける必要があるため，事前準備として治療部位に偏りが無いようにグループ分けされた学習用画像を準備する．課題 2 で述べたように放射線皮膚炎のグレード毎の症例数には偏りがある．この課題を解決する目的として，データの水増し処理を行い，データセットを作成する．データの水増し処理の有無について比較することにより，水増し処理の有用性を示す．次に課題 3 で述べた，特に極少数の重度のグレード 4 については，データ数がわずかである．そのため，水増し処理のみでは学習不足が生じてしまう．この課題を解決する目的として，正常皮膚画像に放射線皮膚炎画像の炎症部を埋め込む手法としてポアソン合成[16)]によって，人工的に症例画像を生成する画像処理（人工症例画像）を施す手法を考案し，水増し画像と人工症例画像をそれぞれ学習用画像としたデータセット，水増し画像と人工症例画像を混合させたデータセット（ハイブリッド生成法）によるデータセットを比較し，検証する．さらに第 2 章では，DCNN の出力結果を分析することを目的に Grad-CAM によるヒートマップを作成可能なニューラルネットワーク（convolutional neural network: CNN）と学習済みの VGG16（visual geometry group-16）[17)]の一部を再利用する fine-tuning を行い，新たな学習モデルを作成する．ハイブリッド生成法による放射線皮膚炎のグレード判定を補助するシステム（Hybrid generation method Radiation dermatitis grading support system: Hyb-RDGS）を開発し，Hyb-RDGS のパフォーマンスを検証し，有効性を示す．

Hyb-RDGS では，課題 1，2 および 3 の検討を行い評価しているが，複数のグレード判定であった放射線皮膚炎画像（課題 4）に対しては，モデルの効率性の

観点から検討に至っていない．そこで，課題 4 を解決する目的として，複数のモデルを作成して，ベイズ定理を適用したグレード判定手法を検討する．複数のモデルの結果を統合することで一つの分類器よりも予測精度の向上および汎化性能に有効な手法としてアンサンブル学習が知られている [18,19]．本研究では，アンサンブル学習の考え方に基づいて，複数のモデルを用いるが効率的な観点から一つのモデルの重み付けを変えたモデルを作成して行う．

複数のモデル作成は，近年 Mingxing ら [20] によって提案された EfficientNet モデルを用いたグレード判定システムを作成する．EfficientNet モデルは，パラメータの少ない特有のチューニング法（複合スケーリング法）を利用するため，効率的な複合スケーリング係数を持つ複数の重み付けの異なるモデルを作成可能である．従来の CNN のスケーリングでは，手動チューニングを行う必要があり，最適な値とはいえない場合もあった．これに対して EfficientNet は，ネットワークの構造を変えずに深さと広さと解像度の比率を固定してスケーリングアップしていく．そのため，他のパラメータ数をあまり増やすことなく精度をあげることが可能なモデルである．パラメータと計算量が他の DCNN よりも小さいため速度も向上している．また，第 3 章で提案する Hyb-RDGS は，ポアソン合成による人工症例画像を作成する必要があるため，効率性に欠けることが課題となった．そこで，第 4 章では，Hyb-RDGS よりも効率的なモデル検証を行うことが可能であると考え，重み付けされた EfficientNet モデルを用いてベイズ推定に基づく判定結果を出力する手法を提案する．はじめに，EfficientNet モデルに用いるデータセットを検討する．Ekin ら [21] によって自動的に水増し処理を選択してくれる手法として提案された Rand Augmentation（RA）を用いた水増し処理を行い，新たなデータセットを作成する．RA を加えたデータセットを作成し，最適な EfficientNet モデル構成を比較し，評価する．ここで，複数の重み付けの異なるモデルを選定し，課題 4 に対する評価方法にベイズ推定を用いることにより，グレード判定の相違のあった放射線皮膚炎画像の最終的なグレード判定を導く．

第１章

背景

放射線皮膚炎は，放射線治療の最も一般的な急性有害事象の一つである．
放射線皮膚炎のグレード判定は，CTCAE を用いてグレード判定を行うが，
基準の解釈，評価者の経験や知識による個人差が生じる．
近年，医用画像にも広く利用されるようになったディープラーニングを用いて放射線皮膚炎のグレード判定システムを構築できれば，人間の評価で起こる個人差に依存しない判定が可能になると考えられる．

放射線皮膚炎の評価におけるこれまでの研究

本研究の位置付けと概要

第２章

ディープラーニングを用いた放射線皮膚炎の評価

・ディープラーニングを用いた画像識別の研究

課題

1）放射線皮膚炎グレード 1～4 における症例画像の収集と画像の品質
2）不均衡なデータ数と少数画像の取り扱い
3）稀な症例の取り扱い（極少数画像の取り扱い）
4）放射線皮膚炎のグレード判定の相違

課題 1, 2, 3 の検討

・ディープラーニングを用いた放射線皮膚炎グレード判定システムの有効性

第３章

・ハイブリッド生成法によるディープラーニングを用いた放射線皮膚炎グレード判定システムに関する研究

課題 4 の検討

・グレード判定システムの高度化
・複数グレード判定画像に対する判定の導出

第４章

・EfficientNet モデルを用いたベイズ推定に基づく放射線皮膚炎グレード判定手法の開発

第５章

・放射線皮膚炎グレード判定システムの総括
考察 1. 先行研究と比較
考察 2. 開発システムの位置付けと今後の課題

第６章

結論

ディープラーニングを用いた放射線皮膚炎グレード判定補助システムを開発し，有効性の実証.

第2章　ディープラーニングを用いた放射線皮膚炎の評価

2.1　ディープラーニングを用いた画像識別の研究

2.1.1　DCNN 概説

DCNN とは，多層のニューラルネットワークによる機械学習手法である．多くの AI 技術は，この機械学習をベースとして成り立っているが，DCNN が他の機械学習技術として異なる点として，それ自体が特徴量を作り出すことができるようになった点が挙げられる．これまでの技術では，ある環境からパターン（特徴）認識を機械に学習させることが難しかったためである．DCNN は長い間解決されていなかったニューラルネットワーク特有の課題を”多層（ディープ）化”して，畳み込みネットワークやオートエンコーダーという技術などの CNN がディープラーニングと呼ばれている．

CNN は，人間の脳神経の構造を模倣した作りになっていることから，「ニューラル（神経系）ネットワーク」と呼ばれている．図 2-1（a）に示すように人間の脳は，基本的に神経細胞（ニューロン）と神経回路網（シナプス）で構成されており，ニューロンは電気信号として情報を伝達する．その時にニューロンとニューロンをつなぐシナプスのつながりの強さによって，情報の伝わりやすさが変わってくる．CNN において，重み（重みづけ）は，シナプス結合の強さを表し，学習によって重みはシナプスごとにその値が変化していく．

CNN は，「入力層」→「隠れ層」→「出力層」で情報の表現を行うが，それでは単純な情報しか処理，表現できないため，情報の複雑さに対応するように層の数を増やし精度を向上させる必要がある．入力に対して単純な変換を何回も繰り返し，予測結果などを出力する構造である．CNN の精度向上の鍵となるのは，深い構造，すなわち隠れ層を何層も重ねる構造である [22)]．CNN は，図 2-1（b）のように隠れ層は「畳み込み層」と「プーリング層」で構成される．畳み込み層で特徴マップを得た後，プーリング層は，畳み込み層から出力された特徴マップを，さらに縮小して新たな特徴マップを得る処理を行う．つまり，畳み込み層で

画像の局所的な特徴を抽出し，プーリング層は局所的な特徴をまとめあげる処理を行う．画像の持つ情報量を大幅に圧縮しながら特徴を維持することで，入力画像の認識，分類することが可能となる．

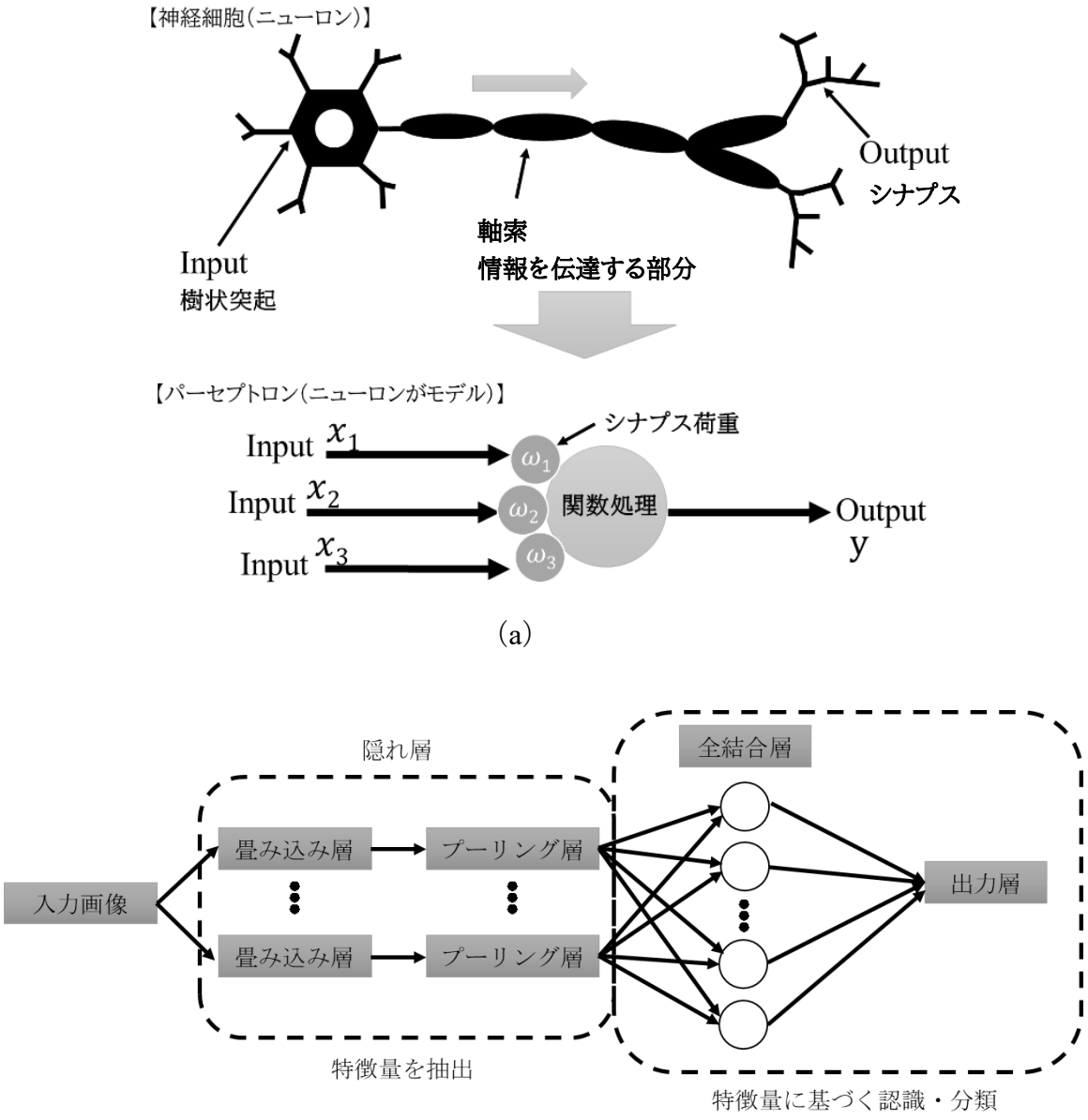


図 2-1 人工ニューロンとニューラルネットワークの構造
（a）人工ニューロン　(b)ニューラルネットワーク

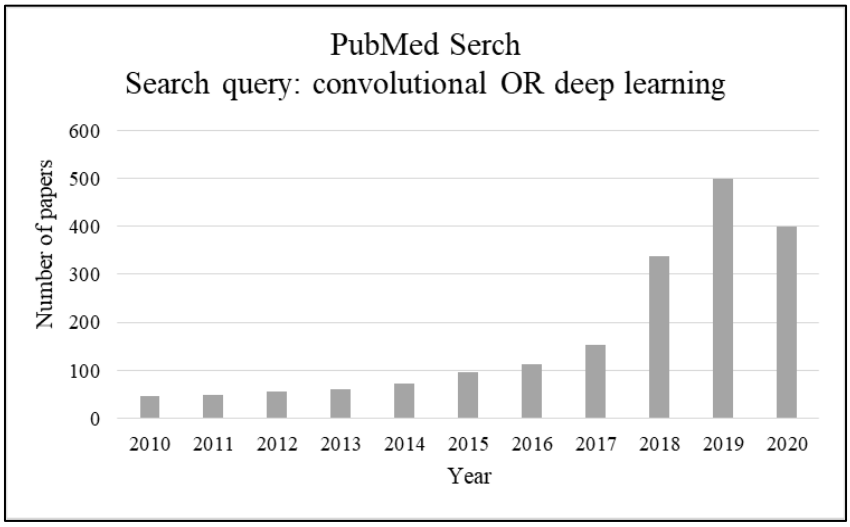
2.1.2　医用画像における DCNN の研究動向

近年のインターネットによるデータベースの利用の簡便さや拡充もあり，医用画像における DCNN の研究は，さらに加速している．2018 年には，それ以前の論文数を 2 倍以上回り，増加傾向にある（図 2-2 （a））．部位に限定されず，主に CT や MRI 画像といった診断領域において，多くの部位でセグメンテーションや分類，検出および生成などのジャンルで医用画像を用いた研究が行われている（図 2-2 （b），（c））．

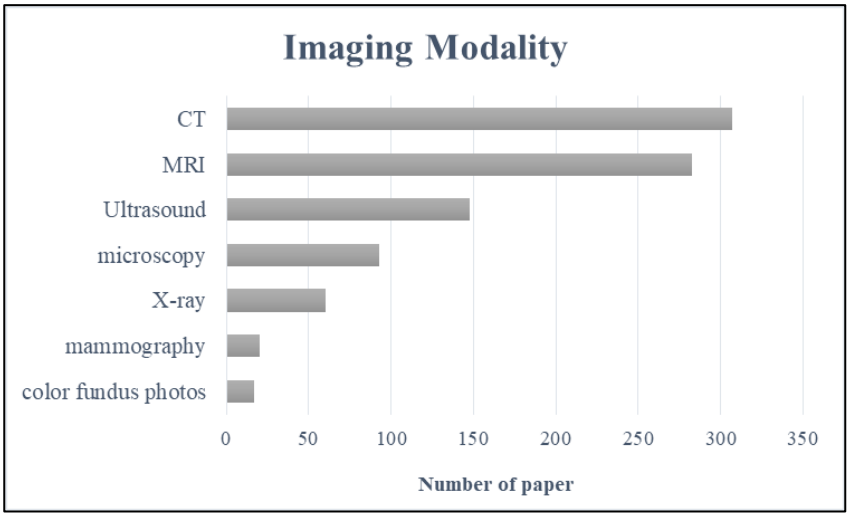
DCNN は多くの領域で活躍しており, AI の進化が果たす役割は計り知れない．一般物体認識と呼ばれる画像認識の分野から CNN や DCNN などを応用した様々な医用画像解析が行われている．医用画像解析には，医用画像から有用な情報を抽出する分類，セグメンテーション，検出そして生成などのタスクがある．その応用は，CT や MRI 画像など，画像の種類や部位について画像診断支援，治療支援として多くの研究が行われ CNN モデルを用いた研究には，多くの CNN モデルが提案されており，それぞれのジャンルで研究が行われている．

画像分類では, VGG や Res-NeT, 近年登場した EfficientNet モデルがある [17,20,23]．病変検出や異常検知では，R-CNN（Regions with CNN features）や YOLO，および CAD（Computer-aided Diagnosis），セグメンテーションには，FCN（fully convolutional networks）や U-Net が代表的なモデルとして有効性が報告されている [24~28]．また，画像生成については，2 つの CNN を使用して画像生成を行う敵対的生成ネットワーク（Genera tive Adversarial Networks: GAN）を用いた研究が盛んに行われている [29, 30]．

(a)



(b)



(c)

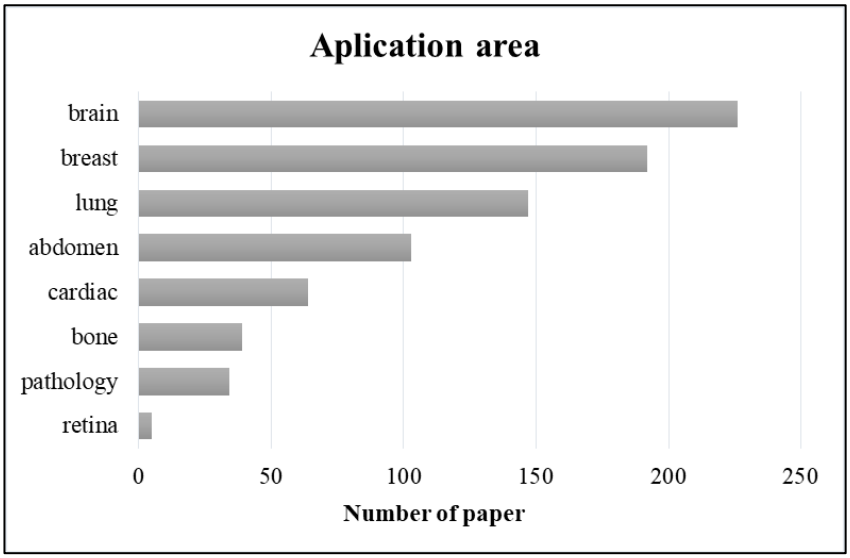


図 2-2 PubMed リサーチ結果
(a) 論文数 (b) 研究のモダリティ (c) 研究されている部位

本研究は，医用画像である放射線皮膚炎画像のグレード判定（分類問題）に DCNN を用いたグレード判定モデルを開発する．図 2-3 に医用画像研究ジャンルと主に用いられている CNN モデルを提示する．本研究では，グレード判定モデルとして, 図 2-3 で示す分類, 生成のジャンルについて, VGG16, EfficientNet, 画像生成法としてポアソン合成（poisson image editing）を用いる．

本研究

		ニューラルネットワークの代表的なモデル		
医用画像	分類 (classification)	VGG	ResNet	EfficientNet
	検出 (detection)	R-CNN	YOLO	CAD
	セグメンテーション (segmentation)	FCN	U-Net	
	生成 (generation)	GAN	画像処理 poisson image editing	

図 2-3 医用画像研究における代表的な CNN モデルと本研究の対象ジャンルと使用するモデル

2.1.3 DCNN を用いた画像識別の研究

米国では，5 人に一人が生涯に皮膚悪性腫瘍と診断され，年間 10,000 人以上の死亡の要因となっている．特に悪性黒色腫の 5 年生存率は初期の段階で発見された場合の 99%から，発見が遅れた場合，約 14%に低下するため早期発見が重要である [31]．このような背景から，発生機序や診断が放射線皮膚炎とは異なるが, 同じ皮膚疾患として皮膚写真やダーモスコピー画像（悪性黒色腫や基底細胞癌, 色素性母斑, 脂漏性角化症などの色素性皮膚病変での有用性は確率されている）[32] を使用して DCNN を用いた画像識別研究が行われている． Esteva ら [33]は，より早期に悪性腫瘍を検出できる DCNN を用いたシステムを構築した．

127,463 枚の学習画像と 1,942 枚の生検ラベル付きテスト画像に分割し，有効性を検証した．3 種類の皮膚腫瘍（良性，悪性，非腫瘍性病変）の良悪性を識別する DCNN を作成し，良悪性識別の正答率の平均が 72.1%，9 つの疾患分類は 2 人の皮膚科医が 53.3%と 55.0%であったのに対して，DCNN は 55.4%であったと報告している．

また，Fujisawa ら[34]は，DCNN を用いた 14 種類の皮膚腫瘍の良悪性を判定するシステムを構築した. 臨床では正確にラベル付けされた大量のものを収集すること，皮膚腫瘍の病理学的な診断は時間がかかり労力であるとして，Esteva ら[33]の学習枚数よりも大幅に少ない約 6,000 枚の臨床皮膚画像を用いた DCNN を構築したと報告している．学習されたシステムの 14 種類の皮膚腫瘍の良悪性識別の正答率は，日本皮膚科認定皮膚科医専門医 13 名と比較したところ，皮膚科専門医が 85.3%であったのに対して，DCNN の識別率は 92.4%，識別の難易度が高い 14 種類の皮膚腫瘍分類においては，皮膚科専門医が 59.7%であったのに対して，DCNN の識別率は 76.5%であったと報告している．

DCNN を用いた画像識別の可能性は，専門医が不足する地域では，写真による診断を可能とし，皮膚がんの早期発見へ繋がる．また，診療する上で手助けとなりセカンドオピニオンやサードオピニオンといった補助システムとして期待される．これらの皮膚腫瘍の診断補助システムのように，放射線皮膚炎のグレードを AI で判定できる可能性が考えられる．

2.1.4　画像生成手法

これまで述べたように，昨今，DCNN の性能向上は著しいが，それに伴い必要なデータ量も増加しつつあり，一般的に大量の学習データが必要と言われている．事前学習によく用いられる ImageNet は約 100 万枚と，それなりに規模が大きい. データ量が少なければ，データ収集の労力や学習に要する時間や負荷も大幅に節約できる．このような少数のデータ量を用いて学習を行う場合によく用

いられる手法として，学習データを拡張させることによって，データ量を増やす Data Augmentation（DA），いわゆる”水増し”とよばれるデータ拡張手法が知られている．Fujisawa ら [34]は，学習画像を 15 度ずつ回転させ，ぼかしフィルタを 0～5 ピクセル，明るさを-10%～＋10%にランダムに変動させることによって，24 倍に水増しを行っている．水増しとは，元の学習データに変換を加えて画像のバリエーションを増やし，データ量を増やす手法であり，特に CNN などを使った画像処理で効果を発揮する．水増し手法には，以下のようなものがある．

- ガウシアンノイズ（ノイズを増やす）
- ソルト&ペーパーノイズ
- コントラスト強調
- 角度変換
- ガンマ変換（明るさを調整）
- 平均化フィルタ（平滑化）
- 拡大・縮小
- 反転（左右・上下）
- シフト（水平/垂直）
- マスク
- 回転
- 変形

これらの画像処理を組み合わせるなどして，学習に最適なデータ拡張画像を加えながら DCNN の性能を向上させる必要がある．一方，これらのデータ拡張の最適値を探す手法として，Auto Augment などの手法がある．Ekin ら [35] は，最適な拡張となるパラメータを探索するコストが非常に大きいという欠点があるとして，自動的に DA を選択してくれる手法として Rand Augment（RA）を提案している．RA で使用される Augmentation（=transformation）手法は，表 2-1 に示す 14 個であり，最適な 2 つのパラメータ N 選択する transformation の数であり，14 個から N 個を選択），M （Augmentation を強さ（0～10））で制御される．RA は，最適な N と M をグリッドサーチで見つけることで，10 の 2 乗オーダーで最適なデータ拡張を見つけ出すことを実現している [21]．

表 2-1 Rand Augmentation の tansformation

・identuty	・autoVontrast	・equalize
・rotate	・solarize	・color
・posterize	・contrast	・brightness
・sharpness	・shear-x	・shear-y
・translate-x	・translate-y	

また，DA などのデータ拡張とは異なり，敵対的生成ネットワーク（GAN）と呼ばれる乱数からフェイクの学習データを生成する手法が注目されている[29,30]．しかしながら，GAN を適用するためには，特定のクラスの学習用画像をクラスの分布特性を特定できる程度にデータを用意しなければならない．データが少ないとオーバーフィッティングによるモード崩壊が起こりやすいといわれている．本研究では，症例数の少ない画像データが与えられている状況を想定する．すなわち，GAN を利用することができないものとする．そこで，図 2-3 で示したように与えられた画像データを手がかりにして，画像処理を用いた人工症例画像を生成する方法を採用する．

2.1.5 内部特徴の可視化

DCNN のようなモデル使用する場合，信頼できるシステムを作成するためには，透過的なモデルでなければならない．近年では，2 次元または 3 次元マップ内の位置情報を与えることにより，高次元データを視覚化する t-SNE（t-Distributed Stochastic Neighbor Embedding）[36] や VGG を用いて入力画像の推論を行う Grad-CAM（Gradient-weighted Class Activation Mapping）[37] が用いられ，その判断根拠とした特徴を可視化することができる．本研究では，VGG を用いることにより，Grad-CAM で出力された結果に対する判断根拠を可視化する．

Grad-CAM では図 2-4 に示す畳み込み層やプーリング層を繰り返し最後に全結合層に接続してクラス分類を行うようなモデルに対して，全結合層の前の畳み込み層で生成された特徴マップが，予測したラベルに対してどれくらい影響

を与えているかを以下のように勾配を使って計算を行う[38].

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1.1)$$

$$L_{Grad-CAM}^c ReLU(\sum_k \alpha_k^c A^k) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1.2)$$

式(1.1)では，クラス c に対して特徴マップ（ヒートマップ）の各要素が微小に変化した際にクラスの確率がどれくらい変化するかを計算して特徴マップ内で平滑化を行い, そして式(1.2)でその値を使って特徴マップ領域内の重要度を計算する. これにより, モデルの予測に対して入力画像のどの領域が影響を与えるかを計算して可視化するヒートマップが出来上がる. 本研究では，このヒートマップを作成することにより，放射線皮膚炎の特徴を認識して判定結果を出力しているのか，その判断根拠を可視化する.

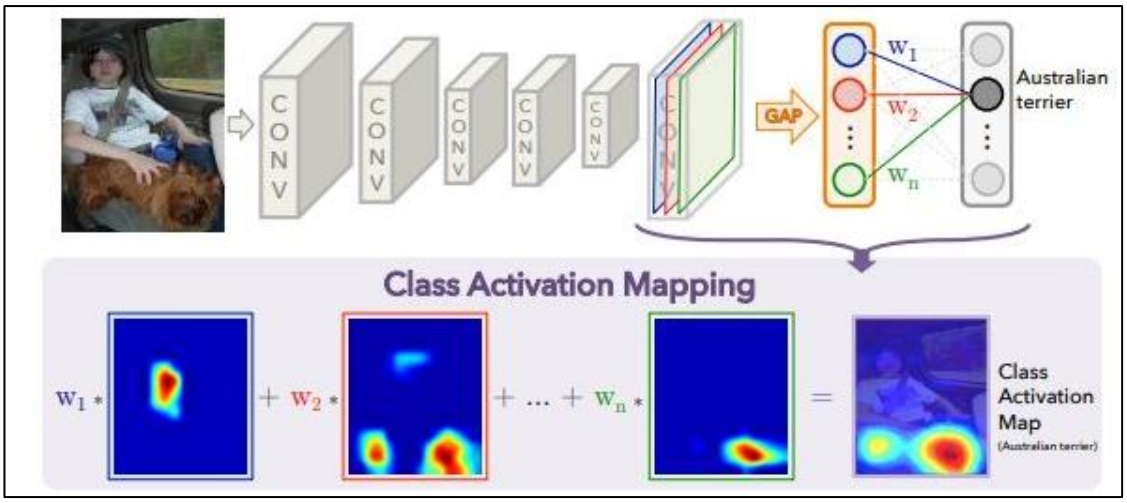


図 2-4 Grad-CAM 概要（Bolei Zhou, Aditya Khosla, Agata Lapedriza et al. Learning Deep Features for Discriminative Localization. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; pp.2921-2929 から抜粋）

2.2　DCNN を用いるグレード判定のシステム化への課題

前述で述べたこれまでの研究によって，放射線皮膚炎の評価統一への様々な取り組みが行われる一方，医用画像の識別に DCNN が用いられるようになり，

人間の識別精度を上回るシステムの研究が報告されている．我々は，データベースを基に放射線皮膚炎のグレード判定を行うシステムを構築すれば，人間の評価で起きる知識や経験の差（個人差）に依存しない判定が可能であると考えた．これまで放射線皮膚炎画像データベースを利用した AI に関する研究は，我々の知る限り報告されていない．しかし，これまでの研究でも明らかになっているように，放射線皮膚炎のグレード判定をシステム化するには幾つかの課題がある．以下にその課題を整理する．

（課題 1）放射線皮膚炎グレード 1～4 における症例画像の収集と画像の品質

写真などの一般的な画像は，インターネットを通じてデータセットが公開されていることも多く，比較的入手は容易であるが，医用画像はあくまで一部のデータセットが公開されているだけで，必要なデータセットが公開されていないことも多い．放射線皮膚炎画像もこのような公開データある訳ではなく，放射線治療の有害事象として症例の割合としては，極少数でもある．公開データセットを用いずに大量の放射線皮膚炎症例を得るためには，多大なコストが必要ともいえる．また，放射線皮膚炎の症例は，発症時期や日々の変化を伴うため，症例も一様ではない．表 1-2 に示したように軽度のグレード 1～4 までグレード分けされ，図 1-2 に示したように，その発症は治療中～治療後に至る．放射線皮膚炎は，放射線の線種や，どれくらいの線量が照射されるかによって，症状の程度は異なる．特に治療終了 1～2 週間後にその症状はピーク時期を迎えるため，治療後も継続して経過観察が重要となる．適切な皮膚ケアにより，ほとんどの放射線皮膚炎は，約 1 カ月前後でほとんどの症状が回復に向かう [4]．しかし，治療後の経過観察は，看護師の指導を受けて，患者自身によるセルフケアが行われる場合がほとんどであるため，治療後の症例画像の収集は容易ではない．

また，Zenda ら [15] は，収集した画像（写真）について，写真の品質をプロトコルに入れるべきと記されている．皮膚画像を評価する際の撮影するカメラの

種類，設定内容（露出，解像度など）や撮影環境または，撮影範囲などは，画質の差異や正しく観察できないなど評価精度に影響を与える．本研究では，放射線皮膚炎画像は，カメラで撮影された写真であるため，写真の品質は，ボケや歪といった低品質のものが含まると予測できる．以上より，本研究では，低品質画像における取り扱いを決めておくことが望まれる．DCNN を用いたシステムの構築において，正確な学習画像選択が重要であり，システムの性能を左右する因子である．

（課題 2）少数画像と不均衡なデータ数の学習

DCNN は，大量の学習画像が必要になる．しかし，医用画像の分野では，インターネットなどによる公開データがあるのは，一部であり，少ないデータ数および不均衡なデータパターンに依存することがいわれている．

本研究で扱う放射線皮膚炎は，近年では，放射線皮膚炎の予防的なケアが行われるようになり，重度の放射線皮膚炎の低減が報告されている [10]．有村ら [39]は，271 名の前立腺がん患者に対して放射線皮膚炎予防目的の処置（フィルムドレッシングの貼付）を行った結果，グレード 0，1，2，3 の放射線皮膚炎が，予防処置無し群それぞれ 0，65，57，4 名，予防処置有り群においては 2 名，122 名，21 名，0 名の患者に起こったと報告しており，予防的なケアにより重度の皮膚炎の減少したことを報告している．このような症状軽減への研究，皮膚炎の初期症状であることからグレード 1～2 の皮膚炎症例が全体のほとんどを占めている．グレード 1～4 全ての症例を均等かつ，より多く収集することは，困難であり多大な労力であることがいえる．Trueman[40] は皮膚へのダメージを考慮した治療計画と皮膚ケアの介入により，グレードの高い放射線皮膚炎の症例が少ないと述べている．このような理由から，放射線皮膚炎のグレード毎の症例は，不均衡なデータ数として対処する必要がある．

不均衡な学習データとは，このように目的とするデータ分布に偏りが生じる

ことをいう．発生する可能性が極めて低い事象を予測対象とする場合に不均衡データが生じる．不均衡データの学習では，各クラスのデータ数が不均衡であることが問題となるため，多い方のクラスのデータ数を減らすか，少ない方のデータ数を増やすことで，各クラスの数を均衡にするという自然な発想である．多い方のクラスのデータ数を減らす手法をアンダーサンプリングといい，少ない方のデータ数を増やす手法をオーバーサンプリングという．不均衡データを取り扱う手法は，大きく 2 つに分けられ，ここでは代表的なアプローチであるアンダーサンプリング，オーバーサンプリングについて述べる．

- アンダーサンプリング

  アンダーサンプリングは，少ない方のクラスのデータ数に合うように多い方のデータ数の中からデータをサンプリングすることで不均衡を抑制する．単純なサンプリング手法を用いると過学習に陥りやすくなるため，クラス毎のパフォーマンスに応じてサンプリングレートを変更する [41]．アンダーサンプリングは，比較的簡単なだけでなく，不均衡さを取り除くことができ，さらに学習データを小さくし，学習コストを減らすことができるといわれている．

- オーバーサンプリング

  オーバーサンプリングの代表的な手法として良く知られている SMOTE（Synthetic Minority Oversampling Technique）がある．SMOTE は，少ない方の各データ点同士を線でつなぎ，その線分上の任意の点をランダムに人工データとして生成する手法である [42]．

本研究では，第 3 章でデータ数の多いクラスに対してアンダーサンプリングを行い，データ数の少ないクラスでは，オーバーサンプリングとして画像処理を

用いたデータ拡張を行う．本研究で扱う放射線皮膚炎の少数データは，部位が限定されていることやデータ数が極少数であるため SMOTE を用いない画像処理によるデータ拡張法と新たなポアソン合成[16]を用いた人工症例画像生成法を提案する．

（課題 3）グレードの高い（重度）症例の学習

Zenda ら[15]のグレーディングアトラス臨床研究において収集された放射線皮膚炎症例のグレード 4 は 1 名であった．放射線皮膚炎の重度症例数は少なく，特にグレード 4 の症例画像を収集することは極めて困難であると考えられる．グレード 4 は，表 1-2 で示すように生命を脅かす，皮膚全層の壊死や潰瘍の状態であり，臨床において稀な事象ともいえる．少数データを用いて学習を行う際に一般的に用いられるデータ拡張が挙げられるが，元画像が少ない場合，学習したモデルは学習データをいわゆる丸暗記してしまう．そのため，学習データに対しては高い予測精度を示すが，学習データに潜む特徴を学習していない未知のデータに対して予測精度が低くなってしまう．より多くの元画像を使用して学習データを増やせば，全てのデータに共通している特徴を学習するようになる．放射線皮膚炎の重度の症例は，極度に少ないため丸暗記状態に陥ってしまうことが考えられる．本研究におけるグレード 4 の症例については，その対処が必要である．このような少ないデータの特異性を手掛かりに検出を行う研究も行われている[43]．

（課題 4）放射線皮膚炎のグレード判定の相違

CTCAE に基づく放射線皮膚炎のグレード判定を行うプロセスは，病理診断などに基づく確定診断と異なり，視覚的評価に基づいている．また，CTCAE は，表 1-2 に示したように単文のみで記載されている．

大薗ら[14]が報告したように放射線皮膚炎のグレード判定のプロセスは，個人

の知識や経験などに左右されやすく，評価者の解釈によって評価結果に相違が生じることがある．明瞭な境界値が存在しない評価基準が曖昧であるため，医師や看護師による評価者によって複数のグレード判定が生じる．グレード判定に相違が生じることは，一般的なことである．Zenda ら[15]の報告でも施設によって判定が異なる．このような場合，最終的なグレード判定は，最終的な判断を下す評価者（主に医師）にて決定されることが多いと考えらえる．

DCNN を用いたグレード判定システムの構築には，正確なグレード判定画像（正解ラベル）を用いて正確な学習を行うことが必要とされ，精度の高い性能をもつシステムとなる．しかし，複数の評価者によって，グレード判定が異なる判定結果であった画像に対する評価方法は，曖昧であり確立されていない．本研究で作成する放射線皮膚炎グレード判定システムが，グレード判定に相違があった場合にセカンドオピニオンの役割として助言（判定結果を示すこと）ができれば，人間の評価と DCNN の評価を合わせた，より正確なグレード判定結果をもたらすことができると考えられる．

このような理由から，複数のグレード判定結果であった放射線皮膚炎画像の評価手法を検討する意義がある．これまで，述べた本研究の課題を図 2-5 に模式図で示す．

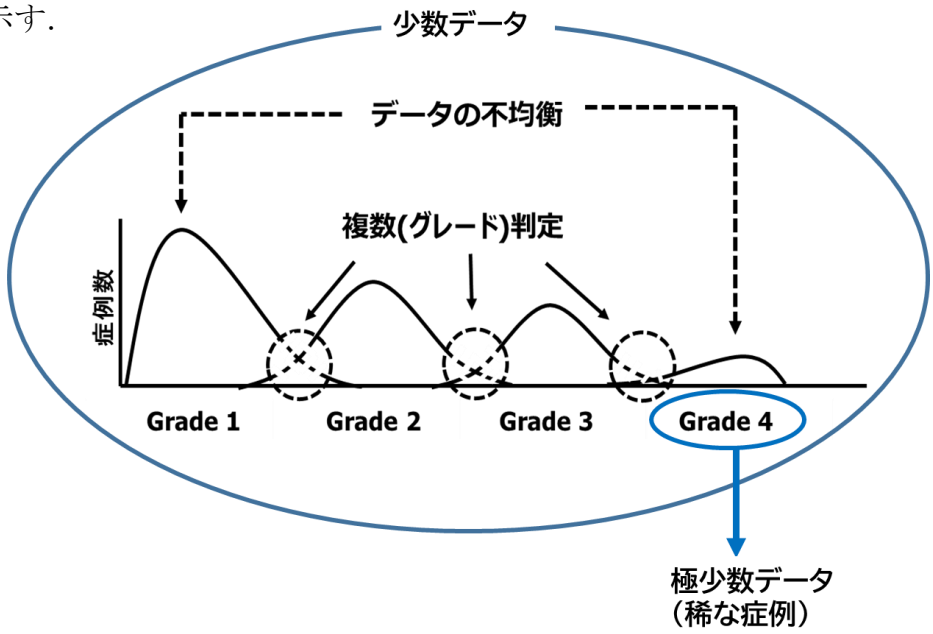


図 2-5 DCNN を用いた放射線皮膚炎グレード判定システムの課題模式図

第3章　ハイブリッド生成法によるディープラーニングを用いた放射線皮膚炎グレード判定システムに関する研究

3.1　緒言

本章では，ポアソン合成によって作成した人工症例画像と一般的に用いられている水増し処理（DA）画像を混合したデータセットを用いる Hyb-RDGS の作成について述べる．

はじめに，課題 1 に対して，放射線皮膚炎データベースから臨床皮膚画像を画像選定プロトコルに沿って評価し，それぞれの画像に対してグレード判定を行い，前処理を行った上で学習画像を作成する．

次にデータセットを作成するため，課題 2 のグレード毎の不均衡データに対して，アンダーサンプリングと概ね同じ割合となるようデータ拡張を加えたオーバーサンプリングを行う．さらに課題 3 の少ない重度の症例（グレード 4）に対しては，ポアソン合成による人工症例画像を生成し，データ数を補う手法を提案する．

DA の有効性を実証し，人工症例画像を用いたデータセットについて検証するため，DA と人工症例画像を学習画像に用いた複数のデータセット，DA と人工症例画像を混合させたデータセットを比較，検証する．さらに，内部特徴を可視化する目的で Grad-CAM を用いて，グレード判定結果を出力し，ハイブリッド生成法による Hyb-RDGS の性能を評価する．図 3-1 に Hyb-RDGS 作成のワークフローを示す．

3.2　Hyb-RDGS 作成のワークフロー

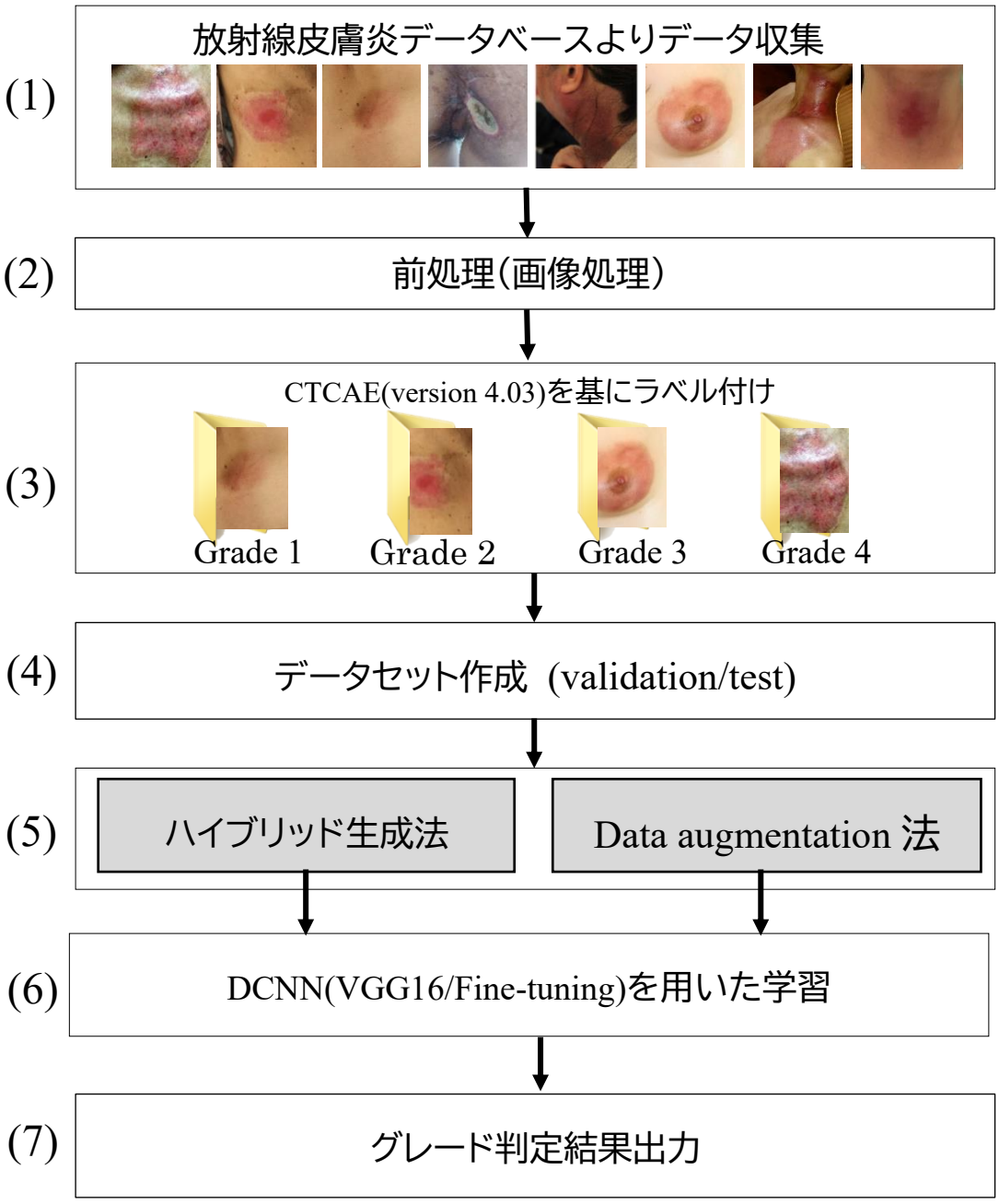


図 3-1 Hyb-RDGS 作成のワークフロー

3.3 構築環境

DCNN は，オペレーションシステムが Windows10 Home64 ビット版であるパーソナルコンピュータに Anaconda Navigator をインストールして Python3.7.3 の仮想環境を構築し，DCNN 用のソフトウェアライブラリである Keras2.2.2 を用いて実施した. CNN と学習済みの VGG16 モデルの一部を再利用する Fine-tuning を行い，新たな学習モデルを作成する．図 3-2 に VGG16 の構造を示す.

DCNN の学習は，バッチサイズと呼ばれるグループに分けられた数に対して，エポック数と呼ばれる回数だけ繰り返す．本研究では，バッチサイズを 32，エポック数を 300 回とする．事前学習でテストデータにおける Accuracy の低下，Loss の上昇が起きていないことから過学習状態に陥っていないことを確認した.

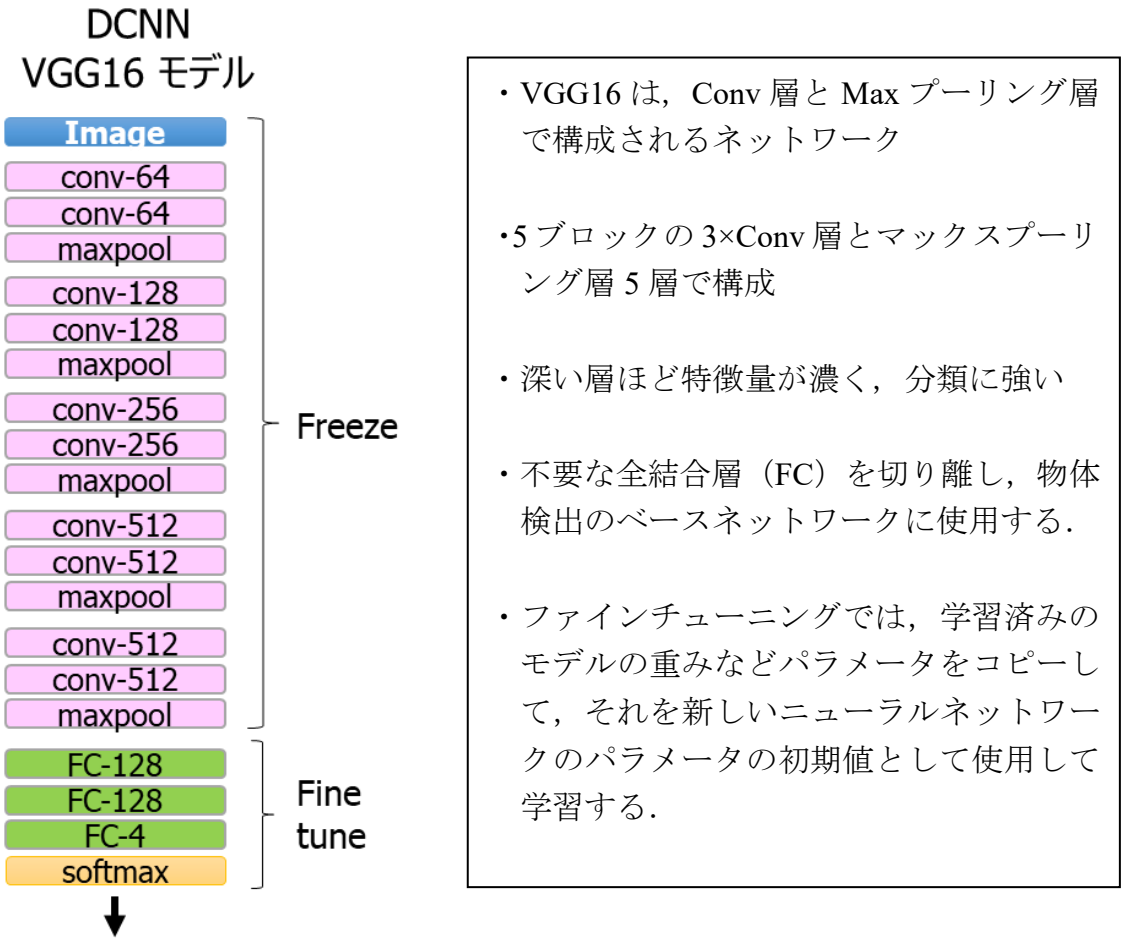


図 3-2　VGG16 の構造

3.4　データセット作成

ここでは，DCNN のデータセット作成するにあたり，前節の課題 1 と 2，および課題 3 に対するデータの取扱いについて述べる．はじめに，放射線皮膚炎の画像データ収集と前処理について述べる．次にデータセット作成に用いる手法について述べた後，ハイブリッド生成法によるデータセット作成手法を提案する．なお，データセットは，3.4.4 項で示す学習用，テスト用に分割する．

3.2.1　データ収集と前処理

事前準備として，DCNN のデータセットに用いる学習画像のデータ収集と前処理を行う．放射線皮膚炎画像は，メディポリス国際陽子線治療センターの陽子線治療情報管理システムのデータベースに治療の経過写真として患者毎に管理されており，治療部位と対比できるようになっている．写真の品質は，カメラの撮影条件や撮影環境が決められていないため，解像度や明るさ，撮影範囲など一定ではない．なお，本研究はメディポリス国際陽子線治療センター倫理審査委員会の承認を得ている（審査番号：R2019-04）．

学習画像は，正確なラベル付けを行わなければ，学習結果の劣化に繋がり，学習画像の境界によって識別性能に影響を及ぼす．例えば，放射線皮膚炎のグレード判定が間違った画像，図 3-3 に示すような炎症部以外の特徴を持つ目や鼻が多く含まれていると識別性能が変わってくる可能性がある．学習画像の与え方によっては不十分な学習になるため，学習画像にどのようなデータを含めて，どのようなデータを含めないかデータ収集と前処理が重要となる．

そこで，本研究では，データ収集における基準を設けるため放射線皮膚炎画像の選定プロトコルを策定した．以下に，本章で行うデータセット作成のデータ収集における放射線皮膚炎画像の選定プロトコルと前処理を示す．

図 3-3 人間の特徴を含む放射線皮膚炎画像例

＜放射線皮膚炎画像選定プロトコル＞

(1) 2011 年から 2019 にかけてメディポリス国際陽子線治療センター（単施設）で放射線治療を受けた患者の臨床皮膚画像であること.

(2) データベースから抽出された放射線皮膚炎画像のグレード判定は，放射線腫瘍学会認定放射線腫瘍医 2 名，放射線治療に習熟した看護師 4 名のそれぞれがグレード 1～4 の分類を行う.

(3) 画質が不鮮明な画像やグレード分けが評価者によって異なる画像は，学習画像から除外する．すなわち，評価者全員が一致したグレード判定を受けた画像のみ使用する.

＜前処理＞

(1) 選定された放射線皮膚炎画像のトリミング（等倍）を行う．トリミングは，炎症部を中心に行い，炎症部以外の特徴を持つ構造体が極力含まれないように注意する.

(2) 撮影のし方によるボケや歪といった画質不良による正確に観察できない画像は除外する.

(3) 目や鼻などの人物の顔の特徴，および乳房形状など放射線皮膚炎以外の人物

の特徴に対する学習を避けるため，学習画像を部位別に頭頸部，体幹部および乳房と胸壁領域の 3 つに分類する．

(4) 画像は，jpg 形式で統一する．

表 3-1 収集データ数と除外したデータ数
() は，複数のグレード判定を受けたデータ数

体幹部	グレード1	グレード2	グレード3	グレード4	除外
	418	50	139	31	31
	(34)				
		(13)			
			0		
頭頸部	グレード1	グレード2	グレード3	グレード4	除外
	192	30	126	0	7
	(0)				
		(5)			
			0		
乳房と胸壁領域	グレード1	グレード2	グレード3	グレード4	除外
	406	48	28	0	48
	(7)				
		(0)			
			(0)		
計	グレード1	グレード2	グレード3	グレード4	除外
1468	1016	128	293	31	86

表 3-1 より，収集データ数は 1,468 枚である．これより，放射線皮膚炎画像選定プロトコルと前処理を行った．グレード 4 は，体幹部領域のみ 31 枚であり全体の約 2%である．また，グレード判定が正確に行うことができない低品質とされる画像は，データ収集時に除外している．

3.2.2　不均衡データと少数データの取り扱い

データ収集，前処理されたデータは，表 3-1，前節の課題 2 で示したように放射線皮膚炎の症例データは一般的な DCNN に必要なデータ量に対して少数である．課題 2 を解決する目的でデータ収集における放射線皮膚炎画像の選定プロトコルと前処理が行われた画像を使用し，データ拡張を施す.

本検証では，一般的に用いられる複数の画像処理を用いた DA 処理を採用する．各グレードの画像に対してコントラスト強調と低減，左右反転，平行移動，角度，画像の一部を長方形上の欠損およびソルト&ペッパーノイズ，ガウシアンノイズ付加処理の中から最大 3 種類の画像処理を行う事で実施する．図 3-4 に DA 処理の一例を示す.

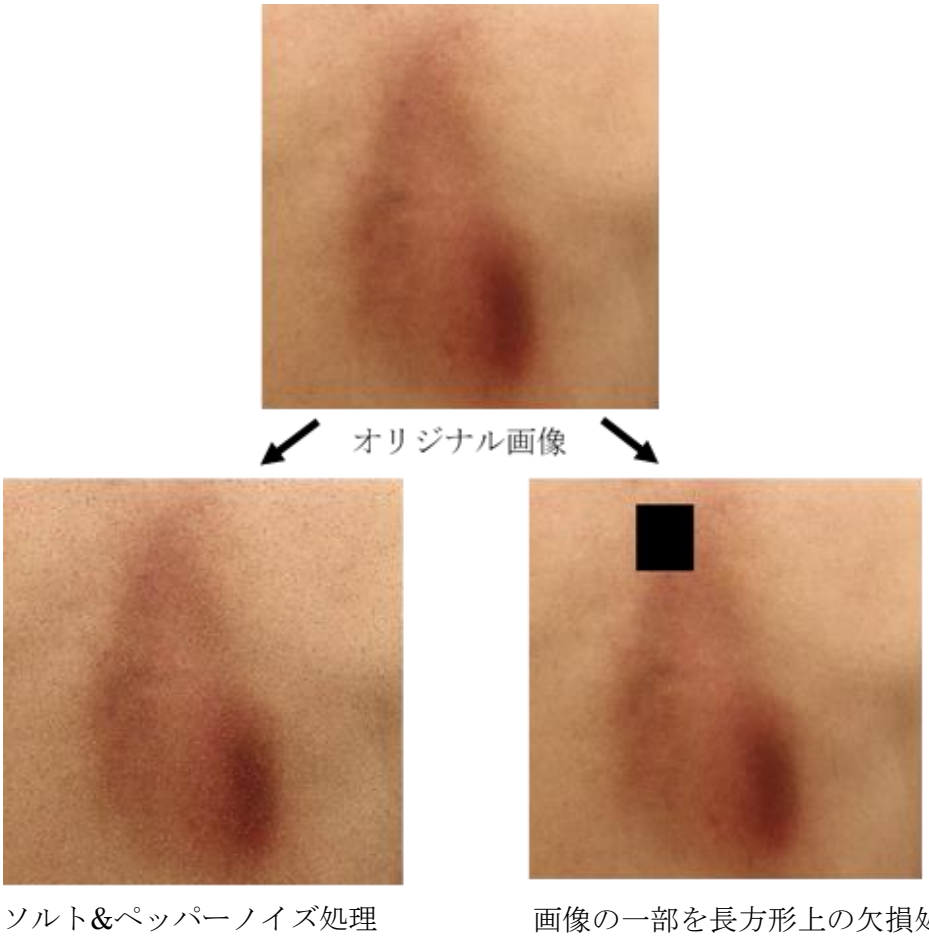


図 3-4 Data Augmentation 法によるデータ拡張の一例

また，不均衡データを緩和する目的にデータレベルでの手法を用いて，データセット作成時にサンプリングを行う．データ拡張操作は，いわゆるオーバーサンプリングとして，データ拡張の割合（サンプリングレート）をグレード毎に変化させるデータ拡張を施す．一方，少数画像に対しては，最もデータ数の多い低グレードの放射線皮膚炎画像をランダムに学習画像から除外するアンダーサンプリングを行う．サンプリングは，治療部位に偏りが生じないようにするため，前処理でグループ分けされた 3 つの領域毎にサンプリングレートを変えながら工夫をすることで，不均衡データの緩和効果を得る．

3.2.3　稀な症例画像の取り扱い（極少数画像の取り扱い）

本研究では，既存の症例画像に対して DA 処理を行うとともに，ポアソン合成 [33] によって生成した人工症例画像も学習画像として利用する手法を提案する．具体的には，正常な皮膚画像に極少数画像であるグレード 4 の放射線皮膚炎の炎症部を埋め込む画像処理を施す．ここで，正常な皮膚画像は，データベースに登録された放射線治療前の正常な皮膚画像を使用する．また，不均衡データを避けるため，放射線皮膚炎画像の部位で分類した頭頸部，体幹部および乳房と胸壁領域のグループ毎に作成を行う．

ポアソン合成は，未知の特定領域画像のラプラシアンからポアソン方程式を作成し，ターゲット境界部分の画素値を境界値としてポアソン方程式に基づいて解く手法である．ポアソン合成は，式（3.1）を用いてターゲット周辺である放射線皮膚炎領域を計算する．計算領域を Ω とし，ある点を p，その近傍の点を q，fq，$f'q$ はそれぞれの画素値を指す．その集合を Np，ターゲットと正常領域の境界領域を $\partial\,\Omega$，点 p と点 q の差を Vpq とし，画素値 fp を求める．

$$f_p = \left(\frac{\sum_{q \in N_p \cap \Omega} f_q + \sum_{q \in N_p \cap \delta\Omega} f'_q + \sum_{q \in N_p} \mathbf{v}_{pq}}{|N_p|} \right) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (3.1)$$

これによって，正常画像の持つ画素値の勾配を維持しつつ，ターゲット領域周辺との境界を滑らかにすることが可能となり，図 3-5 に示すような単純合成に対して，違和感のないポアソン合成画像を生成できる．

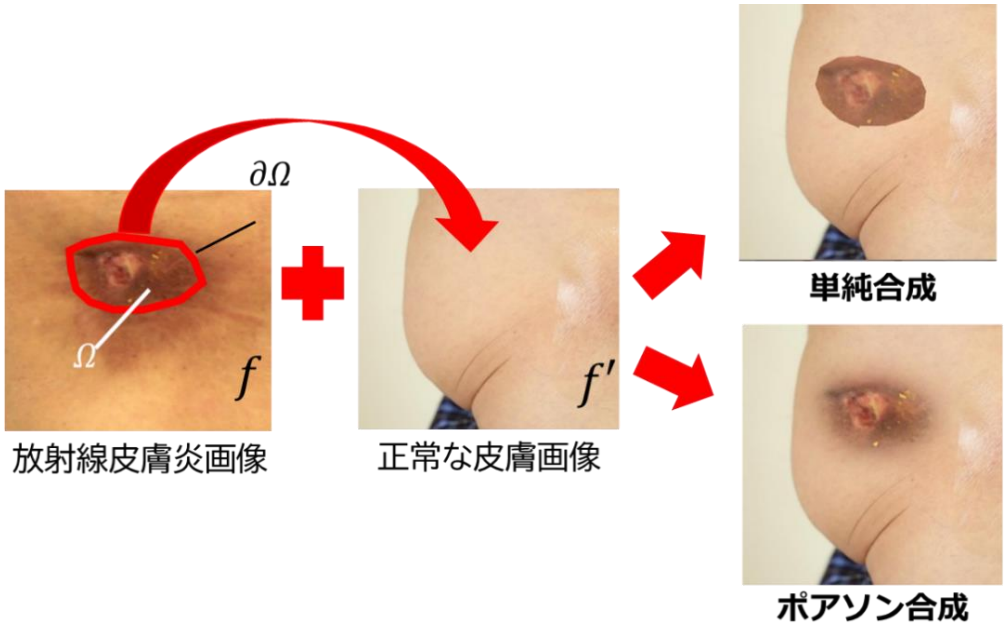


図 3-5 ポアソン合成を用いた人工症例画像の生成

ポアソン合成[16)]：与えられた画像 f の領域 Ω を切り取って，正常画像 f′へ Ω を貼り付ける．ここで，領域周辺の画素値∂Ω から領域内側の画素値を推定し，滑らかに合成する．

3.2.4 ハイブリッド生成法によるデータセット作成

はじめに，DA 処理の有効性を評価するため，3 クラス（グレード 1～3）と 4 クラス（グレード 1～4）の異なるクラス数による比較と DA の有無による比較用データセット概要を表 3-2（a）に示す．

次にグレード 4 を対象にデータ生成手法の異なる比較用データセット概要を表 3-2（b）に示す．ここで，作成したデータ生成手法は DA 処理（D），ポアソン合成（E），ハイブリッド生成による DA 処理とポアソン合成画像の混合（F）である．

表 3-2 データセット概要

（a）異なるクラス分類の比較用データセット

（b）グレード 4 におけるデータ生成比較用データセット

(a)

クラス	グレード	データ拡張の有無	データセット No.
3クラス	Grade1, 2, 3	none	A
		Data Augmentation処理	B
4クラス	Grade1, 2, 3, 4	none	C
		Data Augmentation処理	D

(b)

クラス	グレード	データ生成手法	データセット No.
4クラス	Grade1, 2, 3, 4	Data Augmentation処理	D
		ポアソン合成	E
		DA処理+ ポアソン合成	F

前項で述べた前処理と学習データ（Vlidation data）とテストデータ（test data）に分割したデータセット詳細について表 3-3 に示す. Hyb-RDGS のグレード 4 のクラスの学習画像には，DA 処理で水増しした画像 56 枚とポアソン合成によって水増しした画像 200 枚を使用する.

表 3-3 に表 3-2 で示したデータ拡張，生成法による DA 処理やアンダーサンプリングによって調整した学習画像数，Hyb-RDGS の検証に使用した学習画像数を示す．No. C，D，E，F は，それぞれのテストデータセット（test data）105 枚を 3 セット作成した.

表 3-3 データセットパラメータ

			Original data	Validation data	test data	データ生成
3クラス	A／B	Grade 1	450	450／450	30	N/A
		Grade 2	83	83／249	30	N/A ／ Data Augmentation ×3
		Grade 3	101	101／322	30	N/A ／ Data Augmentation ×3
4クラス	C／D	Grade 1	450	450／450	30	N/A
		Grade 2	83	83／249	30	N/A ／ Data Augmentation ×3
		Grade 3	101	101／322	30	N/A ／ Data Augmentation ×3
		Grade 4	13	**21／78**	15	N/A ／ Data Augmentation ×3
	E	Grade 1	450	450	30	N/A
		Grade 2	83	249	30	Data Augmentation ×3
		Grade 3	101	322	30	Data Augmentation ×3
		Grade 4	13	**213**	15	ポアソン合成 +200
	F	Grade 1	450	450	30	N/A
		Grade 2	83	249	30	Data Augmentation ×3
		Grade 3	101	322	30	Data Augmentation ×3
		Grade 4	13	**256**	15	ポアソン合成 +200 Data Augmentation +56

3.3 Hyb-RDGS の出力

Hyb-RDGS は, Hyb-RDGS が行うグレード判定結果を図 3-6 に示すように対象画像のグレードとその判定結果の内訳を示した円グラフ，対象画像上にグレード判定に用いた内部特徴をハイライト表示するヒートマップと呼ばれる画像の 3 点として出力される. ヒートマップは, VGG16 の特定の畳み込み層において, 入力画像の推論を行う Grad-CAM を実装することで作成する.

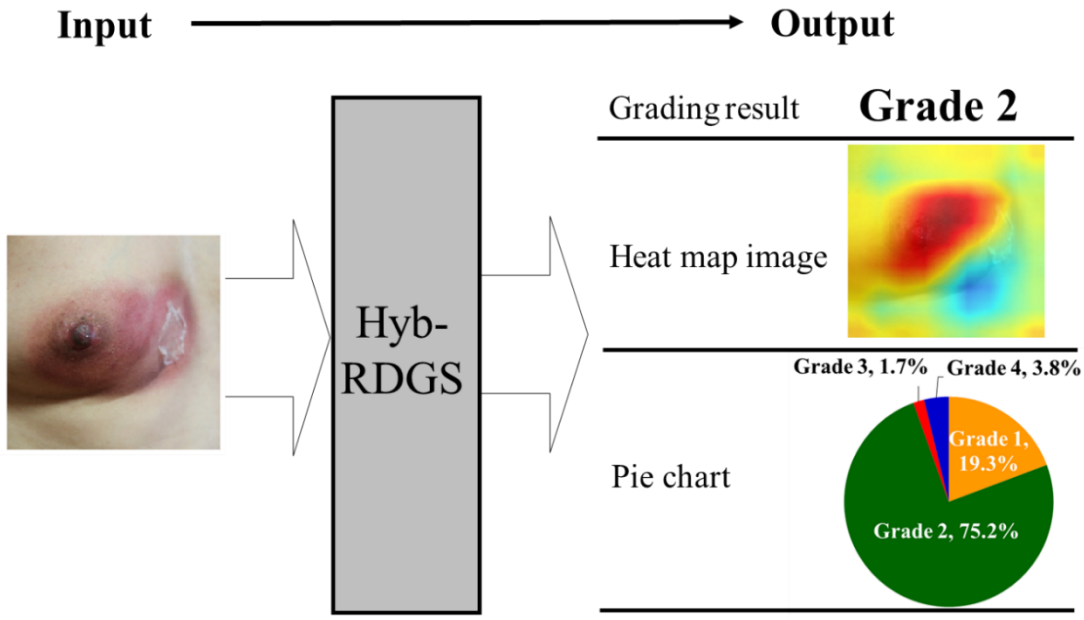


図 3-6　Hyb-RDGS のグレード判定方法の概要

3.4　Hyb-RDGS の学習性能評価，検証方法および精度評価

ここでは，Hyb-RDGS の学習評価性能，検証方法およびモデルの精度評価について述べる．本研究で行うモデルの検証は，前項で述べたぞれぞれのデータセットについて，データ全体を学習データ（validation data）とテストデータ(test data)に分割して行う．

DCNN の目的の一つは，学習データ（既知のデータ）の特徴量（説明変数）を使って目的変数を予測するモデルを作り，結果が未知のデータを予測することである．作成したモデルの性能を表せる数値指標が，アルゴリズムの妥当性などを評価することができる．数値によるモデルの性能評価は不可欠といえる．モデルの評価指標は，幾つかあるが大別すると分類モデルと回帰モデルに分類することができる．回帰モデルの評価は，実測値と予測値の差異（残差）を使った指標が一般的に用いられている．分類モデルの場合，目的変数がどのカテゴリに属するかを予測する．分類モデルの性能評価でよく作成される評価指標に混同行列や ROC 曲線がある．これについては，3.6.3 項で述べる．

本研究の目的は，放射線皮膚炎のグレード判定を行うため，分類モデルとして性能評価を実施する．また，性能評価用の評価値を得るために学習性能評価を行い，汎化性能を検証する必要がある．本節では，学習に使用する検証データとテストデータを用意して学習性能を評価し，解析の妥当性を含む検証に Hold-out 検証，交差検証法（cross-validation）を用いる．

3.4.1　学習性能評価

学習性能評価は，学習曲線（learning curve）を指標として評価される．学習曲線は検証データのサンプル数と予測性能の関係を示したグラフであり，図 3-7 に示すように横軸は学習データのサンプル数であり，縦軸は評価指標である．学習曲線は，予測モデルが過学習を起こしているのか，それとも学習不足になっているのかを判断する指標になる [44)]．CNN の学習においては CNN の出力と正解のラベルの差を意味する loss を最小化することが目標になる．つまり，学習時にニューラルネットワークは train loss を小さくする方向に学習を行う．

図 3-7 の学習曲線を見ると，およそ 50 epoch 付近からわずかに test loss が上昇していることが確認できる．このように train loss と test loss が学習時に乖離していく曲線は，過学習と呼ばれる．特定のデータで過学習してしまい，そのデータに依存したモデルとなってしまう．この汎化誤差が大きくなることをバリアンス（variance）が大きいという．これとは，反対にデータに対して学習不足でモデルが単純すぎて，データ量が十分でない汎化誤差が大きいケースをバイアス（bias）と呼ぶ．このバリアンスとバイアスのバランスを取りながら学習を行うことで有能な学習性能を得られる．本システムでは，事前学習からエポック 300 とした．

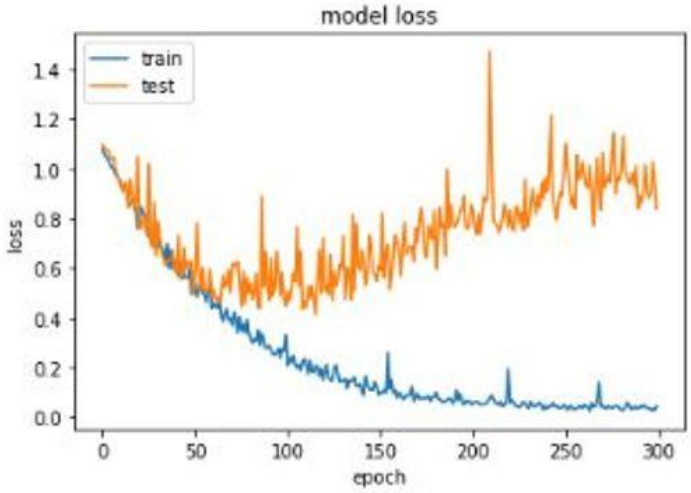


図 3-7 過学習

3.4.2 検証方法（Hold-out 検証と k-分割交差検証（k-hold cross validation））

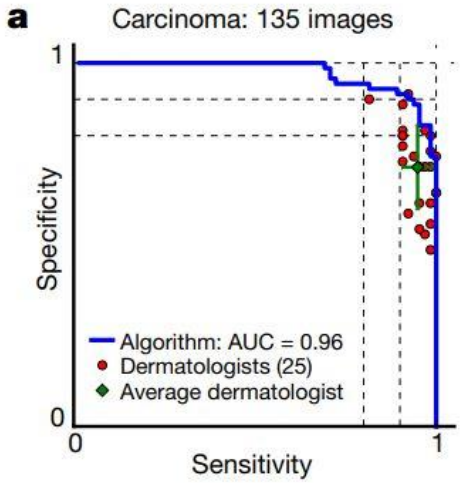
評価方法によって，分割の方法には幾つか手法があるが，本研究では Hold-out 法，k-分割交差検証を用いる手法を用いる．ホールドアウト法は，モデルを作る学習データとモデルを評価するテストデータを 7 対 3 などの割合で 2 分割し，学習したモデルで予測する際に，学習に使っていない未知のデータで予測する．データを分けることで，汎化性能を向上させることができる．ホールドアウト法は大量のデータセットがあり，モデルの推論に時間がかかる場合などに利用されている．k-分割交差検証は，汎化性能を評価する統計的な手法で，分類でも回帰でも用いることができる．k-分割交差検証は，データを k 個に分割してそのうち 1 つをテストデータに残りの k-1 個を学習データとして正答率の評価を行う．これを k 個のデータすべてが 1 回ずつテストデータになるように k 回学習を行なって精度の平均をとる手法である．小規模な画像データを用いて評価する場合に用いられている [45)]．

本研究では，事前に異なる多クラス分類の比較を行い，画像処理によるデータ拡張の有効を確認するためにホールドアウト法を用いて検証する．本検証である Hyb-RDGS には 3 分割交差検証を用いる．

3.4.3　精度評価（混同行列（confusion matrix））

分類モデルの精度評価は，前項で述べた DCNN によらず，機械学習や統計においても混同行列，ROC 曲線（Receiver Operating Characteristics Curve）が評価指標として用いられている．

先行研究における皮膚がんの分類では，Esteva ら [33] は，ROC 曲線と ROC 曲線の下の領域面積を用いて一つの値で表現する AUC（Area Under the Curve）を用いた精度評価を行っている．一方，Fujisawa ら [34] は，混同行列を用いている．図 3-8 に ROC 曲線と混同行列の例を示す．不均衡なモデルに対して分類モデルを評価する場合，目的に合わせて混同行列を使用するべきであり，不均衡データの問題がないとき，ネガティブデータが極端に多い場合，実精度より高く評価されてしまうケースがある [46]．本研究では，不均衡なデータを取り扱うことから，混同行列を用いて評価する．



(a)

Table 2 Performance of the deep convolutional neural network (DCNN) in second-level classification

	Diagnosis			
DCNN output	Malignant epithelial	Melanoma	Benign epithelial	Benign melanocytic
Malignant epithelial	538	16	25	14
Melanoma	10	69	6	6
Benign epithelial	7	4	98	10
Benign melanocytic	7	6	27	299
Total number	562	95	156	329
Accuracy (sensitivity)	95·7%	73%	62·8%	90·9%

The panel colour indicates the percentage within each column:

> 50% | 10–50% | 5–10% | 0–5%

(b)

図 3-8　ROC 曲線と混同行列の例

（a）ROC 曲線：Fujisawa Y, Otomo Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. British Journal of Dermatology 2019; 180: 373-381 より抜粋

（b）混同行列：Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks Nature 2017; 542: 115-118 より抜粋

3.4.4　Hold-out 検証と k-分割交差検証（k-hold cross validation）

混同行列は, 2 値分類問題で出力されたクラス分類の結果をマトリクス表にまとめることによって，学習モデルの性能を評価する指標として利用されている．混同行列は, はじめに「真のクラス」を一つ決めて作成する．例えば, クラス A,B を予測する 2 値分類を考えると,「クラス A を真とした混同行列」,「クラス B を真とした混同行列」が作ることができる．つまり，混同行列はクラスの数だけ作ることが可能である．本研究では，これを利用することで，3 クラス，4 クラスの多クラス分類へ応用する．次に 3 クラス分類の場合の混同行列と評価値の計算方法について述べる．

3 クラス分類における混同行列

ここでは，3 つのクラスがあり，入力がどれに分類されるか推定を行う．グレード判定の分類するタスクを以下の例を使用する．

- TP: True Positive，真のものを真と予測した（正解）
- FP: False Positive，偽のものを真と予測した（不正解）
- TN: True Negative，偽のものを偽と予測した（正解）
- FN: False Negative，真のものを偽と予測した（不正解）
- （neutral，特に分類が無いようなもの）

ここで，混同行列を作成するために一旦「真のクラス」を positive とする．このとき，混同行列は表 3-4 のように作成できる．

表 3-4　3 クラス分類における混同行列

		モデルのtest結果		
		positive(**真**)	negative(**偽**)	negative(**偽**)
真のグレード判定	positive(**真**)	TP	FN	FN
	negative(**偽**)	FP	TN	TN
	neutral(**偽**)	FP	TN	TN

3 値分類を考えた場合，クラス名も 3 種類であり，混同行列も 3×3 になる．一方で，真偽に注目すると，2 種類であるため要素をまとめると混同行列は 2×2 のサイズに凝縮できる．すなわち，FP，FN，TN を合計して一つの要素にまとめることができる．表 3 -5 に示すように「真のクラスが negative」を含む 2×2 に凝縮した混同行列となる．

表 3-5　2 クラス分類に凝縮した混同行列

		モデルのtest結果	
		positive(真)	negative(偽)
真のグレード判定	positive(真)	TP	FN
	negative(偽)	FP	TN

混同行列の評価値

混同行列の評価値は幾つかあるが，本研究では，Hyb-RDGS の各クラスに対する正解率，感度，再現率，F1 値で評価を行う．（3.2）式～（3.5）式を用いて算出する．ここで，感度と再現率はトレードオフの関係にあり，多くの Positive を多くだす学習モデルは，感度は高くなるが，適合率は低くなり，多くの negative を出す学習モデルは，適合率は高くなるが，感度は低くなる．感度と再現率を一方に偏ることなく均等に評価したい場合に，そのバランス指標である F1 値を用いて総合的な精度評価が可能となる．これらは，1 に近いほど精度が高く，0 に近いほど精度が低いことを表す．

ここで，i は対象のグレード，n はその個数である．算出した結果は，Welch t-test または t-test を用いて有意差検定を行い，P-value< 0.05 は有意とみなした．作成したそれぞれのデータセットについて，混同行列を用いた精度評価を行うことで Hyb-RDGS の性能評価を比較することができる．

- 正解率（*Accuracy*）

  全ての予測のうち，正解した予測の割合

$$overall\ accuracy = \frac{\sum_{i=1}^{n}(TruePositive_i + TrueNegative_i)}{\sum_{i=1}^{n}(TruePositive_i + FalseNegative_i + FalsePositive_i + TrueNegative_i)} \quad \cdots\cdots\cdots\cdots (3.2)$$

- 感度（*Sensitivity*）

  実際に陽性であるもののうち，正しく予測できたものの割合

$$sensitivity_i = \frac{TruePositive_i}{TruePositive_i + FalseNegative_i} \quad \cdots\cdots\cdots\cdots (3.3)$$

- 適合率（*Precision*）

  陽性であると予測したもののうち，正しく予測できたものの割合

$$precision_i = \frac{TruePositive_i}{TruePositive_i + FalsePositive} \quad \cdots\cdots\cdots\cdots (3.4)$$

- F1 値（*F1-measure*）

  感度と適合率の調和平均（感度と適合率の両方を加味したバランス指標）

$$F1\ value_i = 2(\frac{sensitivity_i \times precision_i}{sensitivity_i + precision_i}) \quad \cdots\cdots\cdots\cdots (3.5)$$

3.5　画像処理，拡張および生成結果

3.5.1　収集データ

放射線皮膚炎データベースより, 収集した放射線皮膚炎臨床画像は図 3-9 に示すように使用するカメラや撮影環境は様々である．さらに明るさやボケなどの品質不良画像も含まれている．これは，一般的に放射線皮膚炎は，治療終了後は患者自身によって撮影されることもあり，画像の収集に加えて撮影環境の統一の難しさによるものである．

図 3-9　収集した放射線皮膚炎臨床画像

3.5.2　データ拡張（DA 処理）

収集した画像は，放射線皮膚炎画像選択プロトコル従って，グレード付けを行い，前処理を実施した．ここで，評価者によって品質不良画像は，表 3-1 に示したように除外している．また，本節では，複数のグレード判定であった画像は除外し，第 4 章で取り扱う．学習画像として用いる前処理を実施した画像に対して，図 3-10 に示すように DA 処理でデータ数の拡張を行った．

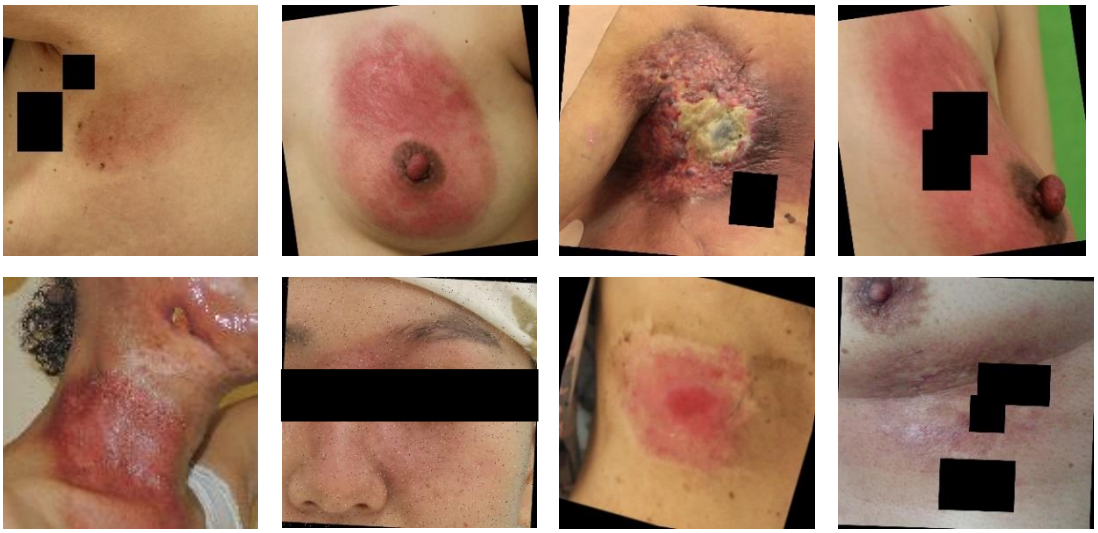


図 3-10　収集した放射線皮膚炎臨床画像の DA 処理例

3.5.3　ポアソン合成手法による人工症例画像生成

本章では，課題 3 に対してポアソン合成により，人工症例画像を生成する手法を用いた．図 3-11 に示すようにグレード 4 の炎症部を正常皮膚画像に埋め込む処理を実施した．

図 3-11　ポアソン合成により作成した人工症例画像

3.6　検証の結果

グレード 1～3 の 3 クラスおよび 1～4 の 4 クラスについて，DA の有無による正答率を比較した．表 3-6 に DA 法の評価結果，図 3-12 に学習曲線を示す，

3 クラスにおいては，DA の有無によって正答率は 86.6%と 76.0%となり，DA によって正答率は 10.6%向上した. 4 クラスにおいては, 正答率はそれぞれ 83.4%, 74.4%となり，正答率は 9.0%向上した．両クラスにおいて，DA の有無によって正答率には有意な差が見られた．

学習画像データ数の DA の効果について，DA を行わない場合に対して DA 法では，3 クラスおよび 4 クラスの正答率が有意に向上したことから，DA が正答率の向上に寄与している事が示された．一方で，DA 法における 4 クラスの正答率は 3 クラスのそれより 3.2%低下した．

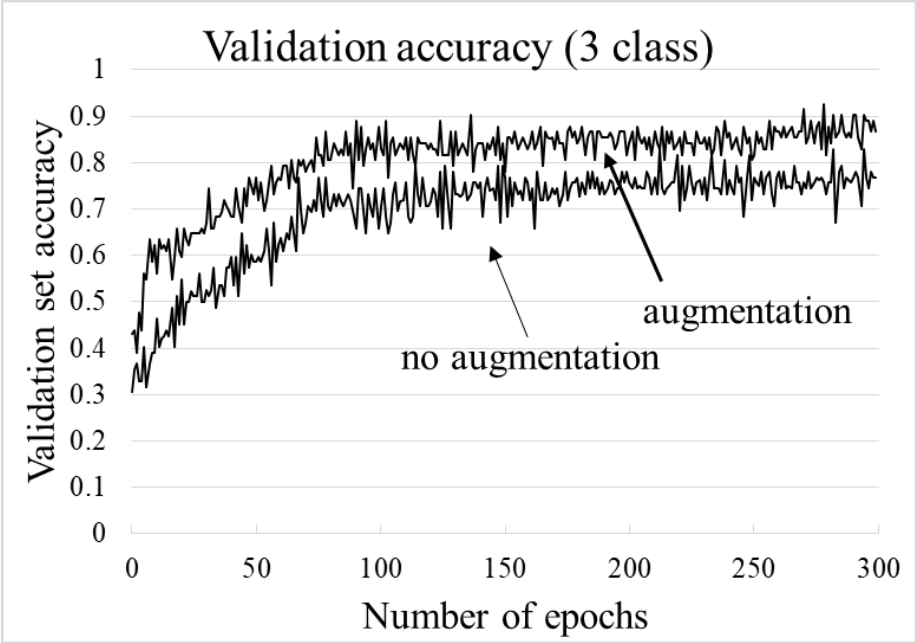
表 3-6　DA 法の性能評価結果

Classification Accuracy

	No Augmentation	Augmentation	
3 classes (Grade1, 2, 3)	76.0%	86.6%	*
4 classes (Grade1, 2, 3, 4)	74.4%	83.4%	*

*; P<0.01

(a)



(b)

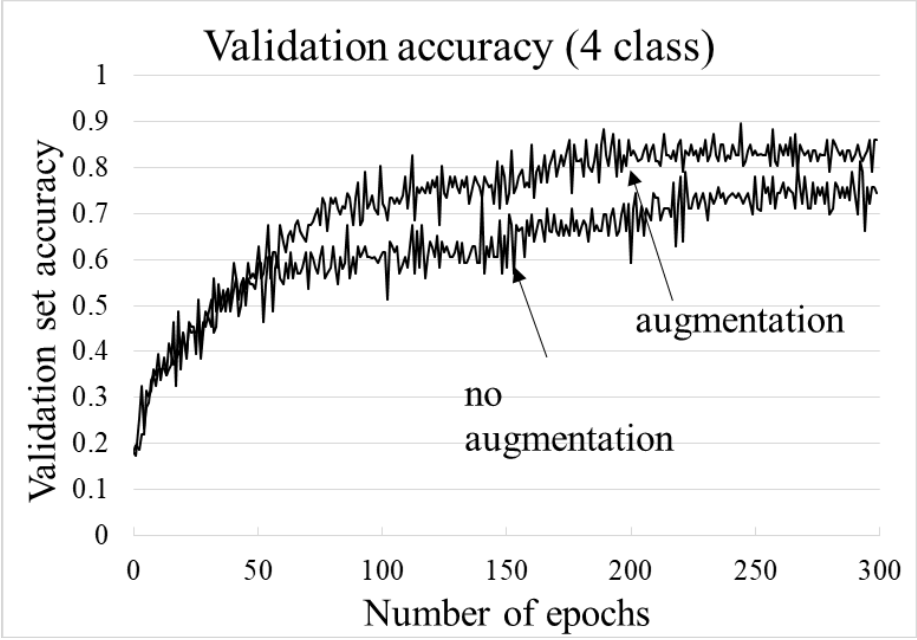


図 3-12　学習曲線（DA 有無の比較）
(a) 3 クラス　(b) 4 クラス

グレード 4 に対してデータ生成法の違いによる学習性能について，表 3-7 に示すように D（DA 法）と E（ポアソン合成）の validation accuracy は，0.834 を示し，最も高い性能であった．一方，validation accuracy は，グレード 4 に対してポアソン合成のみを用いた E は，0.810 であった．これより，ポアソン合成のみで学習した場合より，DA 法を用いた D を下回る性能であることがいえる．

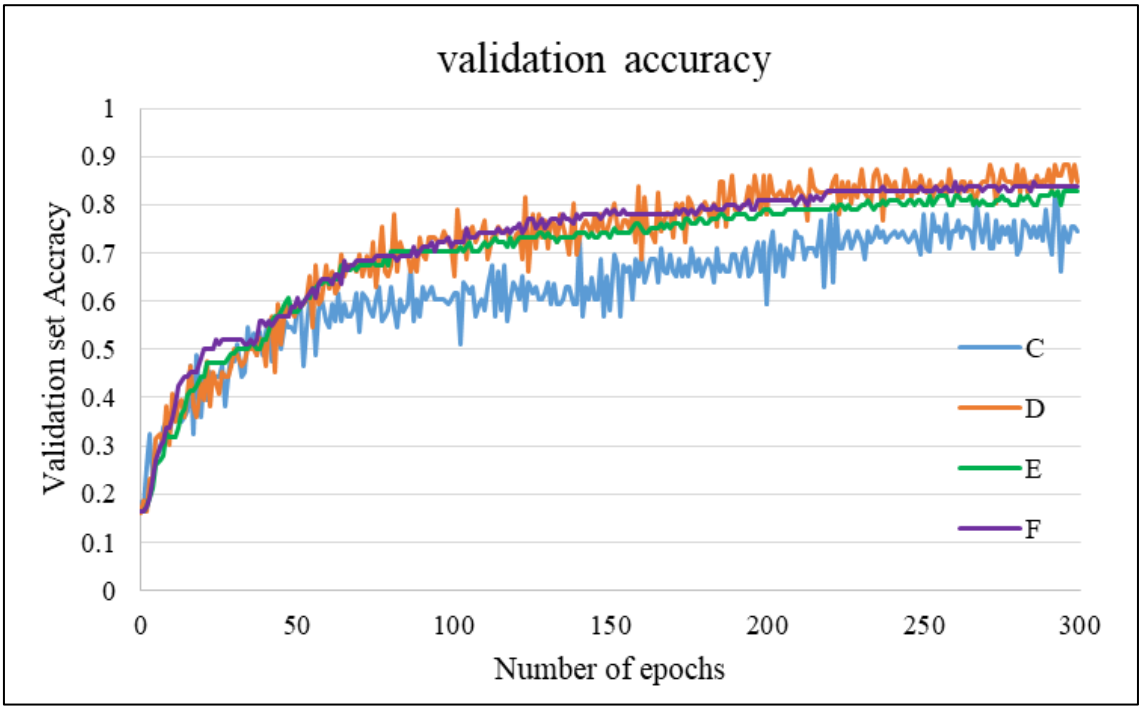
図 3-13 にグレード 4 に対するデータ生成比較の学習曲線を示す．ここで，F（ハイブリッド生成）は，滑らかな曲線であることがわかる．

epoch について，本章では事前学習で過学習を起こさない 300 とした．図 3-13（a）より，Accuracy の低下が発生していないことから学習不足になっていないこと，(b) より，Loss の上昇が発生していないことから，全てのデータセットにおいて過学習が発生していないことが推測される．課題 4 に対して，学習不足を補う目的で作成した人工症例画像は，学習データとして有効であることが推測できる．これらの結果より，D（DA 法）と E（ハイブリッド生成法）について，混同行列を作成し，精度評価値を算出した．

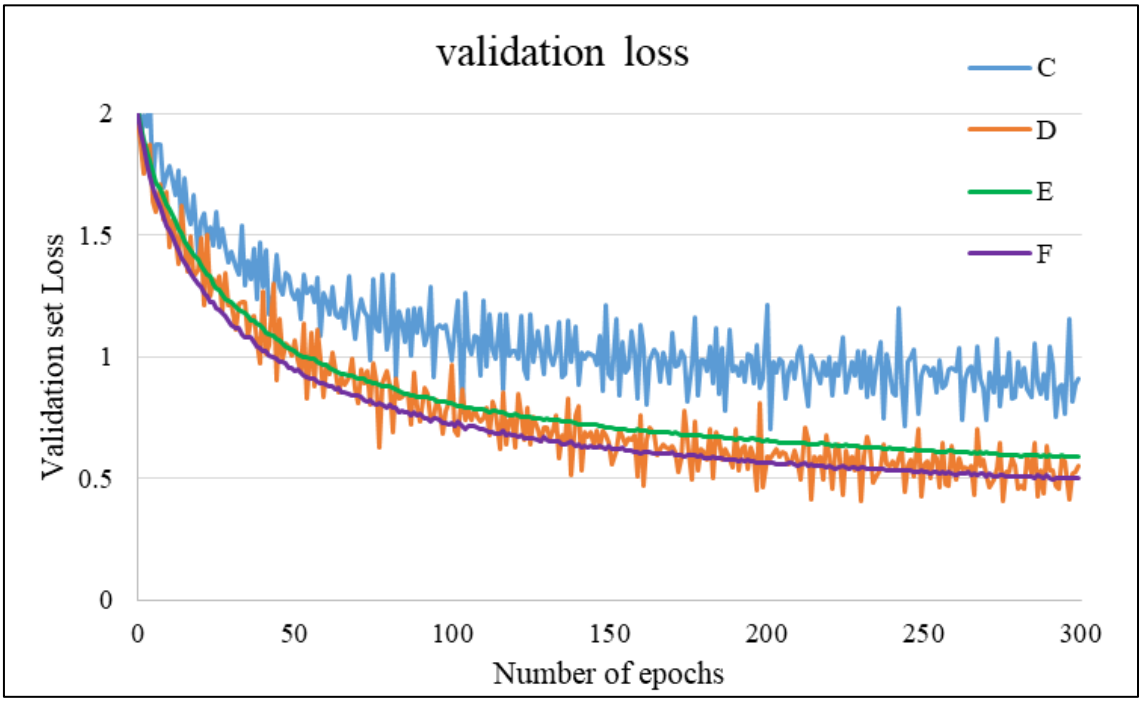
表 3-7　データ生成法の比較検証による性能評価結果

		validation		test	
		accuracy	loss	accuracy	loss
C	Gr4	0.862	0.419	0.744	0.913
D	Gr4_DA	0.942	0.199	0.834	0.527
E	Gr4_PID	0.942	0.192	0.810	0.600
F	Gr4_PID+DA	0.942	0.196	0.834	0.511

PID（Poisson Image Editing）



(a)



(b)

図 3-13　学習曲線（画像生成手法の比較）
(a) validation accuracy　(b) validation loss

DA 法と Hyb-RDGS の混同行列，評価指標の結果を図 3-14，表 3-8 に示す．4 クラスにおける正答率は DA 法が 83.4%であったのに対し Hyb-RDGS は正答率が 85.1%となり，両者には有意な差が見られた．各グレードの評価指標の結果について，DA 法ではグレード 1～3 の感度，適合率，F1 値が 80%を上回る結果を示したが，グレード 4 は感度 77.8%，適合率 76.2%，F1 値 76.9%であった．Hyb-RDGS はグレード 1 では DA 法に比して適合率が小さい結果となったが，有意な差はなかった．グレード 4 では，感度 93.3%，適合率 84.7%, F1 値 88.5%であり，DA 法と比較して，それぞれ 15.5%，8.5%，11.6%向上していた．

DA

		Testing grade						
		Gr 1	Gr 2	Gr 3	Gr 4	recall	precision	F1-measure
True grade	Gr 1	81	4	4	1	0.900	0.890	0.895
	Gr 2	4	72	6	8	0.800	0.819	0.809
	Gr 3	3	10	75	2	0.833	0.834	0.833
	Gr 4	3	2	5	35	0.778	0.762	0.769

Hyb-RDGS

		Testing grade						
		Gr 1	Gr 2	Gr 3	Gr 4	recall	precision	F1-measure
True grade	Gr 1	81	5	4	0	0.900	0.794	0.844
	Gr 2	10	70	4	6	0.778	0.897	0.833
	Gr 3	11	2	75	2	0.833	0.882	0.857
	Gr 4	0	1	2	42	0.933	0.847	0.885

図 3-14 DA 法とハイブリッド生成法の混同行列

表 3-8　DA 法とハイブリッド生成法の精度評価値

Classification Accuracy

		DA method	Hyb-RDGS
	Overall	0.834	0.851
(a)	**P value**	0.024*	

		sensitivity		**precision**		**F1-measure**	
		DA	Hyb	DA	Hyb	DA	Hyb
(b)	**Grade 1**	0.900	0.900	0.890	0.794	0.895	0.844
	P value	0.374		0.347		0.239	
	Grade 2	0.800	0.778	0.819	0.898	0.809	0.833
	P value	0.423		0.02*		0.139	
	Grade 3	0.833	0.833	0.834	0.882	0.833	0.856
	P value	1.000		0.083		0.465	
	Grade 4	0.778	0.933	0.762	0.847	0.769	0.885
	P value	0.025*		0.122		0.0001**	
	Average	0.828	0.861	0.826	0.853	0.827	0.855
	P value	0.475		0.46		0.383	

*; p< 0.05　**; p< 0.01

3.6.2　内部特徴の可視化

図 3-15 にヒートマップの一例を示す．図 3-15 中の上段に示した画像のように，放射線皮膚炎に内部特徴がハイライトされている画像について，Hyb-RDGS は正しくグレード判定していた．しかしながら，低グレードの放射線皮膚炎において，図 3-15 下段のように，Hyb-RDGS は耳や鼻などの放射線皮膚炎以外を内部特徴として捉える場合があり，このようなケースでは放射線皮膚炎をグレード 1 と判定する傾向があった．

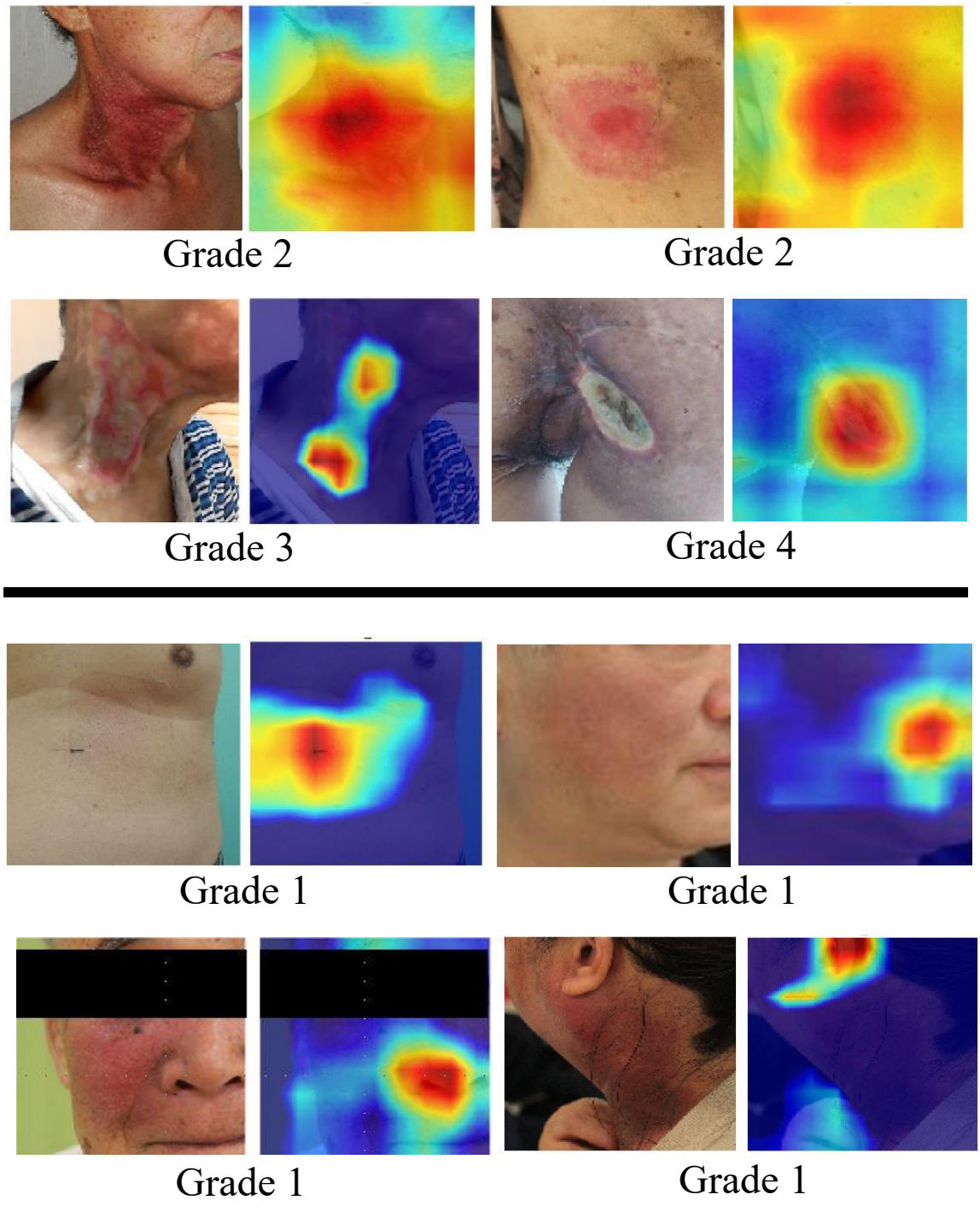


図 3-15 グレード判定の Grad-CAM 特徴画像

（左: オリジナル画像, 右: ヒートマップ画像）

(a), (b), (c) および (d): 放射線皮膚炎の検出例

(e), (f), (g) および (h): 放射線皮膚炎以外の検出例

a b
c d

e f
g h

3.7　考察

本章では, CTCAE に基づく放射線皮膚炎のグレード判定を行う Hyb-RDGS を作成し，その性能を評価した．学習画像数の水増しの効果について，図 3-11 より DA を行わない場合に対して DA 法では，3 クラスおよび 4 クラスの正答率が有意に向上したことから，DA が正答率の向上に寄与している事が示された．一方で，DA 法における 4 クラスの正答率は 3 クラスのそれより 3.2%低下した．これは, 表 3-8 に示すように DA 法のグレード 4 の正答率が他のグレードに比して低いことが原因である．DA による画像数の水増しは，データ数が増えたとしても同じ特徴点を持った学習画像が増えることになる．つまり，DA に用いたグレード 4 の画像数が少ない事が限られた特徴を持つデータによる学習となり，学習画像以外の画像をグレード判定する際には正答率は低下する．この点において, Hyb-RDGS ではグレード 4 の画像数が少ない事に対して，ポアソン生成を用いて特徴点が増えるような画像数の水増しを行ったため, 4 クラスの正答率においても DA 法の 3 クラスとほぼ同じ正答率になった．

Hyb-RDGS の各グレードの評価指標について，グレード 4 の評価指標はポアソン生成による学習画像数の水増しが特徴点の増加をもたらしたため，全ての評価指標が向上した．一方で，グレード 1 の評価指標は DA 法を下回るものもあった．これは，図 3-15 で示した低グレードの画像において特徴点の誤認識が多いことが原因である．本章で用いた検証データセットは，放射線皮膚炎の領域のみをトリミングして使用している．目や鼻，耳，乳首などの近くに発生した放射線皮膚炎画像のトリミングでは，それらを避けることは不可能であり，Hyb-RDGS は特徴点として認識する．しかしながら Hyb-RDGS を使用する場合，ヒートマップで内部特徴がハイライトされている領域を確認することができるので, 判定結果とヒートマップの両者を参照することで, 我々はグレードの誤認識を検知できる．

DCNN での学習は非常に多くの学習データが必要になるものの，放射線皮膚

炎に関しては高グレードの症例を集めることは困難である [14]．本システムで使用した症例画像は 647 枚であり，そのうちグレード 4 の症例画像は 5%程度である．この課題に対して，医用画像の分野では少ないデータ数に対するデータ生成手法の研究が行われている [43, 47]．本章では，データ生成手法として DA とポアソン合成を用いたハイブリッド生成法を用いた．この方法によって学習させた Hyb-RDGS は，グレード 4 の感度，適合率および F1 値は，DA によるデータ生成に対し，15.5%（93.3/77.8%），8.5%（84.7/76.2%），8.6%（88.5/76.9%）にそれぞれ向上し，他グレードのそれと同等レベルの性能に達した（表 3-8）．一般的に感度が高いシステムは適合率が低く，逆に適合率が高いシステムは感度が低い傾向にある．そのバランス指標である F1 値が DA 法より向上していたことから，ハイブリッド生成法によってデータ生成した Hyb-RDGS の有効性が示唆された．

分類器として，本研究の最も高い感度は，93.3%（グレード 4）であった．Fujisawa ら [34] の皮膚腫瘍分類の感度（4 クラス）は，最も高い感度は 95.7%（上皮性悪性細胞腫）であった．Hyb-RDGS のグレード判定は，全ての放射線皮膚炎の画像に対して万能というわけではない．評価に使用する画像は，カメラ等で撮影した写真であるため，どうしても撮影の仕方による画像のボケや歪といった画質に左右されてしまう．不鮮明な画像のグレード判定については，本研究では評価していない．

人間が行う評価では，評価経験が浅い評価者による学習不足，または体調不良などによる思考能力低下が評価に影響を与え，異なるグレード判定をする可能性がある．これに対して，Hyb-RDGS は，データベースを基に人間の評価で起きる知識や経験の差（個人差）に依存しない判定が可能である．これは，放射線皮膚炎の管理を行う上で，セカンドオピニオンやサードオピニオンのような判定の補助システムとなることが期待できる．

3.8　結言

本章では, 放射線皮膚炎のグレード判定システムである Hyb-RDGS を作成し, その有効性を検証した．課題 1（放射線皮膚炎グレード 1～4 における症例画像の収集と画像の品質）に対して，学習画像の精度向上を目的に放射線皮膚炎画像選定プロトコルを決定し，前処理を実施した．収集画像は，放射線皮膚炎の症例部位やグレード分類に偏りがあった．課題 2（不均衡なデータ数と少数画像の取り扱い）に対して，データ数の多いグレードにアンダーサンプリングを実施し，少数画像について DA 処理を用いてオーバーサンプリングを行い，不均衡の解消を図った．さらに課題 3（稀な症例の取り扱い（極少数画像の取り扱い））に対して，ポアソン合成処理により放射線皮膚炎の炎症部を正常皮膚画像に埋め込む人工症例画像を生成する手法を提案した．

DA とポアソン合成のハイブリッド生成法の精度評価を実施した結果，本システムは，DA とポアソン合成を混合して学習画像数を拡張することにより，DA のみを使用した場合に対してグレードの正答率が向上した．症例数が少ないグレード 4 においては，感度が 15.5%向上した．ハイブリッド生成法によって，放射線皮膚炎のグレードがより高精度に判定出来ることを確認でき，本システムが放射線皮膚炎のグレード判定の補助システムとなる可能性を示した．

第4章　EfficientNet を用いたベイズ推定に基づく放射線皮膚炎グレード判定手法の開発

4.1　緒言

DCNN を用いた放射線皮膚炎のグレード判定の作成は，前章で述べた幾つかの課題がある．これらの課題に対処しながら，第 3 章では，DCNN の不足する学習画像を補う手法として，元の学習画像に画像処理を加えてデータ量を増やす DA 処理を行うとともに，人工症例画像と混合したハイブリッド生成法による学習データを作成した．この成果より，Hyb-RDGS のグレード判定の正答率が 85.1%を達成したことを報告した [48]．しかし，Hyb-RDGS は，正しくグレード判定を出力するために，正解画像のみを用いて学習する必要があった．つまり，人間が判断する曖昧なグレード判定（グレード判定の相違）画像は，除外している（課題 4）．

医療分野における DCNN は，第 1 章で述べたように医用画像を画像認識で判断する技術を中心に研究開発が進み，実用化されている．これまでは，人間が曖昧な基準で判断し，大量のデータ処理に時間を要してきた．DCNN は，大量のビックデータより正解を導き，データを効率的に処理できる点に期待されている．高性能な GPU の発展もあり，学習時間（処理速度）の短縮も進んでいる．認識精度の高い DCNN は，アノテーションを繰り返すことで実用化できるシステムが構築される．すなわち，効率性の高い学習モデルが求められる．第 2 章で Hyb-RDGS は，一般的に用いられる DA 処理に加えて人工症例画像を用いる画像生成の手順を追加することで精度を向上させたが，効率的なモデルとは言い難い．そこで，本章では，Hyb-RDGS の性能を得た上で，課題 4 に対して放射線皮膚炎のグレード判システムの効率性を視野に入れた新たな学習モデルを検討する．

本章では，近年 Mingxing ら [20] によって提案された Efficientnet を用いて新たな判定手法を開発し，精度と学習効率の向上を図りながら，課題 4（放射線皮膚炎のグレード判定の相違）に対して，EfficientNet を用いたベイズ推定に基づく

判定手法を提案する[49]．なお，本研究では，グレード判定の相違があった画像は，複数（グレード）判定画像と称する．

4.2　提案手法の概要

4.2.1　提案手法のワークフロー

図 4-1 に提案手法のワークフローを示す．

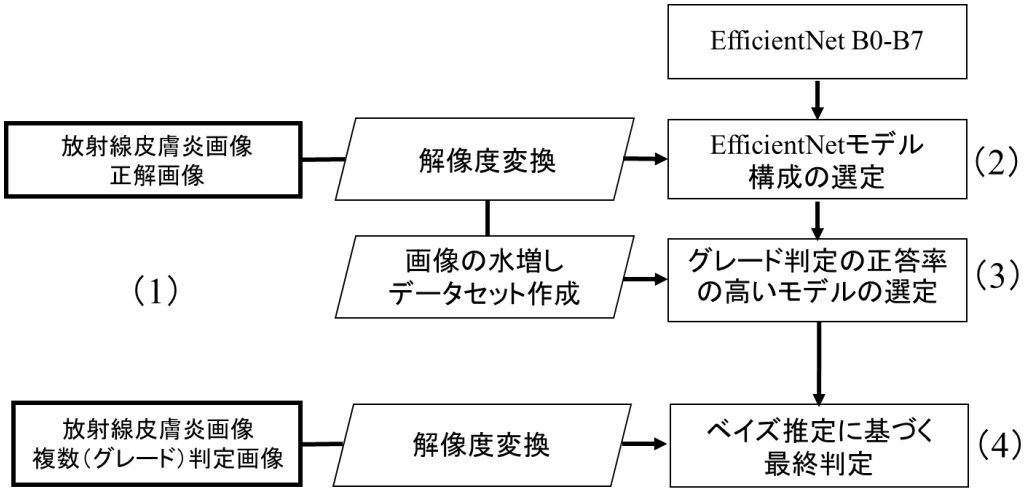


図 4-1．提案手法のワークフロー

(1) はじめに放射線皮膚炎画像を正解画像と複数（グレード）判定画像に分類する．正解画像は，放射線皮膚炎画像判定プロトコルによってグレード判定された画像である．

(2) EfficientNet モデルのスケーリング B0～B7 と解像度の条件を変えたモデル構成を評価し，グレード判定精度の高い EfficientNet モデルを複数選定する．

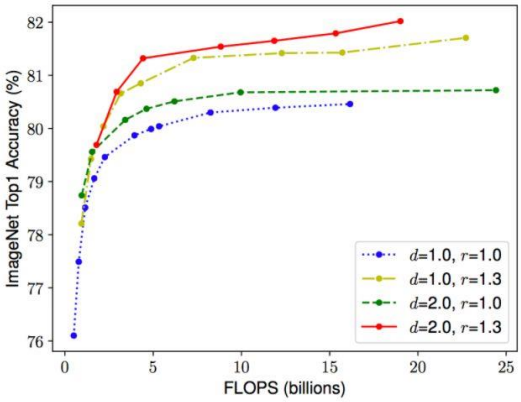
(3) データ拡張の条件を変えた複数のデータセットを作成し，(2) で選定したモデルに用いた．モデル構成を評価し，グレード判定精度の高い EfficientNet モデルを選定する．

(4) (3) で選定したモデルを使用してベイズ推定に基づく，複数（グレード）判定画像に対する最終グレード判定を行う．

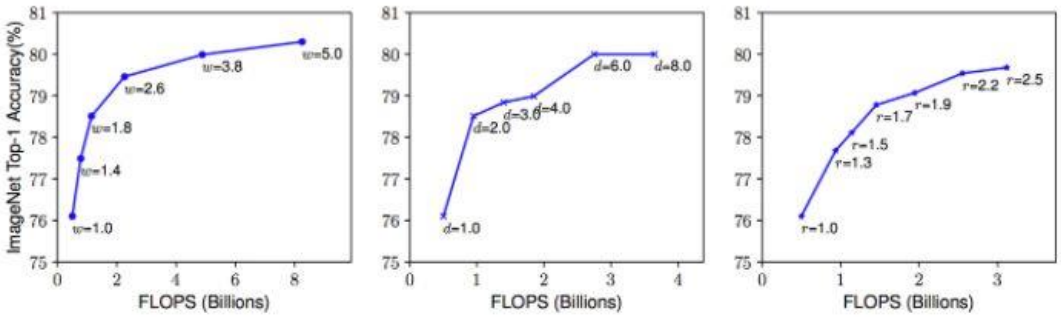
4.2.2　EfficientNet モデル概要

EfficientNet モデルは, 2019 年に Mingxing ら [20]によって提案されたモデルで, 特有のチューニング（複合スケーリング法）を利用することが特徴である. 近年, 高精度な画像認識を達成するため，学習の過程で構造自体を自動的に探索する Neural Architecture Search（NAS）や compound scaling method（複合スケーリング法）を用いてスケーリングを徐々に大きくし，効率的な構造を探索する EfficientNet が登場した．従来の CNN のスケーリングは，CNN の主なパラメータである深さや広さを任意に増大させるか，大きな解像度を使用することであった. 手動チューニングを必要とし, 最適な値とはいえない場合もあった. また, モデルのパラメータ増加に対して精度の向上に限界があった．これに対して EfficientNet は，図 4-2（a）に示すようにスケールアップによって精度はあがるものの広さ, 深さおよび解像度の 3 つのパラメータは, 関係しあっているため, バランスをとりながらスケーリングをした方がよいとされている，各パラメータだけではモデルが大きくなると図 4-2（b）に示すようにその恩恵が受けにくい点に注目し，ネットワークの構造を変えずに深さと広さと解像度の比率を固定してスケーリングアップしていく. そのため, 他のパラメータ数をあまり増やすことなく精度をあげることができる．パラメータと計算量が他の DCNN よりも小さいため速度も向上している． EfficientNet は，比率を固定してスケーリングしているので，そのスケールする値によって B0～B7 まである.

モデルスケーリングについて Sheeela ら [50]は，道路状況カメラ画像の分類に EfficientNet-B0, B4 を使用したと報告している. EfficientNet モデルが分類に有効であったとされているが，リソースの問題から他のモデルのスケーリングは課題とされている．EfficientNet モデルのスケーリングの違いについて研究された報告は十分でなく, 適切なスケーリングや入力画像のパラメータについて, 検討の余地があるといえる.



(a)



(b)

図 4-2 Tan, Mingxing, Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv preprint, 2019: arXiv:1905.11946.から抜粋

(a) 異なるスケーリングによる実験結果

(b) 左から，広さ（レイヤーのサイズ），深さ（レイヤーの数），解像度（入力画像の縦横の大きさ）

4.2.3 アンサンブル学習

本研究では，この相違が生じたグレード判定について EfficientNet を用いた判定モデルを提案し，複数（グレード）判定画像に対する評価方法に複数のモデルを用いて最終判定を行う手法を検討する．一種のアンサンブル学習といえる．本研究では具体的に複数の重み付けによりベイズ推定を適用して行うことから，アンサンブルの考え方に基づき，最終判定を行う手法と捉える．アンサンブル学習について，述べる．

アンサンブル学習は，複数の分類器の結果を統合することにより一つの分類器よりも予測精度の向上及び汎化性能に有効な手法として用いられてきた[18,19]．

CNN を用いた画像認識においてもアンサンブル学習は有効であることが分かっている [51]．一般的なアンサンブル学習では，複数のモデルの予測値の平均値，クラス分類では複数のモデルの予測結果の多数決が用いられる．アンサンブルの学習の形態は多くあるが，代表的な手法にバギング法とブースティング法がある．

バギング法は，学習データを分割するのではなく，図 4-3 に示すように，k組の学習データを作成し，それぞれ並列に学習させ，最後にそれらを統合する手法である．統合は，各分類関数値の平均とする方法と全分類器の多数決とする方法がある．通常は，後者の方が多用されており，並列に学習できる効率よく学習できる利点があるといえる．

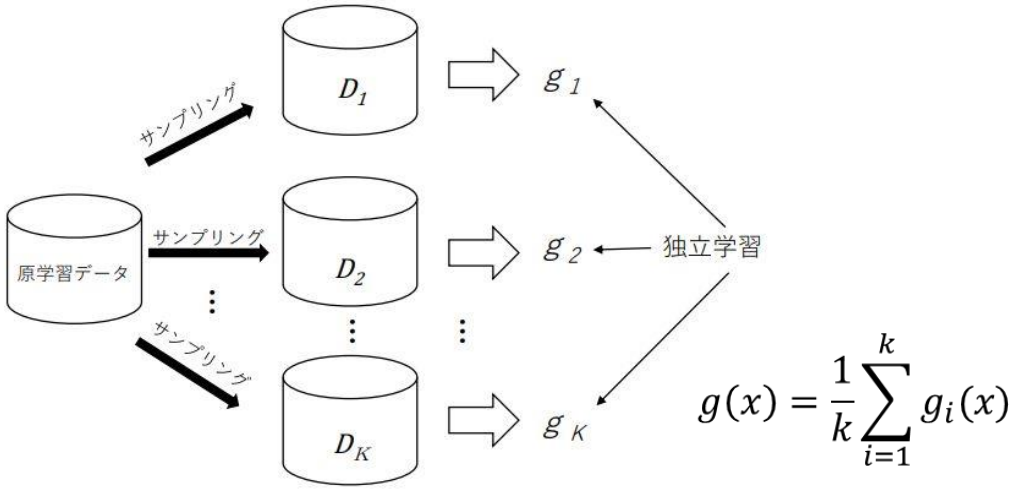
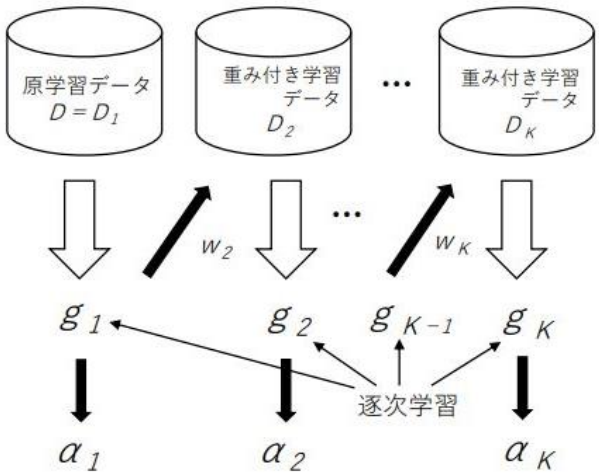


図 4-3 バギング法

ブースティング法は，逐次的に分類器を学習させ，識別性能を向上させる手法である．ブースティング法は，弱学習器を順番に学習して組み合わせていき，前の学習器が誤分類したデータを優先的に正しく分類できるように学習していく．バギング法に対して，並列に学習できないため学習に時間がかかるが，精度がよいとされている．



$$g(x) = \sum_{i=1}^{k} \alpha_i g_i(x)$$

図 4-4 ブースティング法

本研究では，効率さと同種のモデル（複合スケーリングの異なる複数の EfficientNet モデル）を使用するため，バギング法の考え方に基づいている．バギング法は，学習データの変動に対するクラス境界の変動を安定化させる効果があることが理論的に明らかにされており，CNN のような自由度の高い分類器のアンサンブルには効果的であるとされている [52)]．

さらに，本システムの最終グレード判定には，それぞれの独立したモデルの精度に依存する最小誤差を得ることを目的とするため，多数決は採用しない．評価者の判定と誤差の小さいモデルの出力とするベイズ定理に基づく最大事後確率推定法（ベイズ推定）を適用して予測する手法を提案する [53)]．

4.2.4 ベイズ推定

ベイズ推定は，ベイズの定理と組み合わせて確率的推論を行う統計的手法である．ベイズの定理は，尤度関数が理論的に設定できない場合にも適用でき，実用性が高いため臨床医学の診断推論に活かせる定理として用いられている [54, 55)]．図 4-5 に示すようにベイズの定理は，

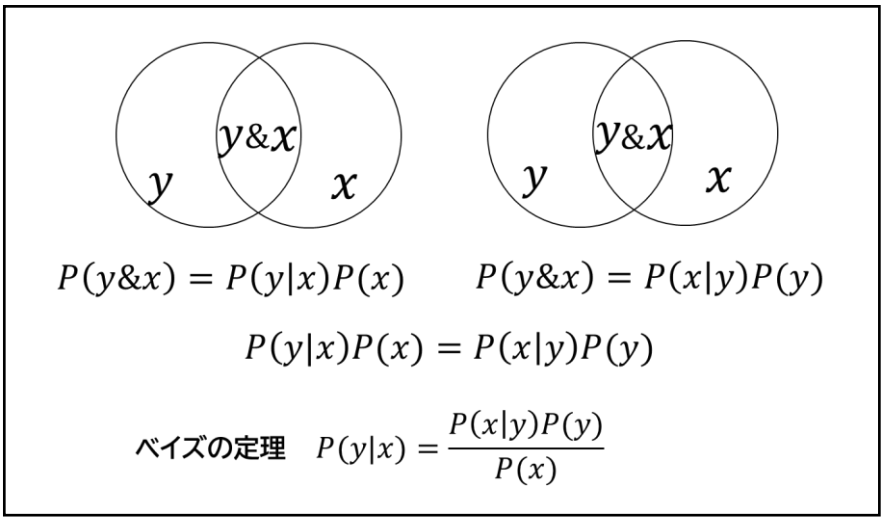


図 4-5 ベイズの定理

$P(y)$ ：yという事象が起こる確率

$P(x)$ ：xという事象が起こる確率

$P(y\&x)$ ：yが起きかつx*が起きる確率*

$P(x|y)$ ：yが起きたという条件のもとxが起きる確率．この条件付き確率は次のように表せる

$$P(x\,|\,y) = P(y\&x)/P(y) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (4.1)$$

条件付き確率の（4.1）の式の両辺に$P(y)$を掛けると，次の乗法定理が得られる．

$$P(y\&x) = P(x|y)/P(y) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (4.2)$$

ベイズの定理は，乗法定理（4.2）式を変形しただけの定理である．（4.2）式から

$P(y\&x) = P(x|y)/P(y)$

y の役割をxに担わせれば

$P(y\&x) = P(y|x)/P(x)$

これらの式は，左辺の$P(y\&x)$が同じなので

$P(x|y)P(x) = P(y|x)P(y)$

$P(y|x)$ について解くと，次の式が得られる．これがベイズの定理である．

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad \cdots\cdots\cdots \quad (4.3)$$

4.3　構築環境

EfficientNet は，ハードウェアは CPU Core i7 7700k (Intel Corporation. Santa Clara,CA USA)，メモリ 32GB, GPU GeForce GTX1060 6GB (NVIDIA Corporation, Santa Clara , CA, USA), オペレーションシステムが Windows10 Home64 ビット版であるパーソナルコンピュータに Anaconda Navigator をインストールして PyTorch の仮想環境を構築し，モデルをインストールして実施した.

4.4　データセット作成

4.4.1　実験データと前処理

本章では，第 3 章 3.4.1 項に示した放射線皮膚炎画像を用いて，データセット作成を実施する．さらに，課題 4 に対して複数（グレード）判定画像を最終グレード判定のテスト画像に用いる（第 2 章，放射線皮膚炎選定画像プロトコルでは，除外されている）.

課題 1，2 に対する対処は，アンダーサンプリングや DA 処理を用いたデータ拡張を実施する．課題 3（グレードの高い（重度）症例の学習）に対して，第 3 章で実施したハイブリッド生成，新たなデータ拡張である Rand Augmentation 法[21] によるデータ拡張を実施する.

表 4-1 EfficientNet モデル構築に使用するデータ数

	グレード1	グレード2	グレード3	グレード4	計
使用データ数（枚）	318	95	205	31	649

表 4.2 最終グレード判定のテストデータ数

	グレード1または，2	グレード2または，3	計
複数（グレード）判定画像データ数（枚）*	30	30	60

※テスト画像に用いる

4.4.2　Rand Augmentation (RA) によるデータ拡張

表 2-1 に示したように，RA は transformation の数 K=14 種類のデータ拡張操作からランダムに N 個サンプルし，それぞれを強さ M で順番に適用することにより，最適なデータ拡張を行う [21]．K＝14 はあまり変えることなく固定値とされている．この N と M の 2 つのパラメータを調整するだけで，チューニングが可能であり，効率よくデータ拡張が行える手法といえる．さらに RA は，本システムで使用する EfficientNet-B7＋RA において，ImageNet の Top1 精度を塗り替えた実績があることからも本研究では，RA を新たにデータセットとして用いることとしている．図 4-6 に RA の例を示す．Ekin ら [21] の実験では，図 4-7（c）に示すように学習データが大きいほど M が大きくなり，(d) では学習データサイズが大きければ大きいほど，最適な M は大きくなる傾向が示されている．つまり，小さい学習データに対して，強めのデータ拡張が必要とはいえないことが示されている．本章で用いるデータ数は，わずか 649 枚と少ないため，極端なデータ拡張が必ずしも最適とはいえないことになる．

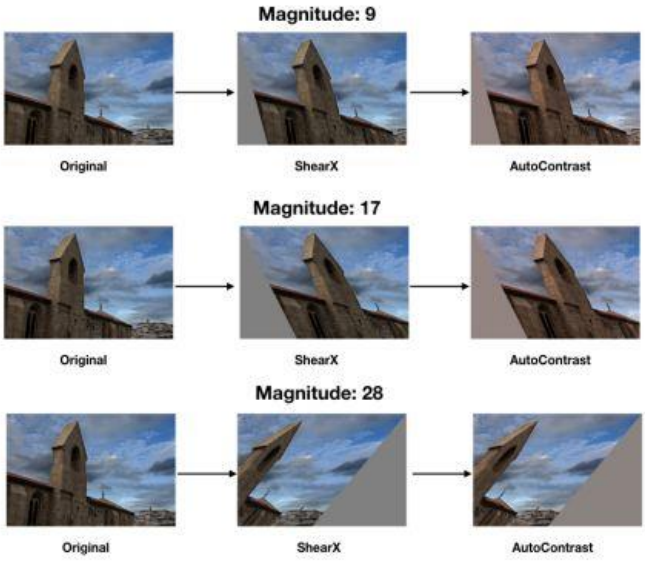


図 4-6 Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. arXiv: 1909. 13719v2 [cs.CV], 2019.から抜粋

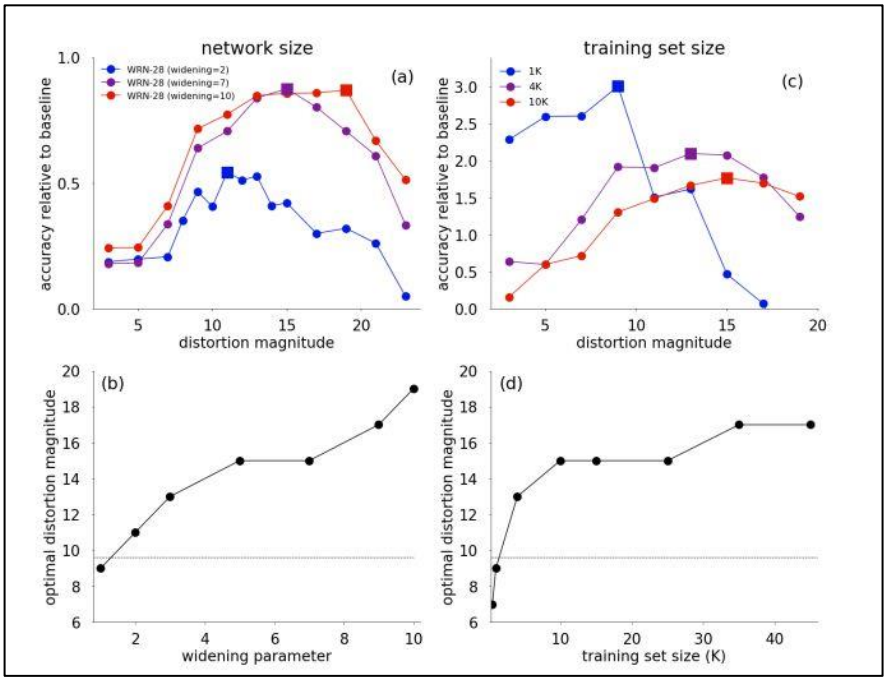


図 4-7 Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. arXiv: 1909. 13719v2 [cs.CV], 2019.から抜粋

（a）3 つの widening パラメータの比較　（b）モデル容量の比較

（c）学習データサイズの比較　（d）学習データと magnitude の比較

4.4.3 データセット作成

本章では，図 4-1 に示したように，はじめに EfficientNet モデルの構成を検証する．モデル構成検証用には，次項で述べる解像度を変換したモデル構成用データセットを作成し，256×256，384×384，512×512 画素の 3 パターンを作成する．モデル構成のパターンが多くなることから，表 4-3 に示す少数データを用いて検証する．

解像度のパターンを変えた選定したモデルについて，それぞれのモデルのデータセットにグレード判定モデル検証用データセットを作成する．ここで，課題 3（稀な症例の取り扱い（極少数画像の取り扱い）に対して，グレード 4 におけるデータ生成を行い，表 3-4 に示すように A～D の 4 つのデータセット作成する．

表 4-2 モデル構成検証用データセット

データ生成手法	学習データ	テストデータ
Rand Augmentation	377	110

表 4-3 グレード判定モデル検証用データセット

データセット No.	データ拡張法	学習データ（枚）	テストデータ（枚）
A	Rand Augmentation	649	110
B	Data Augmentation	649	110
C	Rand Augmentation (ポアソン合成*)	649 (195)	110
D	Rand Augmentation (Rand Augmentation＋ポアソン合成)	649 (183+195)	110

（）は，グレード4に対する生成手法

4.5 EfficientNet モデルの構成比較

EfficientNet モデルは，特有のチューニング（複合スケーリング法）を利用するモデルであり，ネットワークの構造を変えずに深さと広さと解像度の比率を固定してスケールアップしていく．比率を固定してスケーリングしているので，そのスケールする値（スケーリング係数）によって B0～B7 まである [20]．本章では，スケーリング係数と学習データの解像度の関係を最適化するため，解像度変換を行い，それぞれのスケーリング係数を変えたモデルに対して，解像度を変えたデータセットを用いて精度評価（test accuracy）を用いて実施する．

本章では，第 3 章で作成した Hyb-RDGS の性能を得た上で，実用化に必要な効率性の高いモデル作成をもう一つの目的としている．そこで，近年，普及している小型端末のモデルにも乗せられる高性能 CNN としてと，幾つか報告されて

いるデータサイズに注目した[56-58]．Srinivasu ら[59]は，7 つの皮膚疾患分類に MobileNet V2 と Long Short Term Memory（LSTM）をベースとした moblie-size の DCNN を開発したと報告した. Srinivasu らの提案モデルの目標解像度は 224×224 画素とされている．本研究では，Mobile-size の DCNN で報告されている解像度を参考に 256×256，384×384，512×512 画素の 3 パターンの解像度を設定する．はじめに解像度の条件を変えた学習用データを複数作成し，放射線皮膚炎画像のグレード判定を行う学習モデルの構成を比較する. 表 3-5 に検証するモデル構成を示す．また，それぞれのモデルの学習時間についても比較を行う．

表 4-4 EfficientNet モデルの構成パターン（24 通り）

スケーリング係数	B0	B1	B2	B3	B4	B5	B6	B7
学習データの解像度	256×256	256×256	256×256	256×256	256×256	256×256	256×256	256×256
	384×384	384×384	384×384	384×384	384×384	384×384	384×384	384×384
	512×512	512×512	512×512	512×512	512×512	512×512	512×512	512×512

4.6　EfficientNet モデルを用いたグレード判定モデルの作成と性能評価

第 3 章では，元の画像に変換を加えて DA を行うとともに，新たな症例を人工的に生成する人工症例画像も症例画像として利用した Hyb-RDGS を作成している．しかし，DA は画像データにどのような操作が有効なのか試行に時間を要してしまい，人工症例画像を生成する作業を加えなければならない．そこで，本章では，Ekin らによって自動的に DA を選択してくれる手法として提案された RA を用いたデータセットを新たに作成した．ここでは，RA，DA および人工症例画像で作成したデータセットを用いて最適なデータセット検証を行う．それぞれのモデルについて 3 分割交差検証を用いて，学習曲線，グレード判定の正答率および混同行列から性能を評価する．それぞれのモデルの評価値を比較し，アンサンブル学習を行うモデルを探索する．混同行列を用いた評価値の算出は，第 3 章でも示した（3.2）～（3.5）式を用いる．

- 正答率（*Accuracy*）

$$overall\ accuracy = \frac{\sum_{i=1}^{n}(TruePositive_i + TrueNegative_i)}{\sum_{i=1}^{n}(TruePositive_i + FalseNegative_i + FalsePositive_i + TrueNegative_i)} \quad \cdots\cdots\cdots\cdots (3.2)$$

- 感度（*Sensitivity*）

$$sensitivity_i = \frac{TruePositive_i}{TruePositive_i + FalseNegative_i} \quad \cdots\cdots\cdots\cdots (3.3)$$

- 適合率（*Precision*）

$$precision_i = \frac{TruePositive_i}{TruePositive_i + FalsePositive} \quad \cdots\cdots\cdots\cdots (3.4)$$

- F1 値（*F1-measure*）

$$F1\ value_i = 2\left(\frac{sensitivity_i \times precision_i}{sensitivity_i + precision_i}\right) \quad \cdots\cdots\cdots\cdots (3.5)$$

4.7　ベイズ推定に基づく最終グレード判定

本章では，課題 4 に対して EfficientNet モデルを用いてベイズ推定に基づく最終グレード判定を行うことを目的としている．これまで述べた最適な重み付けされたモデル構成，データセット検証および，ベイズ推定用いた最終判定までの過程を図 4.8 に示す.

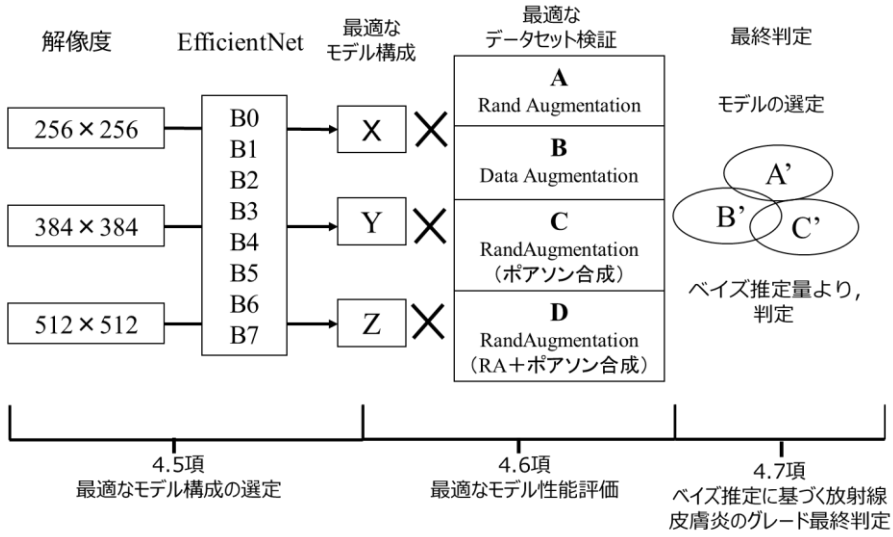


図 4-8 最終判定を行うモデル選定の概要

4.7.1　最適な EfficientNet モデル検証

最終的なグレード判定に用いる EfficientNet モデルは，表 4-3 で示した 4 つのデータ生成手法に対して，表 4-4 より解像度と重み付け変えたモデル検証を 3 分割交差検証法により実施する．これより，最も精度の高い重み付けの異なるモデルを選定し，次項に述べるベイズ推定を適用して最終判定を行う．モデル選定は，正答率と性能評価値（*sensitivity, Precision, F1-measure*）を用いる．

4.7.2　ベイズ推定による放射線皮膚炎のグレード最終判定

放射線皮膚炎画像を判定するモデルでは，画像がある感度/特異度（尤度比）をもつグレードのパターン(x)，各グレードのクラスの正答率をクラス(y)とする．$p(y|x)$はグレードのパターン(x)を推定した後で，クラス(y)の判定確率，$p(y)$はパターン(x)と推定する前のクラス(y)の判定確率を表す．それぞれ事後確率，事前確率とよぶ．

ここで事前確率と尤度比が分かれば事後確率が求められる．図 4.9 に事後確率の決定を示す．

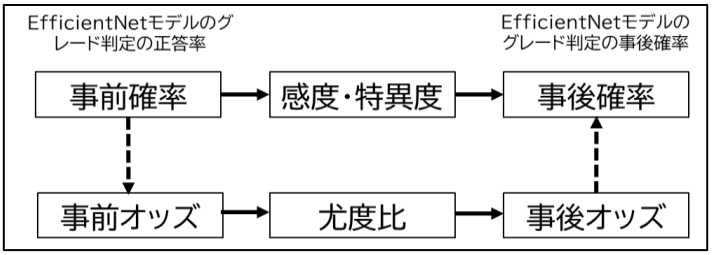


図 4-9 事後確率の決定

事後確率$p(y|x)$が最大となるクラスyにパターンxを分類すれば，認識誤差が最適なパターンになる．ベイズの定理（4.3）式によると

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y) \quad \cdots\cdots\cdots\cdots \quad (4.4)$$

ここで，$p(y|x)$の(y)に関する最大化を$p(x|y)p(y)$の最大化に変換できる．ニュ

ーラルネットワークでは，事後最大化解は，交差エントロピーの損失最小化（ベイズ推定量）で解くことができる．アンサンブル学習の各モデルの事後確率を用いてベイズ推定量を算出し，複数（グレード）判定画像に対する最終グレードを判定する．複数（グレード）判定画像に対するベイズ推定量は，2 つの確率分布 p, q に対して（4.5）式で与えられる．

$$H(p,q) = -\sum_{x} p(x) \log q(x) \qquad \cdots\cdots\cdots\cdots\cdots \quad (4.5)$$

本研究のグレード判定について，p を「評価者の分布」，q を「予測の分布」とすると，EfficientNet モデルによる予測が正解に近いほど p と q の誤差が小さくなると考えることができる．図 4-10 にベイズ推定量による最終判定の導出過程を示す．ここでは，3 つのモデルが，それぞれ正解であると判定したグレードを出力する．ベイズの定理より事後確率が求められ，グレード判定の予測の分布 q となる．また，評価者のグレード判定が 1 または，2 であった複数（グレード）判定画像を評価者の確率分布 p とする．これより，(4.5) 式を用いてベイズ推定量を算出する．各モデルの真のグレード判定におけるベイズ推定量が最も小さいモデルの判定を最終グレード判定とする．

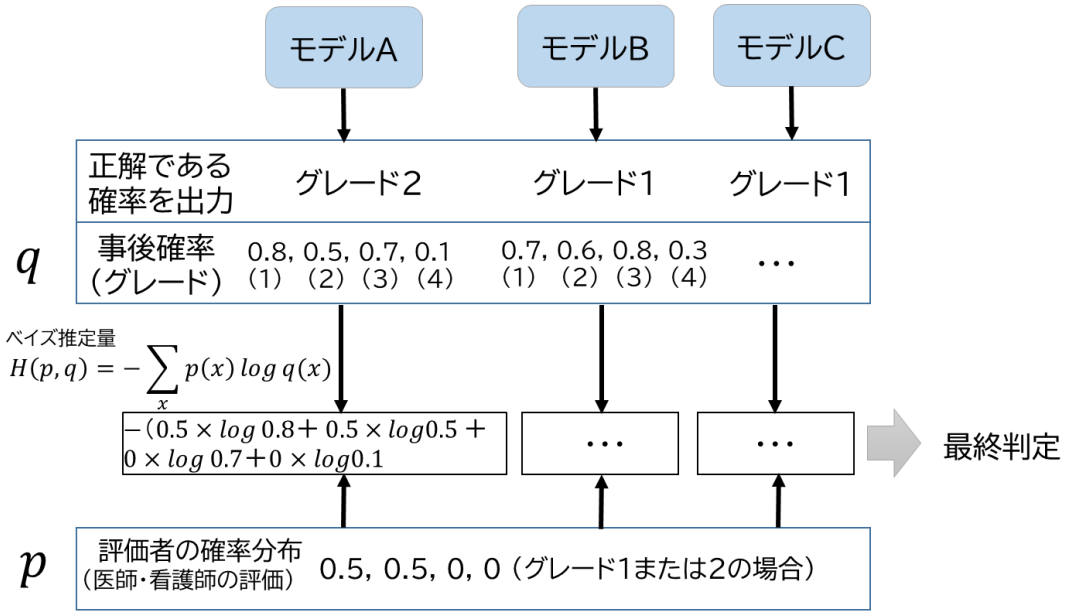


図 4-10 最終判定の導出

4.8　検証の結果

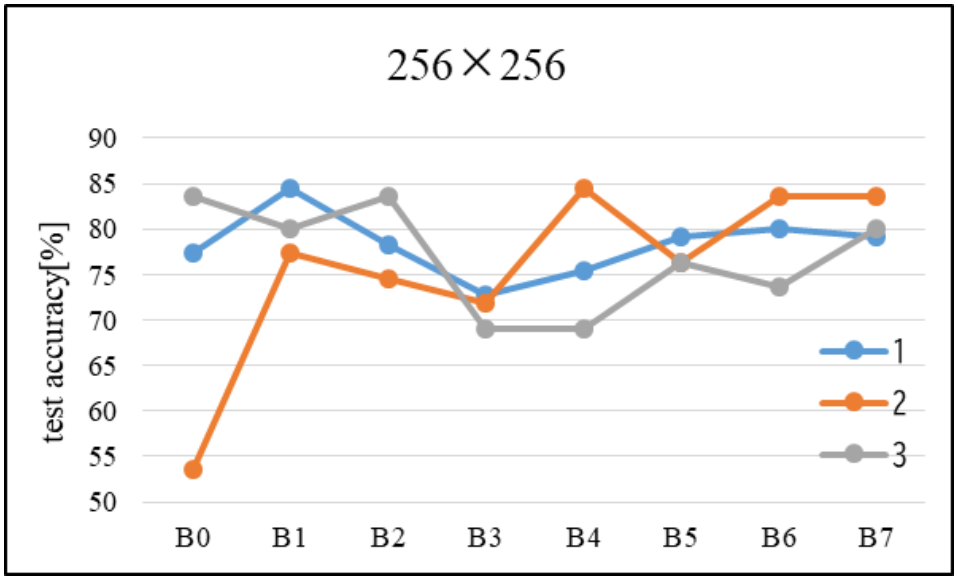
4.8.1　EfficientNet モデルの構成比較

EfficientNet-B0～B7 とそれぞれのデータ解像度の構成の test accuracy を図 4-11，図 4-12 にそれぞれの解像度における構成の test accuracy 比較した．表 3.5 には，各モデルの test accuracy の平均値を示す.

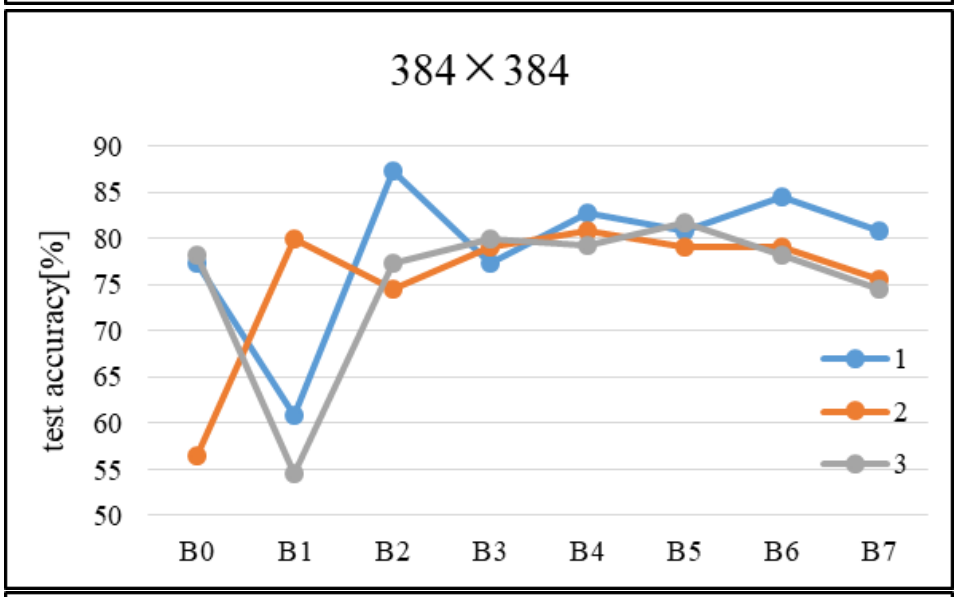
モデルに使用するデータ解像度は, 図 4-11 より, スケーリングが小さい B0~B2, 解像度 256，384 画素では，約 20～30%の test accuracy のばらつきがみられる．B3 以上では，384，512 画素では，安定している傾向がみられる．各解像度の平均値で示した図 4-12 より，スケーリングが小さい B0～B3，および最も大きい B7 では, 解像度によって test accuracy のばらつきが目立つ. また, B4～B7 では, 256×256，512×512 画素において test accuracy の向上がみられた．表 3-6 より，最も test accuracy が高いスケーリングは，B7（80.7±3.6）であった．これに対して，B0，B1（72.7±2.9，72.7±7.7）が最も低い結果を示した.

図 4-15 にスケーリング係数が B7 の解像度を変えた場合の学習曲線を示す. また，本検証におけるそれぞれの EfficientNet モデルの学習時間の計測結果を図 4-16 に示す．スケーリング係数が大きいモデルになるほど，学習時間が長くなっていることがわかる.

(a)



(b)



(c)

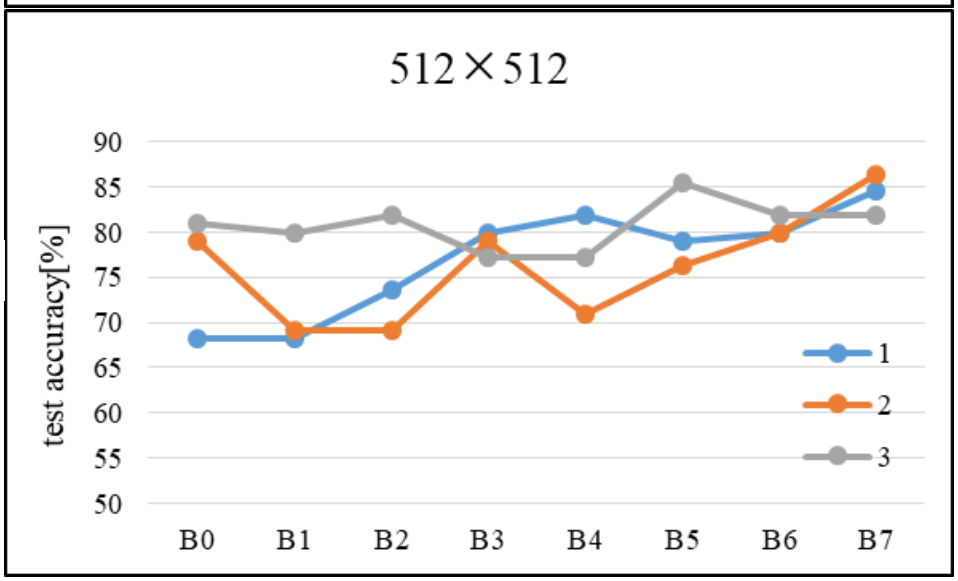


図 4-11 各モデルのデータ解像度における精度
（a）256×256 画素，（b）384×384 画素，（c）512×512 画素

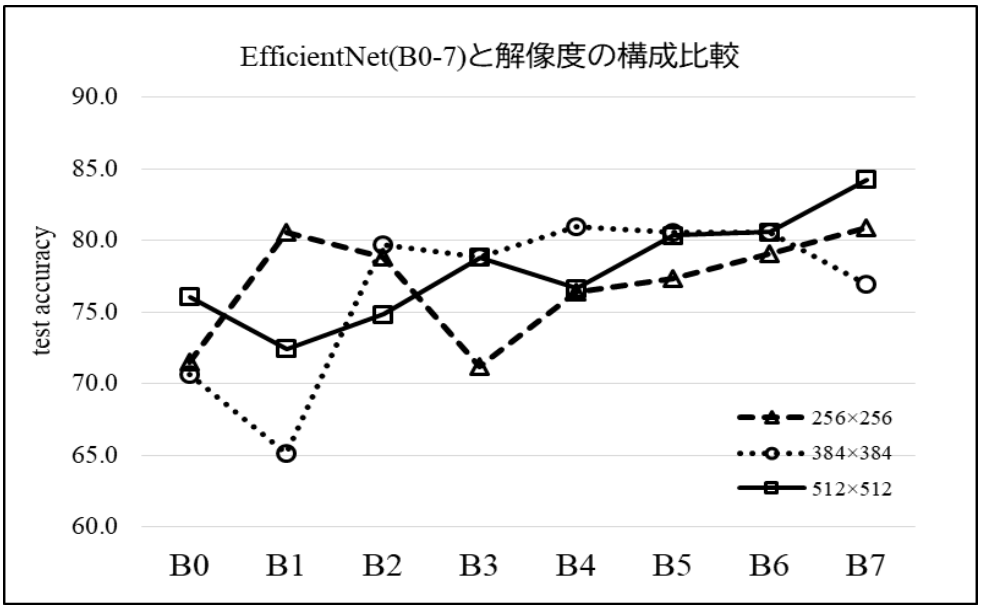


図 4-12 各モデルのデータ解像度における test accuracy の平均値

表 3-6 各モデルの test accuracy の平均値

	B0	B1	B2	B3	B4	B5	B6	B7
平均	72.7	72.7	77.8	76.3	78.0	79.4	80.1	80.7
標準偏差	2.9	7.7	2.6	4.4	2.6	1.8	0.9	3.6

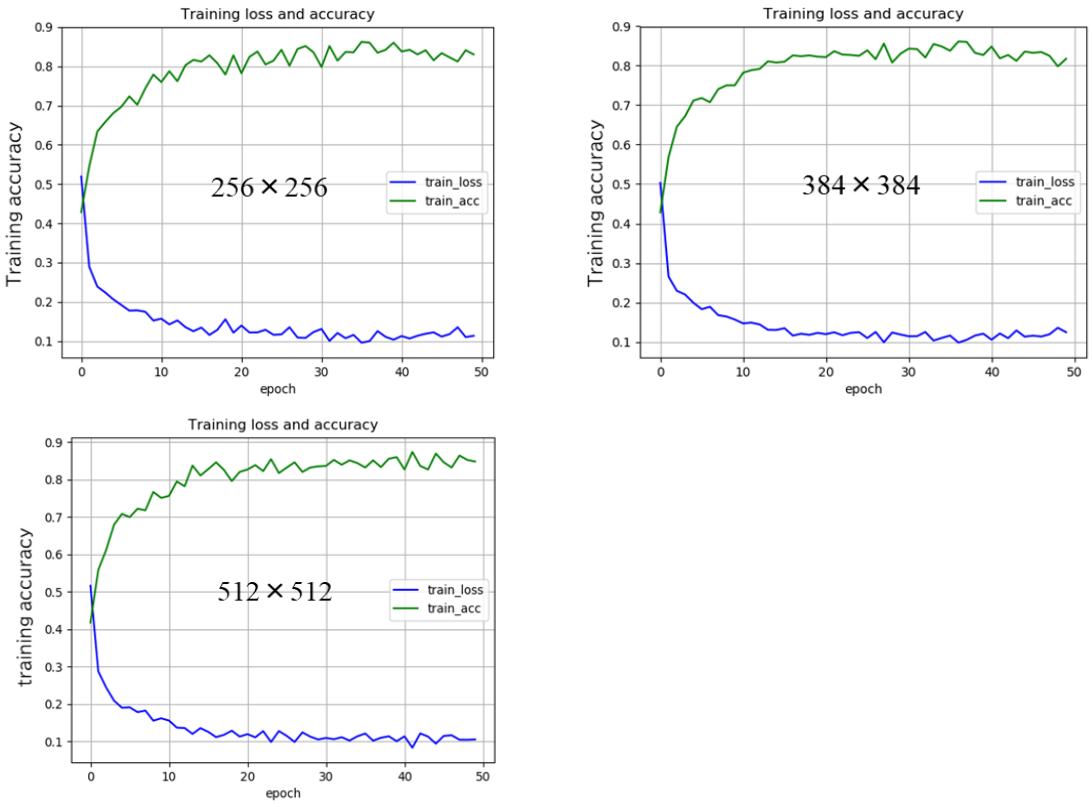


図 4-13 EfficientNet-B7 の学習曲線

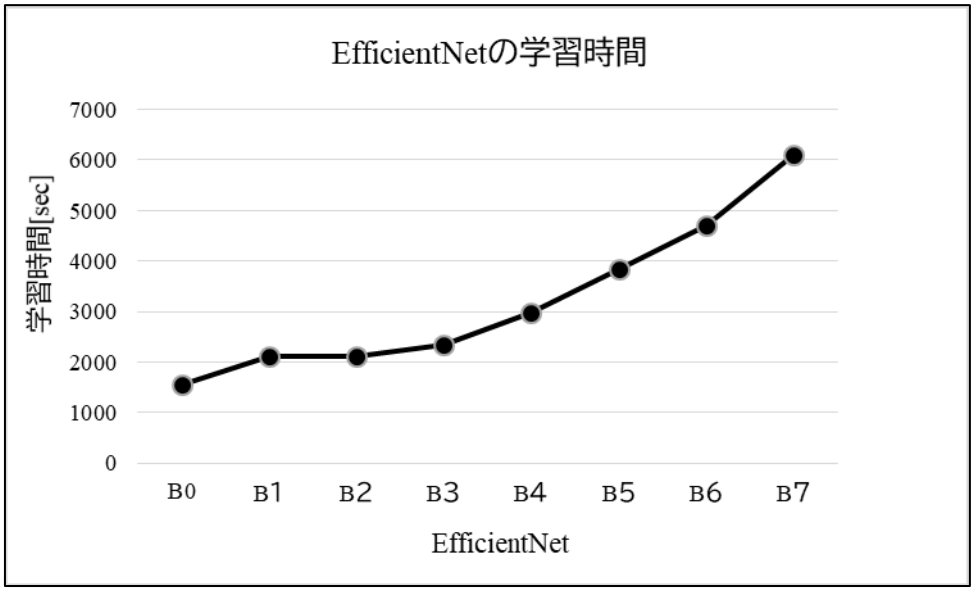


図 4-14 EfficientNet の学習時間計測結果（平均値）

表 4-5 に本検証の結果をまとめる．EffcientNet モデルと解像度の構成について，3 分割交差検証法により比較した．これより，test accuracy の高い 3 つのモデル構成（test accuracy の平均）を選択した．

選定したモデル構成

- EfficientNet-B1，256×256 画素（80.6%）
- EfficientNet-B4，384×384 画素（80.9%）
- EfficientNet-B7，512×512 画素（84.2%）

表 4-5 EfficientNet モデル構成比較表（test accuracy の平均値）

	B0	B1	B2	B3	B4	B5	B6	B7
256×256	71.5	80.6	78.8	71.2	76.4	77.3	79.1	80.9
384×384	70.6	65.1	79.7	78.8	80.9	80.6	80.6	77.0
512×512	76.1	72.4	74.8	78.8	76.7	80.3	80.6	84.2

4.8.2　データセット作成（Rand Augmentation によるデータ拡張）

本章では, 表 4-4 に示したように新たなデータ拡張法として Rand Augmentation 法を用いた．ここで，RA のパラメータは, K=13，N=6，M＝10 とした．図 4-15 に RA でデータ拡張を実施した結果の一例，図 4-16 に Python code を示す.

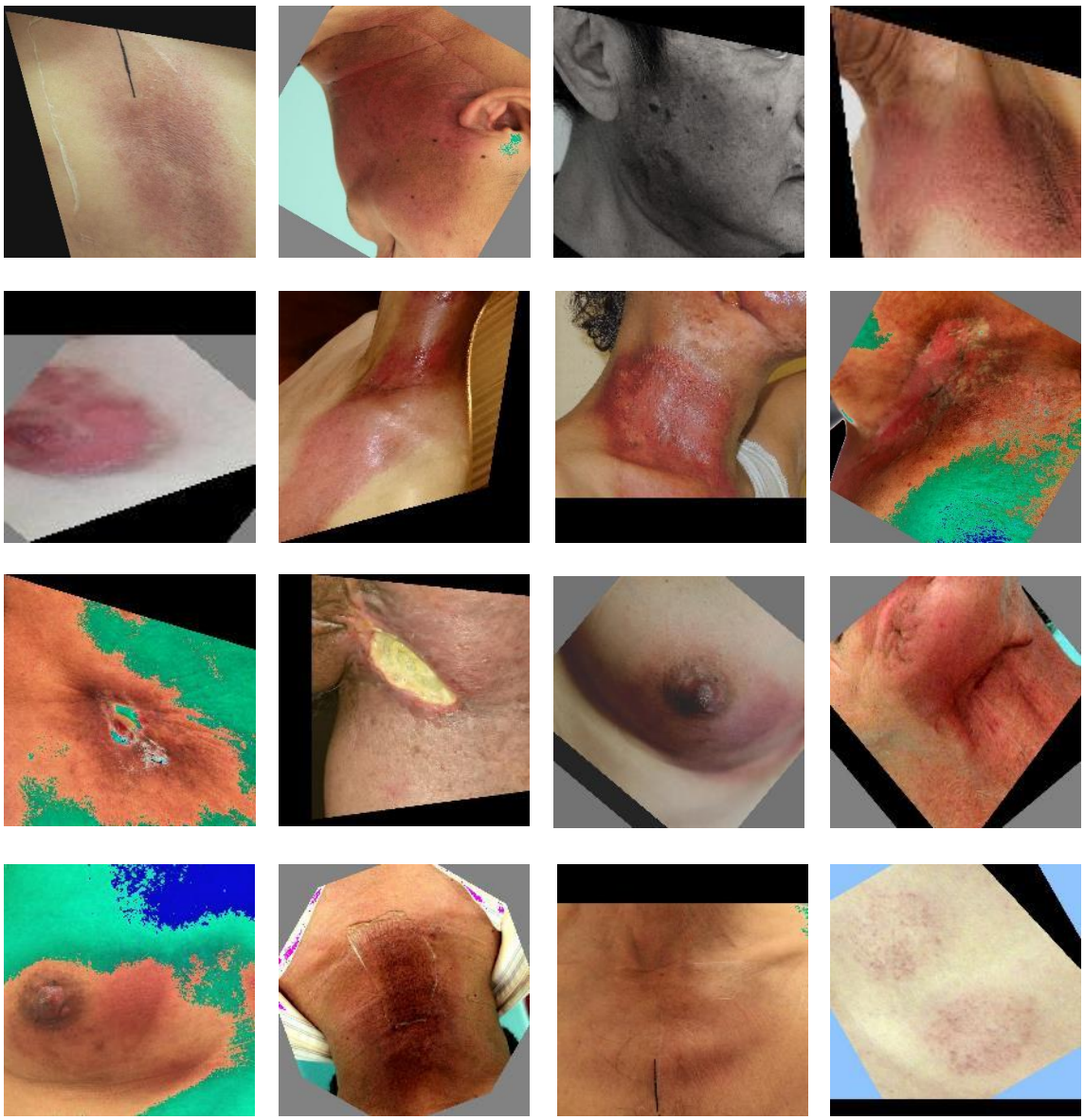


図 4-15 Rand Augmentation の実施結果（K=13，N=6，M＝10）

```
class Rand_Augment():
  def __init__(self, Numbres=None, max_Magnitude=None):
        self.transforms = ["autocontrast", "equalize", "rotate", "solarize",
                           "color", "posterize", "contrast", "brightness",
                           "sharpness", "shearX", "shearY", "translateX",
                           "translateY"]
  if Numbres is None:
            self.Numbres = len(self.transforms) // 2
        else:
            self.Numbres = Numbres
        if max_Magnitude is None:
            self.max_Magnitude = 10
        else:
            self.max_Magnitude = max_Magnitude
        fillcolor = 128
        self.ranges = {
            # Magnitude range
            # np.linspace(最初の値，最後の値，要素数)
            # np.round : 四捨五入を行う
            "shearX":np.linspace(0, 0.3, 10),
            "shearY": np.linspace(0, 0.3, 10),
            "translateX": np.linspace(0, 0.2, 10),
            "translateY": np.linspace(0, 0.2, 10),
            "rotate": np.linspace(0, 360, 10),
            "color": np.linspace(0.0, 0.9, 10),
            "posterize": np.round(np.linspace(8, 4, 10), 0).astype(np.int),
            "solarize": np.linspace(256, 231, 10),
            "contrast": np.linspace(0.0, 0.5, 10),
            "sharpness": np.linspace(0.0, 0.9, 10),
            "brightness": np.linspace(0.0, 0.3, 10),
            "autocontrast": [0] * 10,
            "equalize": [0] * 10,
            "invert": [0] * 10
        }
 def rand_augment(self):
        M = np.random.randint(0, self.max_Magnitude, self.Numbres)
        sampled_ops = np.random.choice(self.transforms, self.Numbres)
        return [(op, Magnitude) for (op, Magnitude) in zip(sampled_ops, M)]
```

図 4-16　Rand Augmentation の Python code

4.8.3　EfficientNet モデルの性能評価（最適なデータセット検証）

表 4-3 に示した A～D のデータセットを使用し，3 分割交差検証法によりそれぞれの EfficientNet モデルを用いたグレード判定モデルを作成した．前項で選定した EffcientNet モデルとデータセットについて，3 分割交差検証法により比較した. 本章で作成した EffcientNet モデルによるグレード判定システムの精度は，86.4%（384 画素の DA 法を用いたデータセット，EfficientNet-B4 モデル）であった．各モデルの全体の正答率を表 4.6，性能評価結果を表 4-7 にまとめる．これより正答率の高い（表 4-6），*F1-measure* 値の高い（表 4-7），3 つの EfficientNet モデルをアンサンブルに用いるモデルとして選択した．ここで，選択されたデータセット（A,B）は，RA，DA によるデータ拡張であった．データセット C，D（ポアソン合成による人工症例画像，RA と人工症例画像のハイブリッド生成）は，解像度が低いモデルの正答率が低下している．また，データセット検証結果について，表 4-8～表 4-10 に各モデルの混同行列，図 4-17～図 4-19 に学習曲線を示す.

選定したデータセットを用いた EfficientNet モデル

- EfficientNet-B1，256×256 画素（データセット A）
- EfficientNet-B4，384×384 画素（データセット B）
- EfficientNet-B7，512×512 画素（データセット A）

表 4-6　各データセットのグレード判定モデルの正答率

	B1 256×256	B4 384×384	B7 512×512
A	77.3±2.7%	75.2±6.8%	82.1±1.9%
B	77.0±7.6%	86.4±4.2%	80.9±3.1%
C	63.6±6.9%	59.4±2.1%	75.5±5.1%
D	64.5±7.2%	68.2±6.9%	76.4±9.2%

表 4-7 各モデル性能評価結果（データセットの比較）

Datasets		EfficientNet-B1　256×256				EfficientNet-B4　384×384				EfficientNet-B7　512×512			
		グレード1	グレード2	グレード3	グレード4	グレード1	グレード2	グレード3	グレード4	グレード1	グレード2	グレード3	グレード4
A	*sensitivity*	0.888	0.674	0.753	0.783	0.637	0.831	0.711	0.852	0.744	0.878	0.800	0.883
	precision	0.669	0.795	0.779	0.979	0.784	0.607	0.831	0.897	0.882	0.725	0.867	0.855
	F1-measure	0.763	0.729	0.766	0.870	0.703	0.701	0.766	0.874	0.807	0.794	0.832	0.869
B	*sensitivity*	0.913	0.630	0.737	0.667	0.911	0.867	0.844	0.817	0.978	0.644	0.833	0.767
	precision	0.613	0.795	0.795	0.955	0.882	0.830	0.826	0.961	0.677	0.879	0.862	0.979
	F1-measure	0.734	0.703	0.765	0.785	0.896	0.848	0.835	0.883	0.800	0.744	0.847	0.860
C	*sensitivity*	0.916	0.767	0.556	0.250	0.578	0.711	0.675	0.480	0.622	0.854	0.744	0.833
	precision	0.585	0.793	0.667	0.484	0.559	0.736	0.767	0.400	0.862	0.717	0.838	0.641
	F1-measure	0.714	0.780	0.606	0.330	0.568	0.723	0.718	0.436	0.723	0.779	0.788	0.725
D	*sensitivity*	0.878	0.500	0.719	0.417	0.844	0.711	0.667	0.417	0.889	0.828	0.615	0.650
	precision	0.585	0.726	0.653	0.735	0.724	0.674	0.625	0.735	0.800	0.700	0.757	0.780
	F1-measure	0.702	0.592	0.684	0.532	0.779	0.692	0.645	0.532	0.842	0.759	0.679	0.709

表 4-8 各モデルの混同行列

（EfficientNet-B1，256×256 画素，データセット A～D）

A	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	79	9	1	0	**0.888**	**0.669**	**0.763**
Gr2	20	62	10	0	**0.674**	**0.795**	**0.729**
Gr3	15	6	67	1	**0.753**	**0.779**	**0.766**
Gr4	4	1	8	47	**0.783**	**0.979**	**0.870**

B	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	84	4	2	2	**0.913**	**0.613**	**0.734**
Gr2	26	58	8	0	**0.630**	**0.795**	**0.703**
Gr3	16	9	70	0	**0.737**	**0.795**	**0.765**
Gr4	11	2	8	42	**0.667**	**0.955**	**0.785**

C	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	76	0	0	7	**0.916**	**0.585**	**0.714**
Gr2	16	69	0	5	**0.767**	**0.793**	**0.780**
Gr3	24	12	50	4	**0.556**	**0.667**	**0.606**
Gr4	14	6	25	15	**0.250**	**0.484**	**0.330**

D	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	79	2	2	7	**0.878**	**0.585**	**0.702**
Gr2	32	45	12	1	**0.500**	**0.726**	**0.592**
Gr3	15	9	64	1	**0.719**	**0.653**	**0.684**
Gr4	9	6	20	25	**0.417**	**0.735**	**0.532**

表 4-9 各モデルの混同行列

（EfficientNet-B4，384×384 画素，データセット A～D）

A	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	58	29	2	2	**0.637**	**0.784**	**0.703**
Gr2	6	74	7	2	**0.831**	**0.607**	**0.701**
Gr3	8	16	64	2	**0.711**	**0.831**	**0.766**
Gr4	2	3	4	52	**0.852**	**0.897**	**0.874**

B	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	82	3	3	2	**0.911**	**0.882**	**0.896**
Gr2	4	78	8	0	**0.867**	**0.830**	**0.848**
Gr3	2	12	76	0	**0.844**	**0.826**	**0.835**
Gr4	5	1	5	49	**0.817**	**0.961**	**0.883**

C	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	52	5	1	32	**0.578**	**0.559**	**0.568**
Gr2	19	64	4	3	**0.711**	**0.736**	**0.723**
Gr3	11	15	56	1	**0.675**	**0.767**	**0.718**
Gr4	11	3	12	24	**0.480**	**0.400**	**0.436**

D	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	76	6	3	5	**0.844**	**0.724**	**0.779**
Gr2	15	64	8	3	**0.711**	**0.674**	**0.692**
Gr3	8	21	60	1	**0.667**	**0.625**	**0.645**
Gr4	6	4	25	25	**0.417**	**0.735**	**0.532**

表 4-10 各モデルの混同行列

（EfficientNet-B7，512×512 画素，データセット A～D）

A	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	67	16	2	5	**0.744**	**0.882**	**0.807**
Gr2	5	79	6	0	**0.878**	**0.725**	**0.794**
Gr3	2	12	72	4	**0.800**	**0.867**	**0.832**
Gr4	2	2	3	53	**0.883**	**0.855**	**0.869**

B	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	88	0	2	0	**0.978**	**0.677**	**0.800**
Gr2	26	58	6	0	**0.644**	**0.879**	**0.744**
Gr3	9	5	75	1	**0.833**	**0.862**	**0.847**
Gr4	7	3	4	46	**0.767**	**0.979**	**0.860**

C	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	56	9	3	22	**0.622**	**0.862**	**0.723**
Gr2	5	76	3	5	**0.854**	**0.717**	**0.779**
Gr3	2	20	67	1	**0.744**	**0.838**	**0.788**
Gr4	2	1	7	50	**0.833**	**0.641**	**0.725**

D	Gr1	Gr2	Gr3	Gr4	***recall***	***precision***	***F1-measure***
Gr1	80	5	1	4	**0.889**	**0.800**	**0.842**
Gr2	6	77	6	4	**0.828**	**0.700**	**0.759**
Gr3	9	23	56	3	**0.615**	**0.757**	**0.679**
Gr4	5	5	11	39	**0.650**	**0.780**	**0.709**

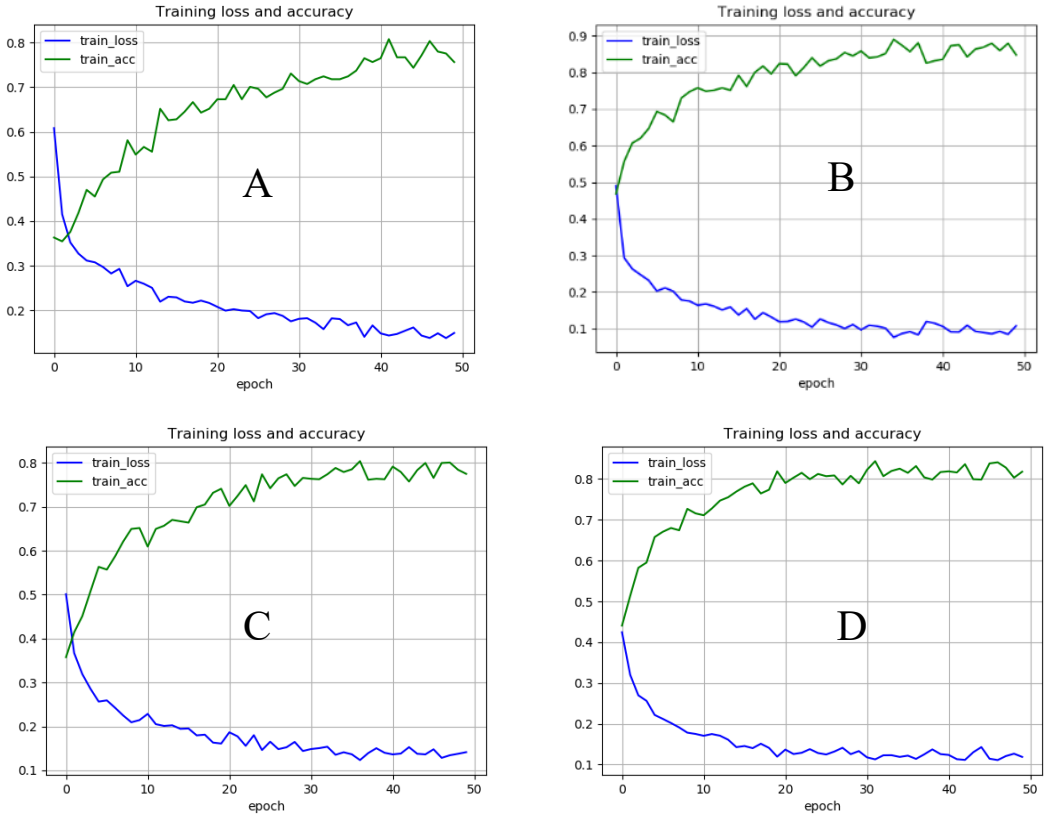


図 4-17 各モデルの混同行列

（EfficientNet-B1，256×256 画素，データセット A～D）

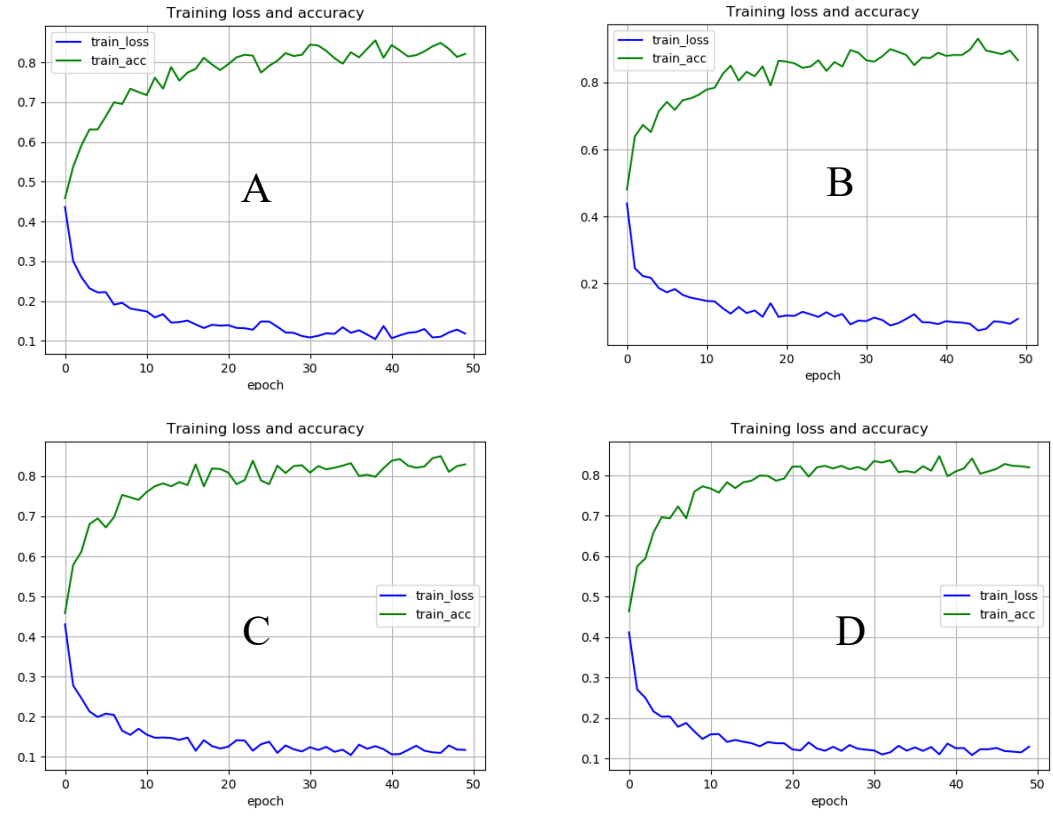


図 4-18 各モデルの混同行列

（EfficientNet-B4，384×384 画素，データセット A～D）

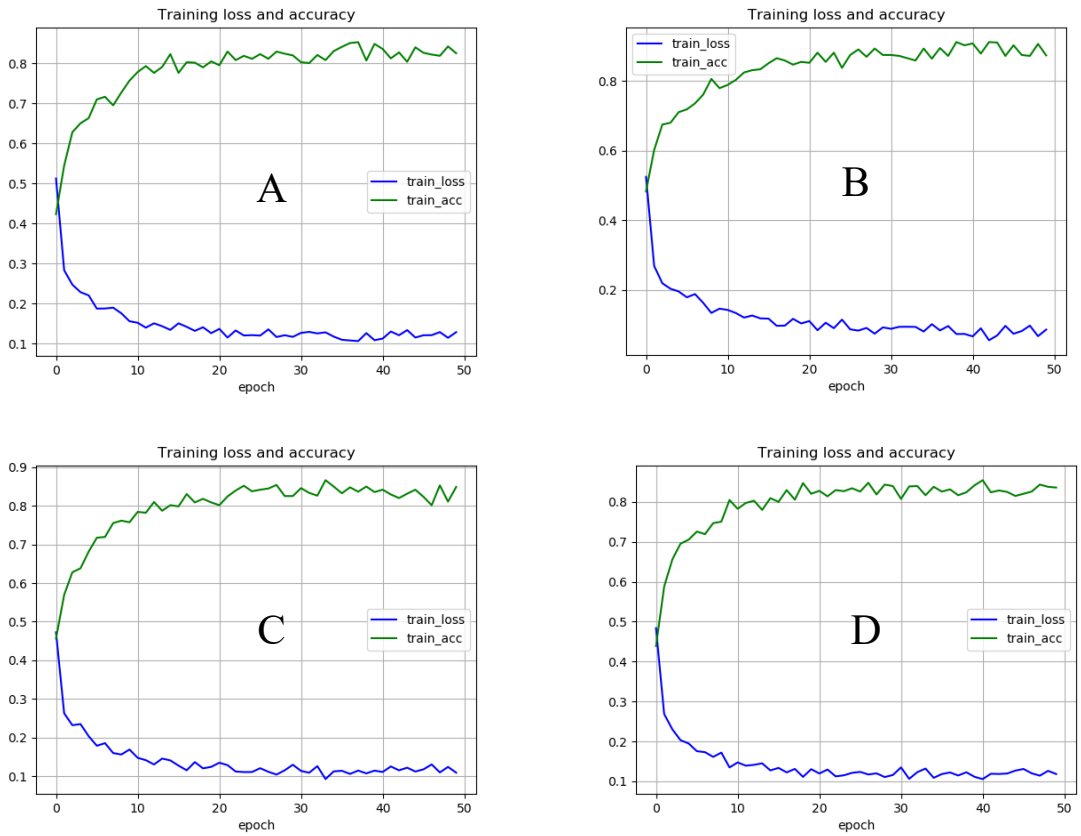


図 4-19 各モデルの混同行列
（EfficientNet-B4，512×512 画素，データセット A～D）

4.8.4 Hyb-RDGS と EfficientNet モデルの性能比較

表 4-7 に示したように本研究で作成した最も正答率の高い EfficientNet のモデルは，データセットに 384 画素の DA を用いた EfficientNet-B4 モデルであった．EfficientNet-B4 のグレード判定モデルの正答率は 86.4%であり，第 3 章で開発した Hyb-RDGS の正答率 85.1%に対して 1.3%上回る正答率を達成した．表 4-11 に Hyb-RDG と本研究で作成したグレード判定モデルと性能比較を示す．Efficient-Net-B4 モデルは，*sensitivity* の性能を上回るが（$p=0.03$），グレード 1～3 に対しては Hyb-RDGS と有意な差はみられず，同等の性能であった（n. s.）．また，グレード 4 に対する性能は，*sensitivity*（$p<0.05$），*F1-measure*（n. s.）は，低下したが，*precision* においては Hyb-RDGS を有意に上回る性能であった（$p<0.05$）．

表 4-11 Hyb-RDGS と EfficientNet モデルの性能比較

(a)

	Classification Accuracy	
	EfficientNet-B4	Hyb-RDGS
Overall	0.864	0.851
P value	0.025*	

(b)

	sensitivity		**precision**		**F1-measure**	
	EfficientNet	Hyb	EfficientNet	Hyb	EfficientNet	Hyb
Grade 1	0.911	0.900	0.882	0.794	0.896	0.844
P value	0.217		0.208		0.111	
Grade 2	0.867	0.778	0.830	0.898	0.848	0.833
P value	0.03*		0.233		0.150	
Grade 3	0.844	0.833	0.826	0.882	0.835	0.856
P value	0.082		0.075		0.511	
Grade 4	0.817	0.933	0.961	0.847	0.833	0.885
P value	0.035*		0.001**		0.088	
Average	0.860	0.861	0.875	0.853	0.853	0.855
P value	0.489		0.35		0.275	

*; p< 0.05 **; p< 0.01

4.8.5 ベイズ推定に基づく放射線皮膚炎のグレード最終判定結果

前項でデータセット検証を実施した結果より，3 つのモデルそれぞれのベイズ推定量を算出した．はじめに各グレードの事後確率を算出し，事後確率を基に評価者のグレードにおけるベイズ推定量を算出した．図 4-20 に示すように算出したベイズ推定量が最も小さいモデルの判定結果を最終判定とした．表 4-12 に事後確率，表 4-13 に各モデルのベイズ推定量を示す．また，評価者がグレード 1 または，2 と判定した複数（グレード）判定画像に対する最終判定の結果を表 4-14，表 4-15 に示す．

最終判定が評価者の判定グレード 1 または，2 と一致した割合は 96.7%，2 または，3 と一致した割合は 93.3%であった．ここでは，評価者により相違が生じている 2 つの対象グレードのうち，どちらか一方の判定と一致した割合を指している． ここで，複数の判定がグレード 1 または，2 の場合，複数（グレード 1，2）画像，グレード 2 または，3 の場合，複数（グレード 2，3）画像とする．

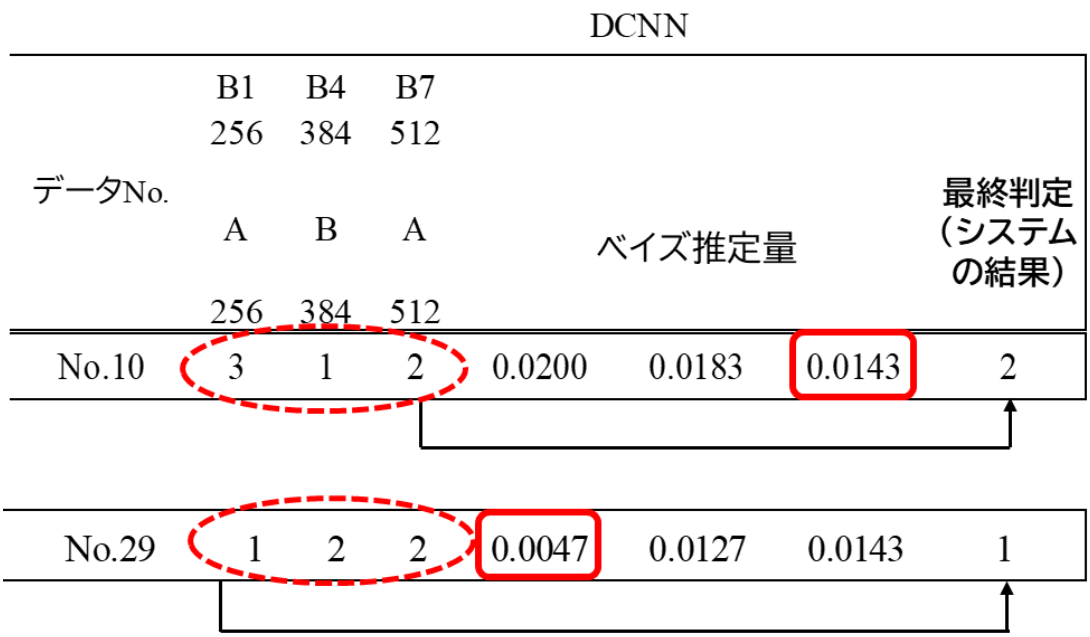


図 4-20 ベイズ推定量を用いた最終判定例

表 4-12 各モデルの事後確率

推定グレード （事後確率$q(x)$）	EfficientNet-B1 256×256 データセットA	EfficientNet-B4 384×384 データセットB	EfficientNet-B7 512×512 データセットA
グレード1	0.9892	0.9942	0.9846
グレード2	0.9293	0.9712	0.9675
グレード3	0.9549	0.9740	0.9675
グレード4	0.9833	0.9871	0.9898

表 4-13 各モデルのベイズ推定量

真のグレード （評価者の確率$p(x)$）	EfficientNet-B1 256×256 データセットA	EfficientNet-B4 384×384 データセットB	EfficientNet-B7 512×512 データセットA
グレード1（1,0,0,0）	0.0047	0.0183	0.0106
グレード2（0,1,0,0）	0.0318	0.0127	0.0143
グレード3（0,0,1,0）	0.0200	0.0114	0.0144
グレード4（0,0,0,1）	0.0073	0.0056	0.0044

表 4-14 複数（グレード 1，2）判定画像の最終グレード判定結果

	各モデル判定結果			ベイズ推定量			
EfficienNet	B1	B4	B7	B1	B4	B7	最終判定
	256	384	512	256	384	512	
データセット	A	B	A	A	B	A	
解像度	256	384	512	256	384	512	
No.1	1	1	2	0.0047	0.0183	0.0143	1
No.2	1	1	1	0.0047	0.0183	0.0106	1
No.3	2	1	2	0.0318	0.0183	0.0143	2
No.4	2	2	2	0.0318	0.0127	0.0143	2
No.5	1	1	2	0.0047	0.0183	0.0143	1
No.6	2	1	1	0.0318	0.0183	0.0106	1
No.7	1	1	3	0.0047	0.0183	0.0144	1
No.8	1	1	1	0.0047	0.0183	0.0106	1
No.9	3	1	2	0.0200	0.0183	0.0143	2
No.10	2	1	2	0.0318	0.0183	0.0143	2
No.11	1	1	1	0.0047	0.0183	0.0106	1
No.12	1	1	3	0.0047	0.0183	0.0144	1
No.13	2	3	2	0.0318	0.0114	0.0143	3
No.14	1	2	3	0.0047	0.0127	0.0144	1
No.15	1	1	2	0.0047	0.0183	0.0143	1
No.16	2	1	2	0.0318	0.0183	0.0143	2
No.17	2	2	1	0.0318	0.0127	0.0106	1
No.18	1	2	2	0.0047	0.0127	0.0143	1
No.19	1	1	2	0.0047	0.0183	0.0143	1
No.20	2	2	3	0.0318	0.0127	0.0144	2
No.21	2	2	2	0.0318	0.0127	0.0143	2
No.22	1	2	2	0.0047	0.0127	0.0143	1
No.23	1	1	2	0.0047	0.0183	0.0143	1
No.24	1	2	2	0.0047	0.0127	0.0143	1
No.25	3	1	2	0.0200	0.0183	0.0143	2
No.26	1	2	1	0.0047	0.0127	0.0106	1
No.27	2	1	2	0.0318	0.0183	0.0143	2
No.28	1	1	2	0.0047	0.0183	0.0143	1
No.29	1	2	2	0.0047	0.0127	0.0143	1
No.30	2	2	2	0.0318	0.0127	0.0143	2
評価者のグレード判定と一致した割合（グレード1または，2）							96.7%

表 4-15 複数（グレード 2，3）判定画像の最終グレード判定結果

	各モデル判定結果			ベイズ推定量			
EfficienNet	B1	B4	B7	B1	B4	B7	最終判定
	256	384	512	256	384	512	
データセット	A	B	A	A	B	A	
解像度	256	384	512	256	384	512	
No.1	2	1	2	0.0318	0.0183	0.0143	2
No.2	3	2	2	0.0200	0.0127	0.0143	2
No.3	2	2	2	0.0318	0.0127	0.0143	2
No.4	2	1	2	0.0318	0.0183	0.0143	2
No.5	2	2	3	0.0318	0.0127	0.0144	3
No.6	2	2	2	0.0318	0.0127	0.0143	2
No.7	3	2	2	0.0200	0.0127	0.0143	2
No.8	2	3	2	0.0318	0.0114	0.0143	3
No.9	2	2	2	0.0318	0.0127	0.0143	2
No.10	3	1	2	0.0200	0.0183	0.0143	2
No.11	2	2	1	0.0318	0.0127	0.0106	1
No.12	2	1	2	0.0318	0.0183	0.0143	2
No.13	2	1	2	0.0318	0.0183	0.0143	2
No.14	2	3	3	0.0318	0.0114	0.0144	3
No.15	2	2	2	0.0318	0.0127	0.0143	2
No.16	3	3	3	0.0200	0.0114	0.0144	3
No.17	3	3	3	0.0200	0.0114	0.0144	3
No.18	3	3	3	0.0200	0.0114	0.0144	3
No.19	3	3	3	0.0200	0.0114	0.0144	3
No.20	3	2	1	0.0200	0.0127	0.0106	1
No.21	3	3	3	0.0200	0.0114	0.0144	3
No.22	3	3	3	0.0200	0.0114	0.0144	3
No.23	3	3	2	0.0200	0.0114	0.0143	3
No.24	3	2	2	0.0200	0.0127	0.0143	2
No.25	3	2	3	0.0200	0.0127	0.0144	3
No.26	3	2	3	0.0200	0.0127	0.0144	3
No.27	3	3	2	0.0200	0.0114	0.0143	3
No.28	3	3	2	0.0200	0.0114	0.0143	3
No.29	3	3	3	0.0200	0.0114	0.0144	3
No.30	3	2	2	0.0200	0.0127	0.0143	2
評価者のグレード判定と一致した割合（グレード2または，3）							93.3%

4.8.6　ベイズ推定と評価者の判定結果の分析

表 4-14，4-15 に示したように，複数（グレード）判定画像に対するシステムは，評価者の判定が相違する場合に DCNN は，どちらか一方の判定を高い精度で示す結果であった．さらに，本項では，評価者 6 名（医師 2 名，看護師 4 名）の判定に対して，本研究で作成したシステムがどのような結果を示したのか，評価する必要がある．

ベイズ推定と評価者 6 名の判定結果の関係について分析した結果を表 4-18，4-19 にまとめ，図 4-22，4-23 にバブルチャートを示す．図 4-21 は，バブルチャートの説明図である．

分析の詳細を以下の表 4-15～4-18 に示す．

＜複数（グレード 1，2）判定画像＞

表 4-18：システム最終判定結果と評価者の判定の割合

図 4-22（a）：評価者とシステムの関係

（システムの判定結果グレード 1 の場合のバブルチャート）

図 4-23（b）：評価者とシステムの関係

（システムの判定結果グレード 2 の場合のバブルチャート）

＜複数（グレード 2，3）判定画像＞

表 4-19：システム最終判定結果と評価者の判定の割合

図 4-22（a）：評価者とシステムの関係

（システムの判定結果グレード 1 の場合のバブルチャート）

図 4-23（b）：評価者とシステムの関係

（システムの判定結果グレード 2 の場合のバブルチャート）

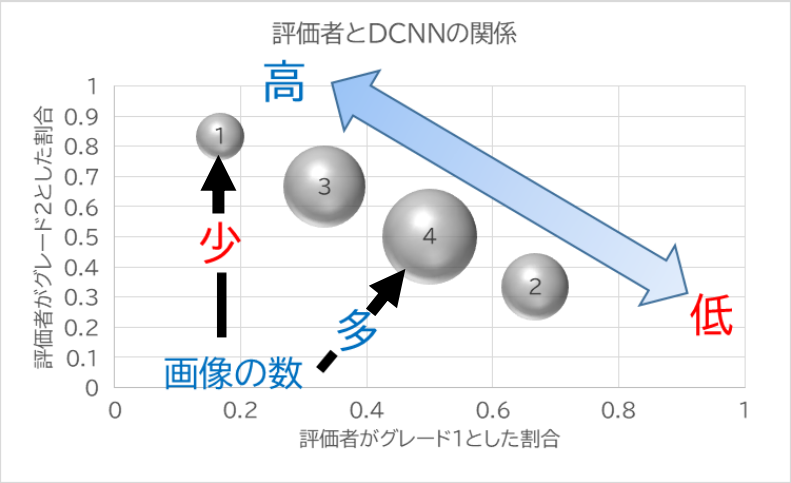


図 4-21 システムと評価者の判定結果の関係の説明図

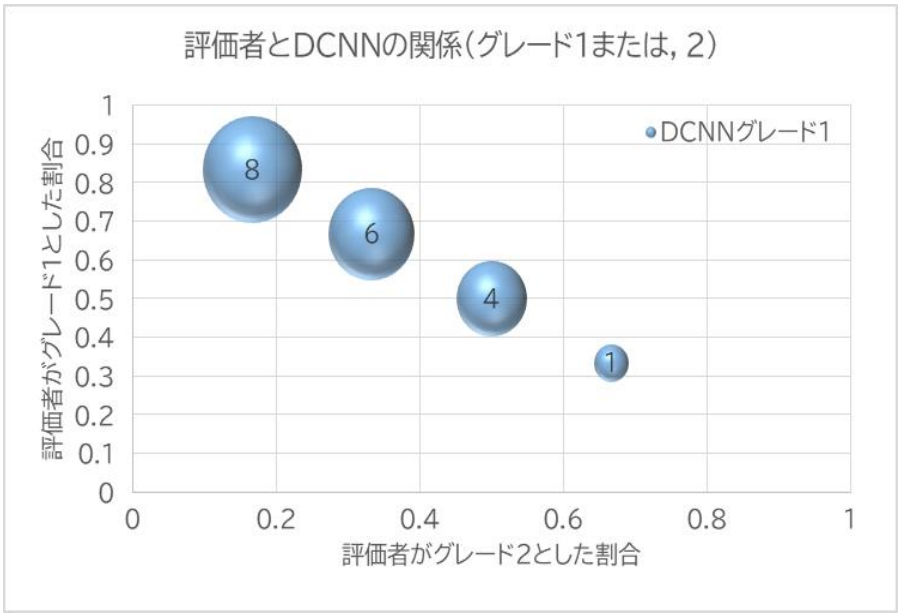
表 4-16 より，システムの判定結果は，評価者 6 名の判定と一致した割合（≧50%．≧66%，≧83%）は，複数（グレード 1，2）画像判定，DCNN グレード 1 に対して 95%，74%，42%であった．これは，具体的にグレード 1 は，6 名中 3 名の評価者と 95%，4 名の評価者と 74%，5 名の評価者と 42%で一致していることを示している．また，表 4-17 より複数（グレード 2，3）画像判定，DCNN グレード 3 に対して 64%，18%，0%であり，グレード 1，2 に対して下回る結果であった．

また，図 4-22，4-23 に示したバブルチャートは，左上になるほど判定者と一致している割合が高いことを示しており，バブルの大きさは，その画像の数である．図 4-22（b）,図 4-23（b）（システムの判定結果グレード 2，グレード 3）より，評価者との関係は，バブルチャートが右下に分布の傾向を示しており，やや判定結果の一致の割合が低い傾向にあることがいえる．

表 4-16 システムと評価者の判定結果一致の割合（グレード 1，2）

DCNNの結果	評価者とシステムの一致した割合		
	≧50%（3/6）	≧66%（4/6）	≧83%（5/6）
グレード1	95%	74%	42%
グレード2	80%	40%	0%

(a)



(b)

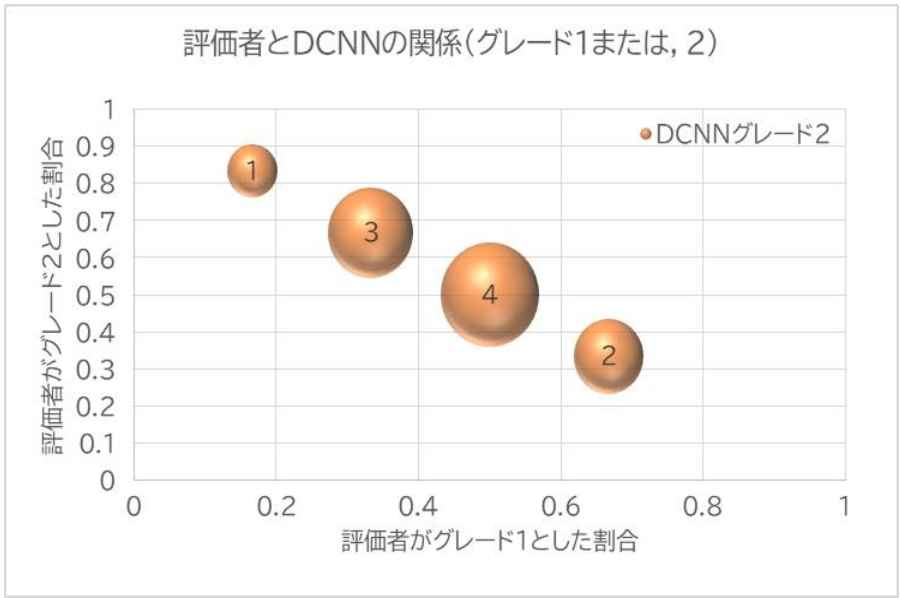


図 4-22 システムと評価者の判定結果の関係

（a）DCNN グレード 1 の場合の評価者の判定との関係

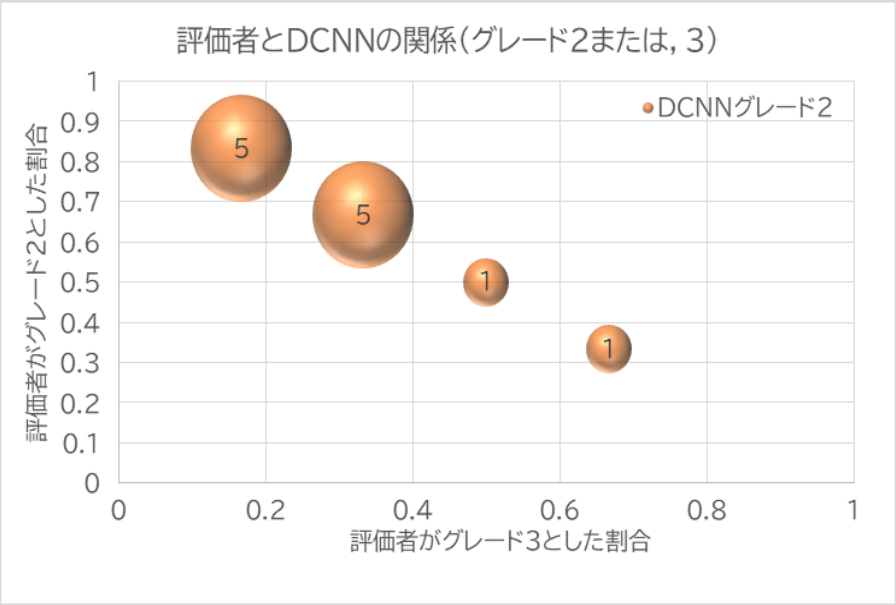
（b）DCNN グレード 2 の場合の評価者の判定との関係

表 4-17 システムと評価者の判定結果一致の割合（グレード 2，3）

複数(グレード2または, 3)判定画像

DCNNの結果	評価者とシステムの一致した割合		
	≧50%(3/6)	≧66%(4/6)	≧83%(5/6)
グレード2	88%	71%	41%
グレード3	64%	18%	0%

(a)



(b)

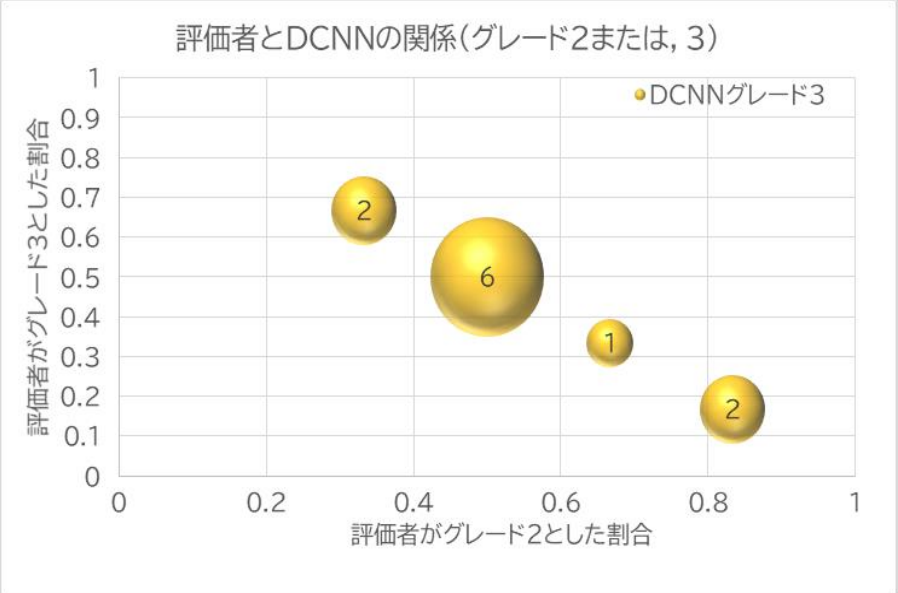


図 4-23 システムと評価者の判定結果の関係

(a) DCNN グレード 2 の場合の評価者の判定との関係

(b) DCNN グレード 3 の場合の評価者の判定との関係

表 4-18 システムと評価者の判定結果一致の割合
（複数（グレード 1，2）判定画像における**グレード 1** の割合））

	DCNN							評者者				
データNo.	B1 256 A 256	B4 384 B 384	B7 512 A 512	ベイズ推定量			最終判定（システムの結果）	評者者の判定		評者者6名がグレード1と判定した割合		
								グレード1の割合	グレード2の割合	≧50% (3/6)	≧66% (4/6)	≧83% (5/6)
										（表示：≧X%=1）		
No.1	1	1	2	0.0047	0.0183	0.0143	**1**	**0.667**	0.333	1	1	0
No.2	1	1	1	0.0047	0.0183	0.0106	**1**	**0.833**	0.167	1	1	1
No.5	1	1	2	0.0047	0.0183	0.0143	**1**	**0.833**	0.167	1	1	1
No.6	2	1	1	0.0318	0.0183	0.0106	**1**	**0.667**	0.333	1	1	0
No.7	1	1	3	0.0047	0.0183	0.0144	**1**	**0.667**	0.333	1	1	0
No.8	1	1	1	0.0047	0.0183	0.0106	**1**	**0.833**	0.167	1	1	1
No.11	1	1	1	0.0047	0.0183	0.0106	**1**	**0.833**	0.167	1	1	1
No.12	1	1	3	0.0047	0.0183	0.0144	**1**	**0.333**	0.667	0	0	0
No.14	1	2	3	0.0047	0.0127	0.0144	**1**	**0.833**	0.167	1	1	1
No.15	1	1	2	0.0047	0.0183	0.0143	**1**	**0.500**	0.500	1	0	0
No.17	2	2	1	0.0318	0.0127	0.0106	**1**	**0.500**	0.500	1	0	0
No.18	1	2	2	0.0047	0.0127	0.0143	**1**	**0.833**	0.167	1	1	1
No.19	1	1	2	0.0047	0.0183	0.0143	**1**	**0.667**	0.333	1	1	0
No.22	1	2	2	0.0047	0.0127	0.0143	**1**	**0.833**	0.167	1	1	1
No.23	1	1	2	0.0047	0.0183	0.0143	**1**	**0.667**	0.333	1	1	0
No.24	1	2	2	0.0047	0.0127	0.0143	**1**	**0.667**	0.333	1	1	0
No.26	1	2	1	0.0047	0.0127	0.0106	**1**	**0.833**	0.167	1	1	1
No.28	1	1	2	0.0047	0.0183	0.0143	**1**	**0.500**	0.500	1	0	0
No.29	1	2	2	0.0047	0.0127	0.0143	**1**	0.500	0.500	1	0	0
システムと評価者の一致する割合								**95%**		**74%**		**42%**

表 4-19 システムと評価者の判定結果一致の割合
（複数（グレード 1，2）判定画像における**グレード 2** の割合））

	DCNN							評者者				
データNo.	B1 256 A 256	B4 384 B 384	B7 512 A 512	ベイズ推定量			最終判定（システムの結果）	評者者の判定		評者者6名がグレード2と判定した割合		
								グレード1の割合	**グレード2の割合**	≧50% (3/6)	≧66% (4/6)	≧83% (5/6)
										（表示：≧X%=1）		
No.3	2	1	2	0.0318	0.0183	0.0143	**2**	0.500	**0.500**	1	0	0
No.4	2	2	2	0.0318	0.0127	0.0143	**2**	0.667	**0.333**	1	1	0
No.9	3	1	2	0.0200	0.0183	0.0143	**2**	0.333	**0.667**	0	0	0
No.10	2	1	2	0.0318	0.0183	0.0143	**2**	0.500	**0.500**	1	0	0
No.16	2	1	2	0.0318	0.0183	0.0143	**2**	0.833	**0.167**	1	1	0
No.20	2	2	3	0.0318	0.0127	0.0144	**2**	0.667	**0.333**	1	1	0
No.21	2	2	2	0.0318	0.0127	0.0143	**2**	0.667	**0.333**	1	1	0
No.25	3	1	2	0.0200	0.0183	0.0143	**2**	0.333	**0.667**	0	0	0
No.27	2	1	2	0.0318	0.0183	0.0143	**2**	0.500	**0.500**	1	0	0
No.30	2	2	2	0.0318	0.0127	0.0143	**2**	0.500	**0.500**	1	0	0
システムと評価者の一致する割合									**80%**		**40%**	**0%**

表 4-20 システムと評価者の判定結果一致の割合
（複数（グレード 2，3）判定画像における**グレード 2** の割合））

	DCNN							評価者				
	B1 256	B4 384	B7 512					評価者の判定		評価者6名がグレード1と判定した割合		
データNo.	A 256	B 384	A 512	ベイズ推定量			最終判定（システムの結果）	**グレード2の割合**	グレード3の割合	≧50% (3/6)	≧66% (4/6)	≧83% (5/6)
										（表示:≧X%=1）		
No.1	2	1	2	0.0318	0.0183	0.0143	2	**0.667**	0.333	1	1	0
No.2	3	2	2	0.0200	0.0127	0.0143	2	**0.833**	0.167	1	1	1
No.3	2	2	2	0.0318	0.0127	0.0143	2	**0.667**	0.333	1	1	0
No.4	2	1	2	0.0318	0.0183	0.0143	2	**0.167**	0.833	0	0	0
No.6	2	2	2	0.0318	0.0127	0.0143	2	**0.833**	0.167	1	1	1
No.7	3	2	2	0.0200	0.0127	0.0143	2	**0.333**	0.667	0	0	0
No.9	2	2	2	0.0318	0.0127	0.0143	2	**0.667**	0.333	1	1	0
No.10	3	1	2	0.0200	0.0183	0.0143	2	**0.833**	0.167	1	1	1
No.12	2	1	2	0.0318	0.0183	0.0143	2	**0.667**	0.333	1	1	0
No.13	2	1	2	0.0318	0.0183	0.0143	2	**0.833**	0.167	1	0	1
No.15	2	2	2	0.0318	0.0127	0.0143	2	**0.833**	0.167	1	0	1
No.24	3	2	2	0.0200	0.0127	0.0143	2	**0.833**	0.167	1	1	1
No.25	3	2	3	0.0200	0.0127	0.0144	2	**0.667**	0.333	1	1	0
No.26	3	2	3	0.0200	0.0127	0.0144	2	**0.667**	0.333	1	1	0
No.27	3	3	2	0.0200	0.0114	0.0143	2	**0.667**	0.333	1	1	0
No.28	3	3	2	0.0200	0.0114	0.0143	2	**0.500**	0.500	1	0	0
No.30	3	2	2	0.0200	0.0127	0.0143	2	**0.833**	0.167	1	1	1
システムと評価者の一致する割合										**88%**	**71%**	**41%**

表 4-21 システムと評価者の判定結果一致の割合
（複数（グレード 2，3）判定画像における**グレード 3** の割合））

	DCNN							評価者				
	B1 256	B4 384	B7 512					評価者の判定		評価者6名がグレード3と判定した割合		
データNo.	A 256	B 384	A 512	ベイズ推定量			最終判定（システムの結果）	グレード2の割合	**グレード3の割合**	≧50% (3/6)	≧66% (4/6)	≧83% (5/6)
										（表示:≧X%=1）		
No.5	2	2	3	0.0318	0.0127	0.0144	**3**	0.500	**0.500**	1	0	0
No.8	2	3	2	0.0318	0.0114	0.0143	**3**	0.500	**0.500**	0	0	0
No.14	2	3	3	0.0318	0.0114	0.0144	**3**	0.833	**0.167**	0	0	0
No.16	3	3	3	0.0200	0.0114	0.0144	**3**	0.500	**0.500**	1	0	0
No.17	3	3	3	0.0200	0.0114	0.0144	**3**	0.500	**0.500**	1	0	0
No.18	3	3	3	0.0200	0.0114	0.0144	**3**	0.667	**0.333**	0	0	0
No.19	3	3	3	0.0200	0.0114	0.0144	**3**	0.333	**0.667**	1	1	0
No.21	3	3	3	0.0200	0.0114	0.0144	**3**	0.833	**0.167**	0	0	0
No.22	3	3	3	0.0200	0.0114	0.0144	**3**	0.500	**0.500**	1	0	0
No.23	3	3	2	0.0200	0.0114	0.0143	**3**	0.333	**0.667**	1	1	0
No.29	3	3	3	0.0200	0.0114	0.0144	**3**	0.500	**0.500**	1	0	0
システムと評価者の一致する割合										**64%**	**18%**	**0%**

4.9　考察

4.9.1　EfficientNet モデルの構成について

解像度とモデルの関係は，図 4-12 に示したように，モデルのスケーリングが大きくなるに従い test accuracy が向上した．しかし，各モデルの構成を比較するとモデルのスケーリングが低い B0～B3 では，解像度によって test accuracy にばらつきがみられた．特に B1 は，画素数によって 65～81％となり test accuracy に 16％の差が生じた．これに対して B4～B7 は，76～84％でありその差は 8％であった．これより，EfficientNet モデルは，スケーリングが低いほど解像度の影響を受けやすいと考えられる．一方，スケーリングが最も高い B7 では，384×384 画素の精度が低下した．本検証は，解像度を一定にした前評価でテストデータにおける accuracy の低下，loss の上昇が起きていないエポック数を選択している．EfficientNet の複合スケーリングは，解像度だけでなくネットワークの深さ，幅を固定された比率でスケールアップする [20]．図 4-13 の学習曲線で accuracy の低下傾向が見られたことから，解像度のみ変化させた場合でも最もスケールアップされた B7 のモデルは過学習を起こしている可能性も考えられる．

これまで DCNN では，層を深く，幅を広く，解像度を上げること，どれか一つの要素を増やすことで精度を向上させてきた．ネットワークの精度は重要な要素だが，パラメータ数が多くなりハードウェアのメモリに限界があった． 前述したように，Srinivasu ら [59] の moblie-size の DCNN 研究報告の目標解像度は 224×224 画素とされている．本研究の EfficientNet モデルの解像度は 256～512 画素を設定した．表 3-7 に示したように必ずしも高解像度の画像を使用した方が，精度が高いというわけではなく，解像度とスケーリングの構成を最適化することで，より汎化性能，精度の高いグレード判定モデルが作成できると考えられる．また，図 4-14 より，解像度の違いは学習時間への影響は及ぼさないが，モデルスケーリングが大きくなると学習時間の延長に繋がると考えられる．

4.9.2　EfficientNet を用いたグレード判定モデルの性能について

表 4-7 に示したように本研究で作成した最も正答率の高い EfficientNet のモデルは，データセットに 384 画素の DA を用いた EfficientNet-B4 モデルであった．EfficientNet-B4 のグレード判定モデルの正答率は 86.4%であり，第 3 章で開発した Hyb-RDGS の正答率 85.1%に対して 1.3%上回る正答率を達成した．

表 4-11 に示した Hyb-RDGS と EfficientNet モデルの性能比較について，課題であった症例数の少ないグレード 4 の性能は，Hyb-RDGS に対して EfficientNet モデルは *sensitivity*，*F1-measure* が，それぞれ 1.14%（0.817/0.933），1.03%（0.855/0.883）低下した．一方，*precision* は，1.13%（0.961/0.847）向上した．これより，EfficientNet モデルのグレード判定は，グレード 4 と予測したものは，正しく予測している割合が高いが，グレードの異なる放射線皮膚炎画像の中からグレード 4 の画像を正しく予測できる割合が Hyb-RDGS に対して，やや劣ると考えられる．しかし，*sensitivity* と *precision* には，トレードオフの関係がある．そこで，*F1-measure* を指標にした場合，両者に有意な差はみられないことから同等の性能であることがいえる．

Hyb-RDGS は，放射線皮膚炎画像データの不均衡に対して DA にポアソン合成による人工症例画像を加えたデータセットを用いて性能が向上した．しかし，EfficientNet モデルでは，表 3-9 に示したように RA と DA を用いたデータセット A，B に対して人工症例を用いた C，D では正答率が 14%（77.3-63.6）～27%（86.4-59.4）低下した．特に C におけるグレード 4 については 512 から 256 画素の解像度低下に伴い，*sensitivity*，*F1-measure* は 58%（0.833-0.250），40%（0.725-0.330）の低下を示した．人工症例画像は，その特徴である炎症部を正常皮膚画像に埋め込み作成される．そのため，本質的な炎症部のデータ拡張とは異なり解像度の低いデータセットでは，学習不足になっていた可能性が考えられる．これに対して，DA は各データに最大 3 種類の画像処理を行い，RA は 14 個の画像処理を対象として，*k* 個の画像処理操作がそれぞれ均等に行われる [31]．炎症部を

含む画像処理がデータ拡張に有効に働いた可能性が考えられる.

本研究で用いた EfficientNet は，MobileNet と ResNet のスケールアップと NAS によって作られた新しいモデルのスケールアップを行った新しいモデルである. EfficientNet は B0～B7 までのスケーリングアップしたモデルが提案されており，ResNet や NAS などの精度比較が報告されている [20]. Srinivasu ら [59] は MobileNet を用いた最小限のパフォーマンス評価として皮膚疾患の分類精度は 85.3%であったと報告している. しかし，特徴抽出のランダム性に欠けており一連の画像テストでは， 80%未満に低下したとされている. EfficientNet-B0～B7 における解像度，データセットの構成など，学習データパターンによる研究報告は少なく十分でない. 本章では，解像度とデータ拡張の条件を変えた EfficientNet を用いたモデルを作成し，評価した. EfficientNet モデルは解像度およびデータセットに依存することから，少なくとも解像度，データセット構成はモデル性能に寄与する. 本章では，mobile-size の低解像度， DA のデータ拡張で高い性能を示した.

4.9.3 ベイズ推定による最終グレード判定について

放射線皮膚炎の複数（グレード）判定画像に対して，3 個の EfficientNet モデルに対してベイズ推定を用いた最終グレード判定を行う手法を提案した.

最終判定に用いた EfficientNet-B1，B4 および B7 モデルのそれぞれの性能は，正答率が 77.3%，86.4%および 82.1%であった. 本研究では，それぞれの EfficientNet モデルのグレード毎の正答率を基に予測された事後確率から，評価者のグレードであるとされた場合のベイズ推定量を 4.7.2 項で示した（4.5）式から求めた. (4.5) 式は，確率分布 p と q の近似性を表現している対数関数である. q は，真のグレードと同じになる確率を表現するので，近似すると誤差が小さくなり，近似しなくなると誤差が大きくなる. つまり，それぞれの EfficientNet モデルが判定したグレードについて，評価者の確率分布に近いベイズ推定量を評価することで最も正解に近いグレードを予測できる.

また，本章で最終グレード判定に用いた EfficientNet モデルは，データの不均衡を補うデータセットを複数作成したものである．表 4-7 に示したように，それぞれグレード毎の *sensitivity* が異なる．提案したベイズ推定は，モデルの *sensitivity* が異なれば，事後確率も異なりその結果，ベイズ推定量も変わることになる．ベイズ推定による判定例をあげるとモデル A，B，C の判定がグレード 1，2，2 と判定されたとする．単純多数決では，グレード 2 と判定してしまう．ここでベイズ推定から真のグレードに対する確率で評価する．B，C はグレード 2 に対するベイズ推定量は小さいが，グレード 1 に対しては，グレード A の方がベイズ推定量の小さいモデルであるとする．評価者のグレードに最も近いモデル A の判定（グレード 1）結果が最終判定となる．

本章で提案したベイズ推定を用いた複数（グレード）判定に対する判定結果は，グレード 1～4 の分類を行うモデルにおいて，例えば，評価者がグレード 1 または，2 とした判定（3 または，4 と判定しない）と一致した割合は 96.7%であった．これは，評価者により相違が生じている 2 つの対象グレードのうち，どちらか一方の判定と一致した割合を指すが，正解画像で作成した EfficientNet モデルの出力結果が，曖昧な評価者の判定より正解画像との誤差が小さいと解釈できる．これにより，本システムは，評価に相違が生じた場合にベイズ推定量に基づいて最終判定結果を示し，評価者のグレード判定の補助システムとなることが期待できる．

さらに，本項で提案したベイズ推定は，評価者の判定に基づいてモデルを作成し，モデルの性能を高めた上でモデルの判定の確率分布を計算し，不確かさを定量化した（ベイズ推定量）．ベイズ推定は，モデル自体の不確かさも定量化できるといえる．

ベイズ推定による判定結果が，複数（グレード）判定画像の評価に有効であると述べたが，曖昧な判定となった画像に対して，評価者の判定とどのような関係があり，どのような分布を示しているのか，4.8.6 項で結果の本システムの最終

判定結果と 6 名の評価者の判定について分析した．図 4-22，4-23 で示した評価者とシステムの評価結果の関係より，それぞれの複数（グレード）画像の評価は，グレードの高い画像の一致する割合が低い傾向がみられた．特に表 4-19 と表 4-20 では，同じグレード 2 を含む判定において，グレード 1 と評価値の境界付近にある場合，一致の割合（≧50%，≧66%，≧83%）は，それぞれ 8%，31%，41%の評価に差が生じている．これらの結果より，グレードの境界値と思われる放射線皮膚炎を評価する場合，評価者は低い方のグレードを選択する傾向にあることが考えられる．そのため，図 4-22（b）では，グレード 2，図 4-23（b）では，グレード 3 の一致する割合が低下していることになる．表 4-22 にグレード 2 に対する評価者と一致した割合の比較表，表 4-23 に複数（グレード 2，3）判定画像の評価者 6 名の判定を示す．

表 4-22 より，複数（グレード 1，2）判定の高グレード側であった場合と複数（グレード 2，3）判定画像の低グレード側であった場合でシステムと一致する割合は異なる．また，表 4-23 より，全体の評価をみるとグレードが高くなるにしたがって，評価者とシステムの一致する割合は低下している．これより，システムの評価に対して，評価者は判定する際，グレード境界値付近であった場合，重症度の高い評価を付けにくい思考が働きグレードを低く評価する傾向にある可能性も考えられる．

特に表 4-23 より，医師は看護師に対して低グレード側に評価する傾向がみられており，評価者や職種でも基準が異なる結果であったことが分かる．医師が低い判定を行う傾向について，本研究で扱う放射線皮膚炎は，放射線治療の有害事象である．有害事象をできる限り抑える治療計画を立て，治療を行うことが求められる．医師は，治療の有害事象である放射線皮膚炎を扱う上で重症度の高い症例には，より慎重にならざるを得ない．すなわち，評価する場合は，より重症度の低いグレードを選択することが推測される．

また，EfficientNet モデルの作成については，解像度の条件を変えた 24 パター

ンとデータ拡張の条件を変えた 4 パターンの構成を検討した．最終的に重み付けの異なる 3 個のモデルを選択した．少ないモデルであっても，ベイズ推定を用いて推論することで不確実なグレード判定の推定に有効であることを確認した．

表 4-22 評価結果グレード 2 の割合の比較（低グレード側 vs.高グレード側）

複数(グレード)判定画像におけるグレード2の割合の比較

	DCNNの結果	評価者とシステムの一致した割合		
		≧50%(3/6)	≧66%(4/6)	≧83%(5/6)
グレード2	低グレード側	88%	71%	41%
	高グレード側	80%	40%	0%

表 4-23 評価者 6 名の判定結果

評価者6名の判定					
Doctor1	Doctor1	Nurce1	Nurce2	Nurce3	Nurce4
2	2	2	3	3	2
2	2	2	3	2	2
2	2	2	2	3	3
2	3	3	3	3	3
2	2	3	3	3	2
2	2	2	2	3	2
2	2	3	3	3	3
2	2	2	3	3	3
2	2	3	3	2	2
2	2	2	3	2	2
3	3	2	3	2	2
2	3	2	3	2	2
2	2	2	3	2	2
2	2	2	3	2	2
2	2	3	2	2	2
2	2	3	3	3	2
2	2	3	3	2	3
2	2	3	2	2	3
2	3	2	3	3	3
2	3	2	3	3	3
2	2	2	2	2	3
2	2	3	3	2	3
2	2	2	3	2	2
2	2	3	2	2	2
2	2	3	2	2	3
2	2	3	2	2	3
2	2	3	3	2	2
2	3	3	3	2	3
2	3	3	3	2	3
2	2	2	3	2	2
低グレード	**29**	**23**	**15**	**8**	**20**	**16**
高グレード	**1**	**7**	**15**	**22**	**10**	**14**

4.10 結言

本章では，アンサンブル学習に基づく EfficientNet モデルを用いた放射線皮膚炎のグレード判定システムを開発した [49]．作成した EfficientNet モデルは，Mobile-size の低解像度，DA によるデータ拡張でグレード判定の正答率は 86.4%を達成した．

第 3 章で述べた Hyb-RDGS は，人工症例画像の手順が必要であり，課題 4 の解決には至らなかった．これに対して，本章では，EfficientNet モデル構成の最適パターンを検証し，同等性能を持つグレード判定モデルを確認した．さらに，課題 4 に対して，ベイズ推定を用いて最終グレード判定を行う手法を提案した．EfficientNet モデルと評価者の評価結果を分析することにより，評価者によるグレード判定に対する傾向から，高グレードにしたがってシステムの評価と乖離していく可能性が考えられた．一方で，システムの判定性能が評価の半数以上で高い性能（95%）で一致した．最終判定に用いる分類器の数が少ない場合にも有効であり，特に評価者の判定が相違するグレード 1 と 2，2 と 3 において，評価者の判定と高い精度で一致する性能を持つことを確認した．

本研究で提案した EfficientNet モデルを用いたアンサンブル学習は，評価者のグレード判定が相違した場合の手助けとなり，汎化性能，効率化が期待できると考えられる．

第5章　放射線皮膚炎グレード判定システムの総括（考察）

5.1　先行研究と比較

5.1.1　放射線皮膚炎評価に関する先行研究と比較

第 1 章では，放射線皮膚炎のグレード判定の統一を目的とした 2 つのアプローチによる研究について述べた．一つは，大薗ら[14]による人間が学習することで評価精度を向上させるクイズ形式の学習ソフトの作成である．もう一つは，Zenda ら[15]の放射線皮膚炎グレーディングアトラスの作成である．

前者は，評価者の教育，学習のアプローチから評価レベルの統一と向上を目的として，看護師 10 名による学習ソフトを用いて評価している．学習した結果，最も高い正答率は，89.9%であったと報告している．学習ソフトを用いた手法は，評価の統一手法として有用なツールとされているが，学習を重ねるにつれ，丸暗記してしまうため新しい画像を入れ替える必要があると課題が示されている．

後者は，CTCAE が単文のみで記載されている点が個人の解釈の違いを招くとされ，図 1-7 に示したように CTCAE に基づく頭頸部腫瘍を対象とした放射線皮膚炎アトラスによる視覚的な基準写真を選定し評価統一を行うものである．しかし，アトラスは，作成されたもののその評価までは，至っていない．また，症例の提示は，頭頚部領域に限定され，我々の研究課題に共通して，重症度の高い症例の収集が困難であり，写真の品質が課題とされている．

本研究は，放射線皮膚炎の評価統一という目的に変わりはなく，そのアプローチとして DCNN を用いた．これまで，放射線皮膚炎のグレード判定に DCNN を用いた研究は，報告されておらず DCNN 特有の課題も生じた．先行研究と本研究について表 5-1 に示す．先行研究のアプローチはそれぞれ，学習ソフトは評価者の教育が可能であり，グレーディングアトラスは，基準写真と評価媒体が同じであるため，評価レベルの向上に有効であると考えられる．しかし，人間が行う評価に変わりはなく，主観的評価のままである．これに対して，我々の作成したシステムは，客観的評価であるため個人差のない評価が可能である．

表 5-1 先行研究と本研究の比較

	先行研究		本研究	
	Zendaら[15]	大薗ら[14]	Hyb-RDGS	EfficientNetモデル
アプローチ	視覚的基準の作成	評価者（人間）の教育・学習	機械学習	機械学習
特徴	CTCAEに基づく放射線皮膚炎アトラスによる視覚的な基準写真による評価統一を行う.	評価者の教育、学習することで評価レベルの向上を行う.	DCNNを用いて，放射線皮膚炎画像の特徴を学習し，評価を行う.	DCNNを用いて，放射線皮膚炎画像の特徴を学習し，評価を行う.
分類クラス数	3クラス	3クラス	4クラス	4クラス
精度	評価なし	89.9%	85.1%	86.4%
分類の対象グレード（対象部位）	グレード1 グレード2 グレード3 （頭頸部）	グレード1 グレード2 グレード3 （全部位）	グレード1 グレード2 グレード3 グレード4 （全部位）	グレード1 グレード2 グレード3 グレード4 （全部位）
評価手法	作成のみであり，評価は主観的のまま.	正答率 （合格ライン90%に定めている）	正答率・混同行列 （ヒートマップ）	正答率・混同行列 （複数（グレード）判定画像に対して，ベイズ推定に基づく最終評価）
データ数	1600枚からグレーディングアトラス選定委員会による選定 38枚を選択	医師の選定 100枚	医師，看護師の選定 647枚	医師，看護師の選定 649枚
システム	放射線皮膚炎グレーディングアトラス	クイズ形式学習ソフト	VGG16 fine-tuning	EfficientNetを用いた放射線皮膚炎グレード判定システム
モデル評価	他施設共同研究であるため，施設が異なっていても評価統一が可能.	評価の統一手法として有用なツール. 学習を重ねるにつれ、丸暗記するため新しい画像を積極的に入れ替える必要がある.	DCNNを用いて，放射線皮膚炎画像の特徴を学習し，評価を行う. 極少数症例に対して人工症例画像生成し，ハイブリッド生成によるモデル構築	DCNNを用いて，放射線皮膚炎画像の特徴を学習し，評価を行う. Data augmentationを用いた学習モデル構築.
重症度の評価	評価なし（1枚のみ） 重症度写真のアトラスを含めることで可能	評価なし 学習ソフトに含めることで可能	少ない重症度症例に対して人工症例画像を生成して識別	データ拡張，データセット工夫により識別
課題	重症度の高いグレード症例が少ない. 頭頸部に限定しているため，他部位の症例の作成が必要. 写真の質が一様でないため，写真の品質プロトコルが必要.	継続した学習が必要 月一回程度の定期的学習が望まれる. 画像の入れ替えが必要.	単施設のみで作成されたモデル. 人工症例画像を生成しなければならない. 複数（グレード）判定に対する評価を除外.	単施設のみで作成されたモデル. 内部特徴を検証する必要がある.

5.1.2　DCNN を用いた病変識別の先行研究と比較

第 4 章では，本研究で作成した 2 つのグレード判定システムについて性能比較を示し，考察した．第 2 章で述べたように本研究の放射線皮膚炎の評価は，病理学的検査と異なり，視覚的評価に基づくものである．皮膚腫瘍の DCNN を用いた先行研究と本質的な比較とならない点があるが，ここでは，DCNN を用いた分類問題における性能や課題について，先行研究と比較し本研究で作成したシステムを総括する. 表 5-2 に本研究で作成した 2 つのモデルと先行研究の比較をまとめる.

Esteva ら [33]は，より早期に悪性腫瘍を検出できる DCNN を用いたシステムを構築した．3 種類の皮膚腫瘍（良性，悪性，非腫瘍性病変）の良悪性を識別する DCNN を作成し，良悪性識別の正答率の平均が 72.1%，9 つの疾患分類は 2 人の皮膚科医が 53.3%と 55.0%であったのに対して，DCNN は 55.4%であったと報告している．しかし，写真画像のズーム，角度，照明などの要因にばらつきがあるため，ビックデータを使用したと述べられ，完全な一致精度にはさらなる研究が必要であると報告している.

Fujisawa ら [34]は，Esteva らの学習枚数よりも大幅に少ない約 6,000 枚の臨床皮膚画像を用いた DCNN を構築したと報告している．皮膚腫瘍の良悪性を判定するシステムを構築し，学習されたシステムの 14 種類の皮膚腫瘍の良悪性識別の正答率は，日本皮膚科認定皮膚科医専門医 13 名と比較したところ，皮膚科専門医が 85.3%であったのに対して，DCNN の識別率は 92.4%，識別の難易度が高い 14 種類の皮膚腫瘍分類においては，皮膚科専門医が 59.7%であったのに対して，DCNN の識別率は 76.5%であったと報告している． しかし，患者選択バイアスが考えられたと述べられ，皮膚腫瘍と関連する状態の画像のみを選択しているため，炎症性の画像は除かれていること，ポロマおよび脂漏性角化症は正しく分類できていないと課題を挙げている.

それぞれの研究結果は，良悪性の識別については高い識別率（72.1%，92.4%）

であることを報告しているが，病変腫瘍分類については，55.4%，76.5%であったと報告されている．また，Fujisawa らの研究では，全体的な精度は 76.5%であったが良性病変の半分以上が正確に分類されていないと報告しており，最も高い精度は 95.7%（悪性上皮腫），最も低い精度は，62.8%（良性上皮腫）であった．これに対して，我々の作成した EfficientNet モデルは，全体的な正答率 86.4%であった．正答率の低いグレード 4 の症例に対して，人工症例画像やモデル構成を検討した結果，全てのグレードにおいて，*sensitivity*，*precision* および *F1-measure* を指標にした場合，グレード間の判定精度に特異的な性能低下は認められなかった．このことから，放射線皮膚炎のグレード判定に DCNN を応用でき，かつ先行研究を上回る性能であると考えられる．また，分類クラスは，Esteva ら，Fujisawa らのそれぞれ，9 クラス，14 クラスに対して，本研究は 4 クラスである．DCNN を用いた分類問題として考えた場合，本研究で作成したモデルのクラス数は少ないものの，先行研究の分類性能を上回る精度を達成した．本研究で作成した放射線皮膚炎グレード判定システムが 86.4%の正答率を備えていると考えると，前向き臨床研究における一般的なスクリーニング行為目的に使用できる現在のシステムであると考えられる．

表 5-2 先行研究と本研究の比較表

	先行研究		本研究	
	Fujisawaら[34]	Estivaら[33]	Hyb-RDGS[49]	EfficientNetモデル
対象	皮膚腫瘍		放射線皮膚炎	
分類種類	1.Actinnic keratosis 2.Bowen disease 3.Squamous cell carcinoma 4.Basal carcinoma 5.Melanoma 6.Poroma 7.Sebaceous naevus 8.Seborrhoeic keratosis 9.Blue naevus 10.Congenital melanocytic nevus 11.Naevus cell naevus 12.Spitz naevus 13.Lentigo simplex 14.Naevus Spilus	1. Cutaneous lymphoma and lymphoid infiltrates 2. Benign dermal tumors,cysts,sinuses 3. Malignant dermal tumor 4. Benign epidermal tumors,hamartomas,milia, and growths 5. Malignant and premalignant epidermal tumors 6. Genodermatoses and supernumerary growths 7. lnflammatory conditions 8. Benign melanocytic lesions 9. Malignant Melanoma	グレード1 グレード2 グレード3 グレード4	グレード1 グレード2 グレード3 グレード4
クラス数	14クラス	9クラス	4クラス	4クラス
正解ラベル	病理学的検査	病理学的検査	医師・看護師の評価	医師・看護師の評価
特徴	比較的少数の画像（<5000）で学習したDCNNで皮膚科医の分類を上回る精度を達成した.	皮膚科医よりも優れた性能であった. 3,374ダーモスコピー画像を含む129,450臨床画像の新しい皮膚科医ラベル付きデータセットによる一般化可能な分類を示した.	重度のグレード4に対して人工症例画像を生成し，データセットに用いることで精度の高い分類器を作成した.	Hyb-RDGSと同等の性能を持つ効率的なシステムであることを示した. 曖昧な画像に対して，アンサンブル学習に基づくベイズ推定量により，評価者の判定に，より近い判定を可能とした.
評価手法	感度・特異度 5分割交差検証	感度・特異度・AUC 9分割交差検証	正答率・混同行列 3分割交差検証	正答率・混同行列 3分割交差検証
学習データ	4,867枚	使用データ129,450枚, 学習データ不明	647枚	649枚
システム	Google-Net DCNN	GoogleNet Inception v3 CNNアーキテクチャ	DAと人工症例画像を混合したハイブリッド生成法（Hyb-RDGS）	EfficientNetを用いた放射線皮膚炎グレード判定システム
精度	76.5%	55.4%	85.1%	86.4%
課題	患者選択バイアスが考えられた. 皮膚腫瘍と関連する状態の画像のみを選択しているため，炎症性の画像は除かれている. エラーの傾向は，ポロマおよび脂漏性角化症のは正しく分類できていない.	写真画像のズーム、角度、照明などの要因にばらつきがあるため、ビックデータを使用した. 完全な一致精度にはさらなる研究が必要.	人工症例画像生成の作業が必要なため，効率性に欠ける. 効率性を向上させる必要がある.	単施設のみで作成されたモデル. 内部特徴を検証する必要がある. 重症度の症例収集が必要.

5.2　本研究で開発した放射線皮膚炎グレード判定システムの位置付けと今後の課題

本研究は，DCNN を用いたシステム作成において 4 つの課題に取り組み，その成果を検証した．それぞれの課題に対して，Hyb-RDGS と EfficientNet モデルを作成した．これらのシステムは，先行研究を上回る性能を示したものの，全ての放射線皮膚炎画像に対して万能というわけではない．ここでは，本研究の現時点における 2 つの課題を述べ，本研究で開発した放射線皮膚炎グレード判定システムの現時点の位置付けを述べる．

1 つ目は，不鮮明な画像に対しては，誤判定が生じてしまう可能性が挙げられる．評価に使用する画像は，カメラ等で撮影した写真であるため，どうしても撮影の仕方による画像のボケや歪といった画質に左右されてしまう．不鮮明な画像のグレード判定については，本研究では評価していない． 正しいグレード判定を行うには，より鮮明な画像でなければ評価することが不可能であり，誤判定を招くことも考えられる．作成したグレード判定システムにおいて，評価を可能とするために放射線皮膚炎写真取得の統一したプロトコルが必要である．また，判定が不可能な画像における誤判定を防ぐためにシステムが判定不能であると示すことも対策として考えられる．誤判定を最小限に抑える手法に敵対的特徴学習という DCNN に間違った予測をさせる敵対的サンプルを加える手法が研究されており近年，注目されている[60,61]．つまり，DCNN は，用意したデータの中では，高精度を出したとしても，微小な摂動を加えただけで誤認識してしまう場合がある．そこで，敵対的サンプルにより，分類エラーなどの誤った出力を引き起こし，認識させないようにする一種の攻撃を行うのである．本研究で作成したシステムにおいて，誤判定となるような画像は除外しているが，敵対的特徴学習を加えることで検閲機能を備えたグレード判定システムとなり，判定性能を向上が期待できると考えられる．

2 つ目は，評価統一を行う目的に作成した本システムは，メディポリス国際陽

子線治療センター単施設のみで作成しているため，他施設では評価基準が異なる可能性が挙げられる．Zenda ら [15] が作成したアトラスのように他施設で放射線皮膚炎症例を収集する方法も挙げられるが，本研究の課題 1 で述べたように極めて困難である．また，4.9.3 項，表 4-23 で示した複数（グレード）判定は，単施設であっても職種により評価結果に相違が生じる知見を得た．本研究の評価者は，医師 2 名と看護師 4 名であり，評価結果に我々の施設の評価者の割合で行ったが，施設によっては医師だけが全て評価しているかもしれない．医師の評価が低くなる傾向があるとすれば，評価者の割合が評価結果に影響する可能性も考えられ，システムの汎用性能低下の要因となる．本研究で得られた知見は，本システムで評価する場合，特に複数（グレード）判定画像を扱う場合に一つのシステム特性として認識しておくべきである．

本システムを多くの施設で使用するためには，データ収集だけでなく，本研究で収集した放射線皮膚炎のグレード付けを行ったようにアノテーションを繰り返す必要がある．そのプロセスには，アノテーションから，学習，評価に至る多大な労力を伴う．本研究では，EfficientNet モデルにより効率的な学習性能向上を図り，現時点で有用なモデルといえる．

また，臨床では画像だけでなく，所見や治療計画など多角的な情報の基で評価し，ケアが行われるため患者の症状に応じて医師の評価が最優先である．あくまで医師や看護師の放射線皮膚炎の評価を助言するシステムである．本システムは，放射線皮膚炎画像のみで評価したシステムであり，評価統一目的や複数（グレード）判定においてその性能が有効なサポートを実現する．本システムの現時点の成果として，セカンドオピニオンやサードオピニオンとなるシステム性能であることを実証した．

第6章　結論

本研究では，はじめに DCNN を用いた放射線皮膚炎のグレード判定システムについて，ポアソン合成を用いたハイブリッド生成法によるデータセットの工夫により，Hyb-RDGS を作成した．これにより，DCNN のグレード判定実現の可能性を確認し，さらに効率化と高精度な判定性能を持つ EfficientNet モデルを用いたシステム開発の研究結果をまとめたものである．

第 1 章では，放射線治療の有害事象である放射線皮膚炎のグレード判定の現状について述べた．人間の判定で生じる曖昧な基準や個人差によって生じる不確定なグレード判定を補助することを目的に，2 つの先行研究とその課題について述べた．

第 2 章では，近年，医用画像に応用されるようになった DCNN を用いた画像識別の先行研究について述べた．放射線皮膚炎のグレード判定に DCNN を用いたシステム開発の研究の必要性と DCNN を用いたシステムの実現に 4 つの課題を提示した．

課題 1．放射線皮膚炎グレード 1～4 における症例画像の収集と画像の品質

課題 2．不均衡なデータ数と少数画像の取り扱い

課題 3．稀な症例の取り扱い（極少数画像の取り扱い）

課題 4．放射線皮膚炎のグレード判定の相違

第 3 章では，課題 1～3 に対して，放射線皮膚炎画像選定プロトコルを策定し，DCNN の学習画像の品質を重要視した（課題 1）．さらに課題 2 に対して，収集したデータ数の不均衡の解消のため，アンダーサンプリング，データ拡張法によるオーバーサンプリングを実施した．課題 3 に対しては，ポアソン合成を用いた人工症例画像を生成し，データ拡張を実施した．このようにサンプリングレートを変えながら工夫をすることで，3 クラス（グレード 1～3）においては，DA の有無によって正答率は 86.6%と 76.0%となり，DA によって正答率は 10.6%向上した．4 クラス（グレード 1～4）においては，正答率はそれぞれ 83.4%，74.4%

となり，正答率は 9.0%向上した．DA 法によるグレード判定の正答率の向上を実証し，不均衡データの緩和効果を得た．さらに DA 法と人工症例画像を混合したハイブリッド生成法によって作成した Hyb-RDGS の正答率（4 クラス）は，DA 法が 83.4%であったのに対し， 85.1%を達成した．混同行列を用いた性能評価は，グレード 4 の判定では，感度 93.3%，適合率 84.7%, F1 値 88.5%であり，DA 法と比較して，Hyb-RDGS の性能は，それぞれ 15.5%，8.5%，11.6%向上した．これより，ハイブリッド生成法が有効であることを示した．

第 4 章では，課題 4 に対して複数の重み付けの異なるモデルにベイズ推定を適用して，最終グレード判定を出力するシステム開発を行った．さらに，Hyb-RDGS の精度を踏まえた上で効率よく学習可能な EfficientNet モデルの適応について，精度評価を行った結果について述べた．EfficientNet モデルは，従来のニューラルネットワークで行う手動チューニングを固定化する複合スケーリング法を特徴とする効率的なモデルである．現時点で，最も効率的な DCNN モデルであるが，そのモデルチューニングを研究した報告は少ない．さらにデータ拡張においても EfficientNet モデルで実績のある RA 法を採用し，モデル構成を検討し，効率性の高いシステム実現を目指した．はじめにモデル構成について検討し，データ拡張手法の比較により，最適なモデル作成検証を実施した．その結果，EfficientNet-B4，384×384 画素（データセット：DA）のモデル構成において，正答率は，86.4%を達成した．また，課題であったグレード 4 の検出については，は，*sensitivity*（$p<0.05$）， *F1-measure* （n. s.）は，低下したが，*precision* においては Hyb-RDGS を有意に上回る性能であった（$p<0.05$）．これより，第 3 章で作成した Hyb-RDGS を同等かつ正答率については，上回る性能を持つ効率的な DCNN を用いた放射線皮膚炎のグレード判定システムの可能性を実証した．課題 4 に対して，アンサンブル学習に基づく複数の高精度 EfficientNet モデルを選定し，特に複数（グレード）判定画像については，ベイズ推定により最終判定結果を提案できることを述べた．評価者とシステムの評価結果の関係について，分

析し，評価者の半数以上において 95%で一致する結果を得た．また，評価結果より，評価者の判定傾向が新たな知見として得られた．

第 5 章では，放射線皮膚炎の評価統一目的および類似する DCNN を用いた病変識別の先行研究と比較し，本研究の今後の課題について述べた．放射線皮膚炎の評価統一目的における研究は，それぞれのアプローチは異なるが，先行研究は人間が行う評価に変わりはなく，主観的評価のままである．これに対して，我々の作成した放射線皮膚炎グレーディングアトラスは，客観的評価であるため個人差のない評価が可能であることを示した．また，DCNN を用いた先行研究に対して，分類クラスは少ないものの先行研究を上回る正答率であったことを示し，前向き臨床試験における放射線皮膚炎のスクリーニング行為目的に使用できるシステムに必要な精度であることを述べた．

本研究では，放射線皮膚炎の評価について DCNN を用いたシステムを開発した．研究課程において，DCNN の学習データ拡張手法や効率的な EfficientNet モデル構成を検証し，放射線皮膚炎の評価目的のみならず様々な病変診断にも応用可能なシステムとして貢献できる．例えば，症例数の少ない病変の識別に DCNN で実現する場合，課題とされた少数データ，効率化は必須であり，本研究の人工症例の作成，モデル開発プロセスにおける効率的な EfficientNet モデルの貢献が期待できる．

本研究で開発した EfficientNet モデルを用いた放射線皮膚炎グレード判定システムは，評価者の判定を補助するセカンドオピニオン，サードオピニオンとなる可能性となることを示した．

文献

1) 日本放射線腫瘍学会構造調査 2005-2015 年調査結果
2) 吉田 浩二, 宮地 麻美, 鍜治 朋子, 朝長 さつき, 伊藤 陽子, 川久 保真弓, 中島 香菜美, 佐藤 良信, “放射線治療を受けた咽頭がん患者の有害事象評価－放射線皮膚炎を中心に－”, 日本放射線看護学会誌, Vol.2, No.8(1) , pp.12-18 (2014)
3) 中川 恵一, 青木 幸昌, “放射線治療ガイドブック”, 医療科学社, p.121 (1999.12)
4) 菱川 良夫, 藤本 美生, ”放射線治療を受けるがん患者の看護ケア”, 日本看護協会, pp.133, 152-153 (2008.6)
5) 紺屋 隆一, 中馬 育子, 有村 健, 川畑 道子, 荻野 尚, 菱川 良夫, “放射線皮膚炎に対する適切な看護についての検討”, 日本放射線腫瘍学会第 25 回学術大会報文集, p.227 (2012)
6) Chan Raymond, Javan Mann, Jennifer Tripcony, Lee Keller, Jacqui Cheuk, Robyn Blades, Rae Keogh, Samantha Poole, Christopher Walsh and Christopher, “Natural oil-based emulsion containing allantoin versus aqueous cream for managing radiation-induced skin reactions in patients with cancer: a phase 3, double-blind, randomized, controlled trial”, Int J Radiat Oncol Biol Phys,Vol.90(4), pp.756-64, （2014.11）
7) Jonathan Leventhal, Melissa Rasar Young, “Radiation Dermatitis: Recognition, Prevention, and Management”, Int J Radiat Oncol Biol Phys,Vol.31(12), pp.885-887, 894-899, (2017.12)
8) National Cancer Institute CTEP, ”Common Terminology Criteria for Adverse Events version 4.0”, MedDRA v12.0 Code 10037767, (2010)
9) 丹生 健一, 佐々木 良平, ”目で見て学ぶ放射線療法の有害反応”, 日本看護協会出版会, p.57, (2011.2)
10) 新井 達, “放射線皮膚炎に対する新たな治療展開～保湿剤の有用性について～”, 日本臨床皮膚科医会誌,Vol.37(1), pp.62-66, (2020)
11) Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura and Ronald M Summers, “Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning”, IEEE Trans Med Imaging, Vol.35(5), pp.1258-1298, (2016.2)

12) Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clarea I. Sanchez and Bram van Ginneken, ”Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks”, IEEE Trans Med Imaging, Vol.35(5), pp.1160-1169, (2016.3)
13) Holger R. Roth, Le Lu, Jiamin Liu, Jianhua Yao, Ari Seff, Kevin Cherry, Lauren Kim, and Ronald M. Summers, “Improving computer-aided detection using convolutional neural networks and random view aggregation”, IEEE Trans Med Imaging , Vol.35(5), pp.1170-1181, (2016.5)
14) 大薗 香織, 有村 健, 松山 貢, 荻野 尚, 中馬 育子, 菱川 良夫, “放射線皮膚炎に対する評価基準の統一と評価レベルの向上について”, 日本放射線腫瘍学会第 25 回学術大会報文集, p153,(2012)
15) Sadamoto Zenda, Yosuke Ota, Hiroyuki Tachibana, Hirofumi Ogawa, Shinobu Ishii, Chikako Hashiguchi, Tetsuo Akimoto, Yuichiro Ohe and Yosuke Uchitomi, “A prospective picture collection study fora grading atlas of radiation dermatitis for clinical trials in head-and-neck cancer patients”, Journal of Radiation Research, Vol.57(3), pp.301-306, (2016.2)
16) Patric Perez, Michel Gragnet, and Andrew Blake, ”Poisson Image Editing”, ACM Transactions on Graphics (SIGGRAPH'03); Vol.22(3), pp.313-318, (2003)
17) Karen Simonyan, Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition”, The IEEE international Conference on Computer Vision (ICCV), arXiv: 1409. 1556v6 [cs.CV], (2015.4)
18) Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcrofe, Kilian Q Weinberger, “Snapshot ensembles”, Machine Learnng, arXiv, 1704.00109[cs.LG] , (2017.4)
19) 上田 修功, ” アンサンブル学習−識別器の識別性能向上法および情報統合の数理− “, 社団法人情報処理学会, CVIM, Vol.145 (26) , (2004)
20) Tan, Mingxing, Quoc V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”, Machine Learnng, arXiv, 1905.11946, (2020.9)
21) Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, Quoc V. Le, “RandAugment”, Practical automated data augmentation with a reduced search space; Computer Vision and Pattern Recognition, arXiv: 1909. 13719v2 [cs.CV] , (2019.9)
22) 岡谷 貴之,”ディープラーニング”, 映像情報メディア学会誌. Vol.68 (6), pp.466-471, (2014)

23) Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, Computer Vision and Pattern Recognition, arXiv:1512.03385 [cs.CV] , (2015.11)
24) Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation” , Computer Vision and Pattern Recognition, arXiv:1311.2524 [cs.CV] , (2014.10)
25) Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “You Only Look Once”, Computer Vision and Pattern Recognition, arXiv:1506.02640 [cs.CV] , (2016.5)
26) 村上 佳菜子, 橋本 典明, 木戸 尚治, 平野 靖, 間普 信吾, 近藤 堅司, 小澤 順, “CNN, FCN, U-Net を用いたびまん性肺疾患の領域抽出の比較”, The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, (2018)
27) 周 向栄, 藤田 広志, “深層学習に基づく CT 画像からの複数の解剖学的構造の同時自動認識と抽出”, MEDICAL IMAGING TECHNOLOGY, Vol.35(4), (2017.9)
28) Olaf Ronneberger, Philipp Fischer, Thomas Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”, Computer Vision and Pattern Recognition, arXiv:1505.04597 [cs.CV], (2015.5)
29) Xuerong Xiao, Swetava Ganguli, Vipul Pandey, “VAE-Info-cGAN: Generating Synthetic Images by Combining Pixel-level and Feature-level Geospatial Conditional Inputs”, Computer Vision and Pattern Recognition, arXiv: 2012. 04196v1 [cs.CV] , (2020.12)
30) Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger and Hayit Greenspan,”GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification”,Computer Vision and Pattern Recognition, arXiv: 1803. 01229v1 [cs.CV] , (2018.3)
31) Howard Dermatol, Martin A Weinstock, Steven R Feldman, Brett M Coldiron, ”Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population,2012”, JAMA Dermatology 151(10), pp.1081-1086 , (2012.10)
32) 福田 英嗣, 向井 秀樹, ”色素性皮膚病変以外の疾患に対するダーモスコピーの活用法”,日本臨床皮膚科医誌, 32(2), pp.181-186, (2015)
33) Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M.Swetter, Helen M. Blau and Sebastian Thrun, ”Dermatologist-level classification of skin cancer with deep neural networks” ,Nature, 542, pp.115-118, (2017.1)

34) Y. Fujisawa, Y. Otomo, Y. Ogata, Y. Nakamura, R. Fujita, Y. Ishitsuka, R.Watanabe, N. Okiyama, K. Ohara and M. Fujimoto, “Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis” ,British Journal of Dermatology,180, pp.373-381, (2019.2)
35) Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, Quoc V. Le, “AutoAugmnet: Learning augmentation strategies from data” Computer Vision and Pattern Recognition, arXiv: 1805. 09501 [cs.CV] , (2019.4)
36) Laurens van der Maaten, Geoggrey Hinton, “Visualizing data using t-SNE”, Jounal of Machine Learning Research 9, pp.2579-2605, (2008)
37) Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das and Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization” , Computer Vision and Pattern Recognition, arXiv: 1610. 02391 [cs.CV] , (2019.12)
38) Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva and Antonio Torralba, “Learning Deep Features for Discriminative Localization” ,Computer Vision and Pattern Recognition arXiv: 1512. 04150 (2015.12)
39) Takeshi Arimura, Takashi Ogino, Takashi Yoshiura, Naoaki Kondo, Shinichi Nagayama and Yoshio Hishikawa, “Effect of film dressing on acute radiation dermatitis secondary to proton beam therapy” ,Radiation Oncology Biology Physics, Vol.95(1), pp.472-476, (2015.10)
40) Ellen Trueman, “Managing radiotherapy-induced skin reactions in the community”, International Journal of Palliative Nursing, Vol.27(4): pp.16-24, (2013)
41) Samira Pouyanfar, Yudong Tao, Anup Mohan, Haiman Tian, Ahmed S. Kaseb, Kent Gausen, Ryan Dailey, Sarah Aghajanzadeh, Yung-Hsiang Lu, Shu-Ching Chen and Mei-Ling Shyu, “Dynamic Sampling in Convolutional Neural Networks for Imbalanced Data Classification” , IEEE Conference on Multimedia Information Processing and Retrieval, pp.112-117, (2018.4)
42) Luis Torgo, Paula Branco, Rita P. Ribeiro, and Bernhard Pfahringer, "Resampling strategies for regression” , The Jounal of Knowledge Engineering, Vol.32(3),pp.465-476, (2015.6)
43) Ryuhei Hamaguchi, Ken Sakurada, Ryosuke Nakamura, “Rare Event Detection using Disentangled Representation Learning” ,Computer Vision and Pattern Recognition, arXiv:1812.01285v1 [cs.CV], (2018,12)

44) Sebastian Raschka, Vahid Mirjalili, “Python Machine Learning”, Second Edition, Chapter 1. Packt Publishing. (2017)
45) 藤田 広志，福岡 大輔, “医用画像のためのディープラーニング-入門編-“ ,オーム社, p146, (2020.4)
46) Takaya Saito, Marc Rehmsmeier, “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets” , PLOS ONE Vol.10(3), (2015.3)
47) Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton, “ImageNet classification with deep convolutional neural networks” ,Neural Information Processing Systems, Vol.60, pp.84-90, (2017.6)
48) 和田 清隆，渡邊 睦，新地 真大，野口 康介，向吉 登貴恵，松山 貢，有村 健，荻野 尚, “ハイブリッド生成法によるディープラーニングを用いた放射線皮膚炎グレード判定システムに関する研究” ,日本放射線技術学会誌, Vol.77(8), (2021.8)
49) 和田 清隆，渡邊 睦，新野 将史，野口 康介，荻野 尚, “EfficientNet を用いたベイズ推定に基づく放射線皮膚炎グレード判定手法の開発”（in press）,日本医用画像工学会誌 Vo.40(2), (2022)
50) Sheela Ramanna, Cenker Sengoz, Scott Kehler and Dat Pham, “Near real-time map building with labelling and classification of road conditions using convolutional neiral networks” ,Computer Vision and Pattern Recognition, arXiv:2001.09947v1 [cs.CV] , (2020.1)
51) Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition” ,Computer Vision and Pattern Recognition, arXiv: 1512.03385[cs.CV], (2015.12)
52) Jerome H. Friedman, “On bias, variance, 0/1-loss, and the curse of dimensionality” , Data Mining and Knowledge Discovery, Vol.1(1), pp.55-77, (1997.3)
53) 杉山 将，鈴木 大慈, “機械学習のための数学” ,情報処理, Vol.56(5), pp.10-15, (2015)
54) 手良 向聡, “希少がんに対する臨床試験デザイン, “The Japanese Journal of Pediatric Hematology/Oncology, Vol.56(5) , pp.425-428, (2019)
55) 佐々木 春喜, “診断推論と確率-ベットサイドでのベイズ定理-“ ,日本プライマリ・ケア連合学会誌, Vol.36(3) , pp.191-197, (2013)
56) Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang,

Tobias Weyand, Marco Andreetto and Hartwig Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications” , Computer Vision and Pattern Recognition, arXiv:1704.04861[cs.CV], (2017.4)

57) Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov and Liang-Chieh Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks” ,Computer Vision and Pattern Recognition, arXiv: 1801.04381[cs.CV], pp.4510-4520, (2018.1)

58) Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V and Le, Hartwig Adam, “Searching for mobilenetv3” , Computer Vision and Pattern Recognition, arXiv:1905.02244[cs.CV], (2019.5)

59) Parvathaneni Naga Srinivasu, Jalluri Gnana SivaSai, Muhammad Fazal Ijaz, Akash Kumar Bhoi, Wonjoon Kim and James Jin Kang, “Classification of skin disease using deep learning neural networks with mobilenet V2 and LSTM” ,Sensors, Vol.21(8): 2852, (2021.4)

60) Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry and Dimitris Tsipras, “Adversarial Examples Are Not Bugs” , Computer Vision and Pattern Recognition, arXiv:1905.02175v4, (2019.8)

61) 岩澤 有祐, 松尾 豊, “類似度学習を用いた敵対的訓練による特徴表現の検閲” ,The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, (2018)

謝辞

本研究の遂行および本論文の作成は，鹿児島大学大学院理工学研究科 渡邊睦教授のご指導の下に遂行されたものであり，終始懇切丁寧なるご指導とご鞭撻を受け賜りました．ここに深く感謝の意を表します．

鹿児島大学大学院理工学研究科 王 鋼教授，朱碧蘭准教授，渕田孝康准教授，鹿嶋雅之准教授，福元伸也助教には，本研究の遂行と論文のとりまとめにあたり，有益なご討論と貴重なご助言を賜りました．ここに熱く御礼申し上げます．

また，放射線治療の基礎から研究開発業務にいたるまで懇切丁寧にご指導いただき，本研究を遂行するための基礎を与えて下さいましたメディポリス国際陽子線治療センター 荻野尚医師，有村健医師，向吉登貴恵看護師に深く感謝いたします．また，社会人として博士後期課程に進学し研究を行う機会を与えて下さいました一般社団法人メディポリス医学研究所 永田良一理事長，福﨑好一郎副理事長に厚く御礼申し上げます．そして，日頃からお世話になったメディポリス国際陽子線治療センター 技術部と看護部の関係各位に感謝申し上げます．

また，博士後期課程の専攻ゼミナールや本研究にご協力をいただいた鹿児島大学大学院 理工学研究科 人物処理グループの新地真大氏，野口康介氏，新野将史氏ならびに研究室の皆様に深く感謝いたします．

最後に，博士後期課程に進学し研究を行う生活を支えてくれた，家族に深く感謝いたします.最後まで私のために本当にありがとうございました．