修士論文　　2019 年度

内視鏡画像からの
簡易カルテ自動生成システム

内藤　慎一郎
(学籍番号：81821276)

指導教員　教授　萩原 将文

2020 年 3 月

慶應義塾大学大学院理工学研究科
開放環境科学専攻

論文要旨

本論文では深層学習モデルを用いて内視鏡画像からの簡易カルテ作成を自動化する医師の支援システムを提案する。従来の研究では画像単位での学習と予測を行っており、患者単位での予測が不可能であった。本研究では、内視鏡検査における複数の画像を患者単位に学習と予測を行えるように、データセットとモデルを作成した。また従来の研究では一種類の病変のみの予測が可能なシステムであった。本研究では患者に病変が存在するかの判定や複数の病変情報をまとめたマルチラベルデータを作成した。これらの改善によって、患者単位での複数の病変の予測が可能なシステムとなった。さらに既存研究では全ての画像に手動でラベルデータを付与していたのに対し、本研究では、ラベルデータを医師が実際に作成した簡易カルテから自動で生成した。この学習データを用いて二次元と三次元の畳み込みニューラルネットワークをそれぞれ複数モデルずつ訓練し比較した。そしていくつかのモデルで、患者に病変が存在するかの判定で F1-Score が 98%以上、複数の病変情報の予測で F1-Score が 54%以上と、高い推論性能を記録した。また推論時の予測ラベルの生成において、複数の信頼度を持つ予測を生成した。複数の病変情報の予測において、信頼度の高い予測では適合率が 77%以上、信頼度が低い予測では再現率 85%以上を記録した。これによって本研究におけるシステムは医師の判断の支援を行うものとしても利用できるようになった。

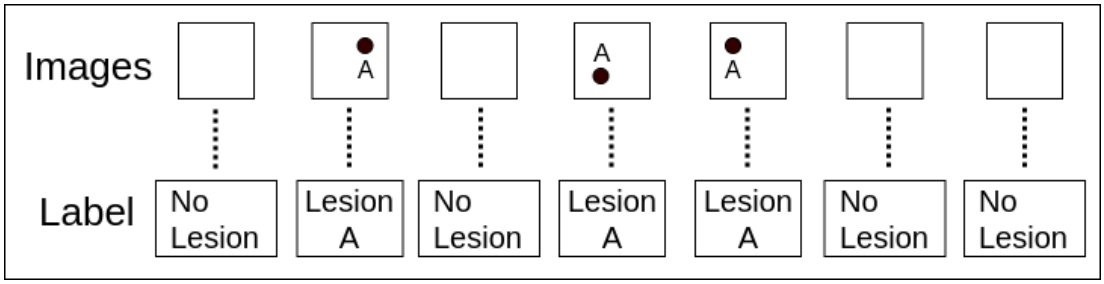
Thesis Abstract

In this paper, we devised an automatic creation support system for a draft of medical charts from endoscopic images using a deep learning model. In previous studies, learning and prediction were performed on an image basis, and prediction on a patient basis was impossible. In this study, we created datasets and models so that multiple images in endoscopy could be learned and predicted on a patient basis. The system in the previous study could predict only one type of lesion. In this study, label data consisting of a label indicating whether one or more lesions exist and a plurality of lesion labels was used. These improvements made it possible to predict multiple lesions on a patient basis. In addition, in the existing study, label data was manually assigned to all images, whereas in this study, multi-label data was automatically generated from a simple chart actually created by a doctor. Using this learning data, we trained and compared several models of 2D and 3D convolutional neural networks. And some models recorded high inference performance. The F1-Score was 98% or more in determining whether a patient had a lesion, and the F1-Score was 50% or more in predicting multiple lesion information. We also generate predictions with multiple reliability levels when predicting labels. In the prediction of multiple lesion information, the precision was 87% or more for highly reliable predictions, and the recall was 81% or more for lowly reliability predictions. As a result, the system in this study can be used to support the judgment of doctors.

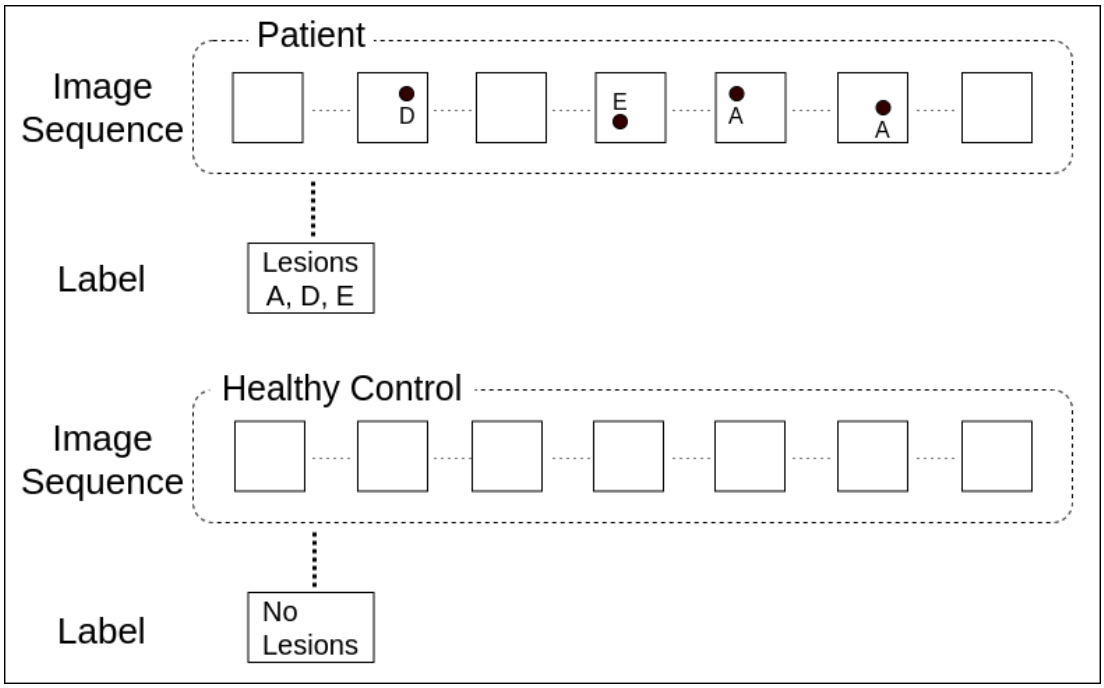
目 次

第 1章

序論



(a) 従来手法 (Lesion A のみの予測)



(b) 提案手法 (Lesion A, B, 他多数, の予測)

図 1.1: 全体概要

現在機械学習の分野において深層学習による画像処理 [1–4]、系列データ処理 [5–8] が高い精度を記録している。このため文字や画像や音声などの認識 [9,10] や生成 [11,12]、翻訳 [8,13,14] や検索 [15] といった自然言語処理に幅広く用いられるようになっている。また化合物の性質予測 [16,17] やゲノム構造の解析 [18]、株価の予測 [19] のように、生物学や化学や経済学などの情報工学分野を超えた様々な分野においても用いられてきている。

この流れの中で医学分野においても様々な用途で深層学習が用いられてきている。例としては皮膚画像による皮膚ガンメラノーマの判定 [20] や放射線画像による深層学習解析 [21] などがあげられる。

他にも医学分野において深層学習を用いた自動化によって診断の精度向上や作業の効率化が期待できる用途がある。その中でも内視鏡画像の解析が挙げられる。既存研究には、内視鏡画像による胃癌判定 [22] や食道癌 [23] の判定がある。これらの研究は内視鏡画像を一枚ずつラベル付けして学習したもので、内視鏡検査における画像のすべてを学習しているわけではない。このためより効果的な用途として内視鏡画像からの簡易カルテの生成が挙げられる (図 1.1)。現在簡易カルテの生成は人手によって行われており、各患者ごとに数分の時間がかかっている。さらに大量の 2 重チェックを行う必要がある。したがって医師の貴重な時間の多くを簡易カルテ作成に費やさなければならないという大きな問題があった。これらの作業が自動化されることによる医師の負担削減が大きく期待できる。また病変見落としが医師によっては 2 割以上といったデータ [24] もある。自動化により精度の高い予測が可能になると、病変の見落としなども減らすことが可能になる。

しかしながら深層学習を医学分野に用いることに関していくつかの問題がある。一つ目は倫理的な問題である。医療における診断という大きな責任を伴う判断において、深層学習のような機械学習モデルを用いることには責任の所在がなくなるという課題がある。現在の深層学習モデルはブラックボックスであり、判断の根拠が不明となっている。このため医療分野における責任ある判断をまかせることにはまだ至っていない。二つ目はプライバシーの問題である。患者の診断データや内視鏡画像は重要度の非常に高いプライバシーデータであ

り、実際の推論システムにおいて患者のデータが特定可能な形で出力されてしまうことを必ず避けなければならない。

本研究では簡易カルテという精密検査や診断の確定がなされる前の作業における支援システムとして導入するため、最終的な判断の責任は医師にある。また学習データの作成では、簡易カルテから学習データのマルチラベルを作成するに当たって、質的診断の項目を自動で収集してマルチラベル化するため、プライバシー情報が完全に学習データに含まれない。学習においても、内視鏡画像とマルチラベルデータのみを深層学習モデルに通した。

既存研究では各患者に対して数十枚ある内視鏡画像すべてに手動で病名情報をラベリングしている。したがって深層学習を用いるためには数十万の学習データを用意する必要があるために膨大な時間がかかっている。全ての画像においてラベリングをした教師あり学習であるために高い精度を記録しているが、推論時も各画像単位で行われるため、画像を用いた診断として用いることはできるが簡易カルテの作成の自動化には至っていない。このため本研究では医師が作成した簡易カルテに対して自然言語処理を行い、教師データとしてマルチラベルを作成した。このマルチラベルデータは二つの部分から構成されている。一番目のラベルは内視鏡画像の中に病変があるか否かを示す値が格納されており、二番目以降は頻出病名に対応した複数のラベルが並んでいる。マルチラベルデータは簡易カルテの病名情報を要約したものとして用いることができる。内視鏡画像とマルチラベルデータの関係性を直接学習させることによって、学習データの作成と学習全ての工程を自動化することに成功した。

学習手法に関しては、大きく分けて二つの手法で実験した。一つ目の手法は各患者における数十枚の画像をそれぞれ一枚ずつ入力し、出力をマルチラベルから損失を計算し、モデルを学習するものである。二つ目の手法では各患者の全ての時系列画像を三次元データとして用いてモデルを学習した。

本論文では作成した学習データを二つの手法においてそれぞれ複数の深層学習モデルで学習と推論を行い、結果を比較した。

以下、第 2 章で関連研究について述べ、第 3 章で提案手法について、第 4 章で評価実験、第 5 章で考察、第 6 章で結論を述べる。

第 2 章

関連研究

深層学習モデルに内視鏡画像を学習させた研究として、胃部分の内視鏡画像から胃癌の有無を判定したもの [22] や食道部分の内視鏡画像から食道癌の有無を判定したもの [22] がある。これらの研究では画像を一枚ずつラベル付けし、これらを Convolutional Neural Network (CNN) [1] を用いて学習した。用いたモデルは Single Shot Multibox Detector (SSD) [25] であり、これは 16 層の CNN で物体検出に特化したものである。またどちらの研究も一つの病変に特化したものである。食道癌の検知の研究では 384 人の患者から得られた 8,428 枚の画像を訓練データとし、47 人の食道がん患者と 50 人の食道がんではない患者から得られた 1,118 枚の画像をテストデータとした。この研究では食道がんの早期発見のために 10mm 以下の小さな病変を検知できるモデルの学習に特化した。その結果 95%という高い再現率を記録したが、適合率は 40%ほどとなっている。またこの研究では通常の内視鏡画像である white-light image (WLI) と内視鏡機器に搭載されている狭帯域光観察によって病変を見やすくした画像である narrow-band imaging (NBI) をそれぞれ分けたものと包括的に扱ったものの 3 つの場合で比較している。この比較ではどの場合でも再現率は高いが適合率が低くなっている。また胃癌の検知の研究も同様に適合率が低く精度があまり良くない。

これらの研究の課題としては大きく 3 つのことが挙げられる。一つ目は各画像への手動でのラベリングである。これにより学習データの準備に非常に時間がかかる上に、データ拡張に対応できない。これらのモデルは診断の補助には利用できるが医師の作業効率化の支援にはなっていない。二つ目は一つ目とも関連するが、学習データが少ないことである。現在深層学習モデルの学習には

数十万の画像を用いるのが一般的であるが、これらの研究では 1 万に満たないものであり、結果の適合率が非常に低いのはこの課題によるものでもあると考えられる。三つ目は SSD というモデルの選択である。SSD は 16 層という比較的浅い CNN モデルであるが、物体検出というタスクに特化することで高い精度を記録したものである。機械学習における物体検出とは景色のなかにある動物の検出といったタスクであり、内視鏡画像の中にあるわずかな病変を検出するような高難度なタスクとは異なっている。この研究での結果はこのモデルの選択に大きく影響を受けてしまっていると考えられる。このため用いる CNN のモデルはより深い特徴抽出ができる多層のモデルかつ、大きさの異なる複数の特徴の組み合わせを学習できる残差構造をもっているものが望ましいと考えられる。本論文においては以上の 3 つの課題を解決することを目標とする。

第 3 章

提案手法

提案手法をデータセットの作成、モデルの作成、モデルの学習、評価指標から説明する。すべての実装コードは筆者のgithubのレポジトリ (https://github.com/shinn1r0/endoscopic_images2karte) にある。実装には主に Python を用い、深層学習フレームワークにはPytorchを使用した。またプライバシー情報のためデータセット作成の元となる内視鏡画像と簡易カルテのデータ、それらから生成したデータセットは含まれていない。

3.1 手法の分類

本論文では図 3.1 のように、画像を個別にマルチラベルと関連付けで学習する①の手法と各患者ごとの全ての画像をまとめてマルチラベルと関連付けで学習する②の手法の二つの手法で実験を行った。そのため以下に続くデータセットとモデルの作成においても、それぞれ 2 種類用意した。学習手法は 2 つの手法で異なっているが、推論時はどちらの手法においても患者単位の画像を全て入力し、 1 人の患者のラベルを予測した。

3.2 データセットの作成

2 つの提案手法において、簡易カルテからのマルチラベル生成と各画像における処理は共通であるが、その後の処理はそれぞれの手法で異なる。

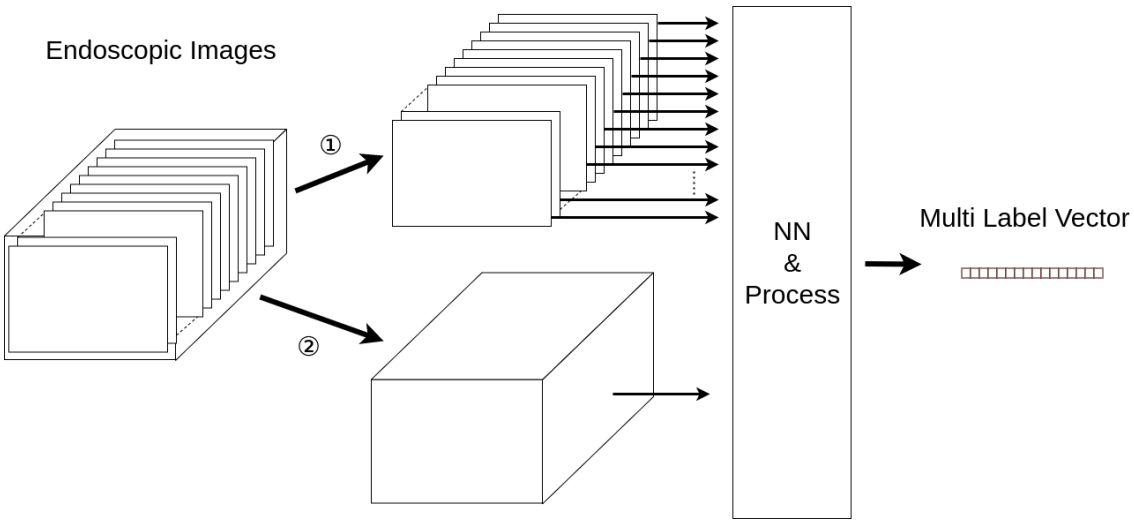


図 3.1: 提案手法

3.2.1 簡易カルテからのマルチラベル生成

簡易カルテからのマルチラベル生成を図 3.2 に示す。また手順は以下のように行った。

1. すべて異常なしの患者を抽出
2. 簡易カルテ内の質的診断を言語処理
3. 処理後の内容を全ての患者でまとめてカテゴリー化
4. 頻出病名カテゴリー選択
5. 各患者ごとにマルチラベル生成
6. マルチラベルの one-hot ベクトル化

詳細を述べる。(1) では簡易カルテ内の全ての行で異常なしとなっている患者を抜き出し、 (2) から (4) の処理には通していない。(2) では簡易カルテ内の質的診断の項目のみを取り出し、言語処理を行った。この言語処理では文書であったり単語や数値の羅列であったりする質的診断をまず単語の系列データへと変

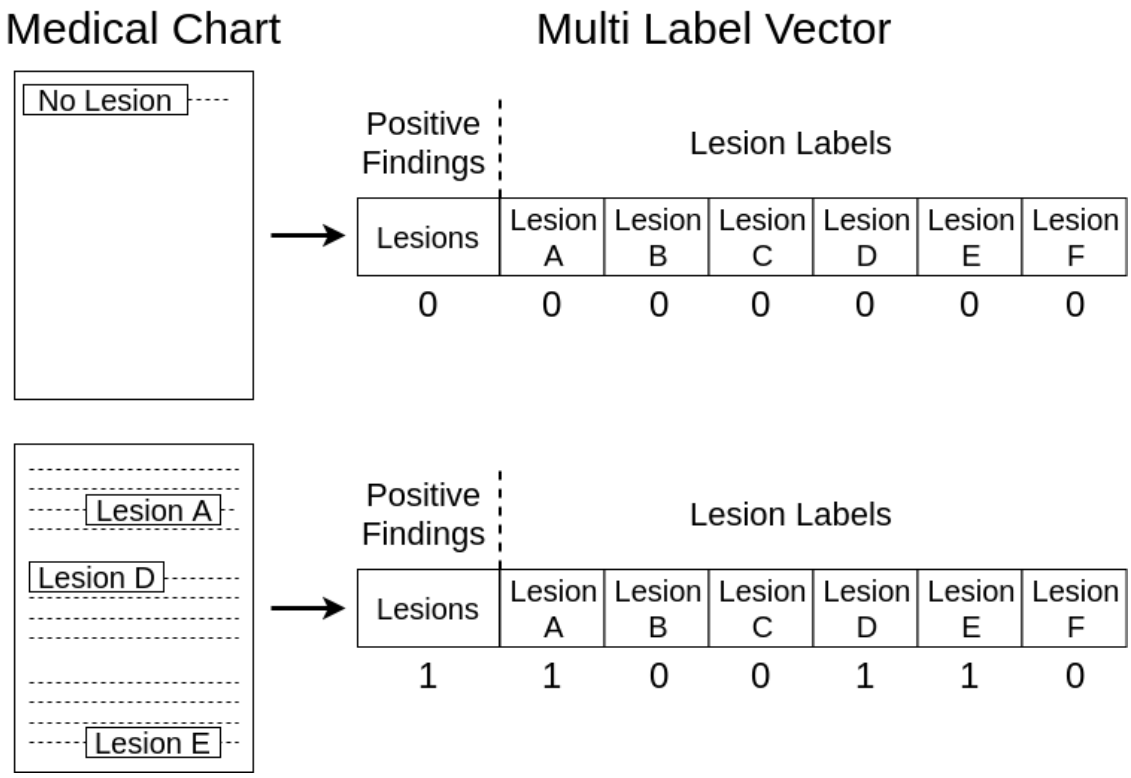


図 3.2: データセット作成

換した。次にこれらを形態素解析 [26] や事前に用意した病名リストを用いた検索などを通して病名を表す単語とそれらに付随する情報を示す単語を抽出した。その際に複雑な付随情報を持つような病名は固有の処理を加えている。これにより簡易カルテの質的診断の項目では、各行が、先頭に病名単語があり、その後に付加情報単語が続く形式になる。(3) では (2) の処理後の内容を全ての患者を通して集計し、同じ単語を同じラベル付けをして辞書に格納した。(4) では辞書内で出現回数を計測し、指定した回数以上のラベルを頻出病名ラベルとして決定した。本研究ではすべて頻出病名を 15 個に設定した。(5) (6) では各患者ごとに one-hot 化したマルチラベルを生成した。マルチラベルの 1 番目のラベルには異常が一つでもある患者は 1 を、それ以外の患者は 0 が格納されている。これには (1) で抽出した情報を用いた。マルチラベルの 2 番目以降のラベルには頻出病名ラベルがある。(2) から (4) で作成した情報を元に、各患者の質的診断の項目に頻出病名が存在する場合は 1 を、存在しない場合は 0 が格納されている。

3.2.2 各画像における処理

すべての画像は 256×256 ピクセルにサイズ変更した。その後画像の RGB チャンネルそれぞれを正規化した。正規化のパラメータは R が平均 0.485、標準偏差 0.229、G が平均 0.456、標準偏差 0.224、B が平均 0.406、標準偏差 0.225 に設定した。これらの値はすべて深層学習フレームワークの Pytorch [27] で推奨されている値である。

3.2.3 データセット分割

２つの手法に合わせてデータセットを２つ用意した。それぞれのデータセットは訓練データ、検証データ、テストデータに分割した。

- 画像を個別に入力する手法

  この手法では、学習に用いる訓練データと検証データは画像を全患者に

渡ってランダムにし、それぞれをマルチラベルと関連付けした。それに対して、テストでは患者単位での予測になるために、テストデータは患者単位で検査順に並んだ画像を系列データとしてマルチラベルと関連付けした。

- 画像を患者ごとにまとめて入力する手法
  この手法では、学習とテストに用いる全てのデータを患者単位でまとめた。データは検査順に並べた画像を時間軸で連結させ、三次元データとし、それぞれマルチラベルと関連付けした。

3.3 モデルの作成

2つの手法に合わせてモデルも2つ用意した。どちらのモデルでも残差構造を持っていて、表現力の大きい多層のCNNを用いている。モデルの選択はPytorchで安定的に良い性能があるとされていて、用いられることが多いものを参考にした。

- 画像を個別に入力する手法
  この手法では残差構造を持つCNN [1] で最も一般的なResNet [4] を発展させたDenseNet [28] を用いた。残差構造では、従来のネットワークでは同じ深さの層では同じ大きさの特徴しか抽出できなかったのに対して、ショートカットを用いることで同じ深さの層でも異なる大きさの特徴を抽出できるようになっている。ResNet [4] はネットワークの主要部分で畳み込みを行って、ショートカット部分でその畳み込みを飛ばす構造をしている。DenseNet [28] では主要部分では畳み込みを行わず、分岐した部分に設置したDense Blockによって畳み込みを行い、その出力が主要部分で合流する構造をしている。これによりResNet [4] より少ないパラメータ数で効率良く特徴抽出ができることが確認されている。

- 画像を患者ごとにまとめて入力する手法
  この手法では画像を連結し、時系列データを三次元データとして入力する

ため、CNN を三次元に拡張した 3D-CNN を用いた。3D-CNN の中でも現在最も良い性能を持っている 3D-ResNet [29] を選択した。

3.3.1 DenseNet のチューニング

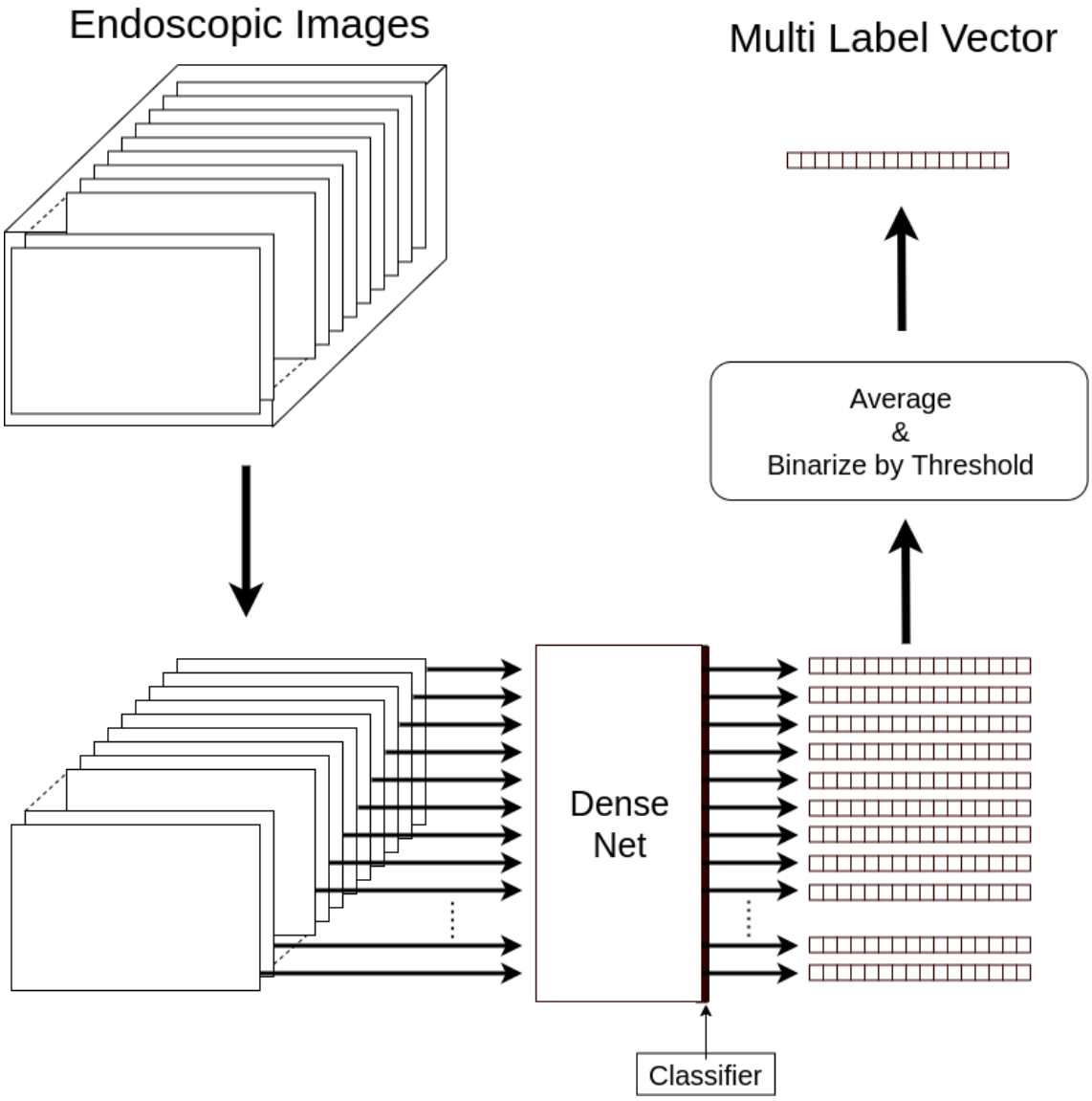


図 3.3: DenseNet を使った手法

DenseNet のチューニングをし、複数のモデルを用意した。これを図 3.3 に示

す。全てのモデルにおいて共通していることは、特徴抽出部分をそのまま用いていることと、分類器部分の最終層のニューロン数をマルチラベルのクラス数と合わせていることである。これに加えて分類器部分を変更したものも用意した。通常は一層の全結合層のみの構造であったが、複数の全結合層とバッチ正規化とドロップアウトをからなる構造に変更した。マルチラベル予測においては、抽出した特徴の組み合わせが 1 つのラベルの予測よりも複雑になる。そのため分類器においても深層の構造を用いた。また今回の実験ではより複雑な特徴にも対応するために、層の数の異なる 2 つの DenseNet で学習した。それぞれ DenseNet121 [28] と DenseNet161 [28] というモデルを用いた。このため合わせて 4 パターンで実験を行った。

3.3.2 3D-ResNet のチューニング

3D-ResNet のチューニングをし、複数のモデルを用意した。これを図 3.4 に示す。特徴抽出部分の最終層において、通常では平均プーリングをして特徴を合算している。その際に特徴が損失する恐れがあったため、これを最大プーリングに置き換えたものも用意した。また分類器部分でも DenseNet のときと同じように 2 つのパターンを用意した。これは一層の全結合層のみのものと、複数の全結合層とバッチ正規化とドロップアウトからなる構造のものである。このため合わせて 4 パターンで実験を行った。

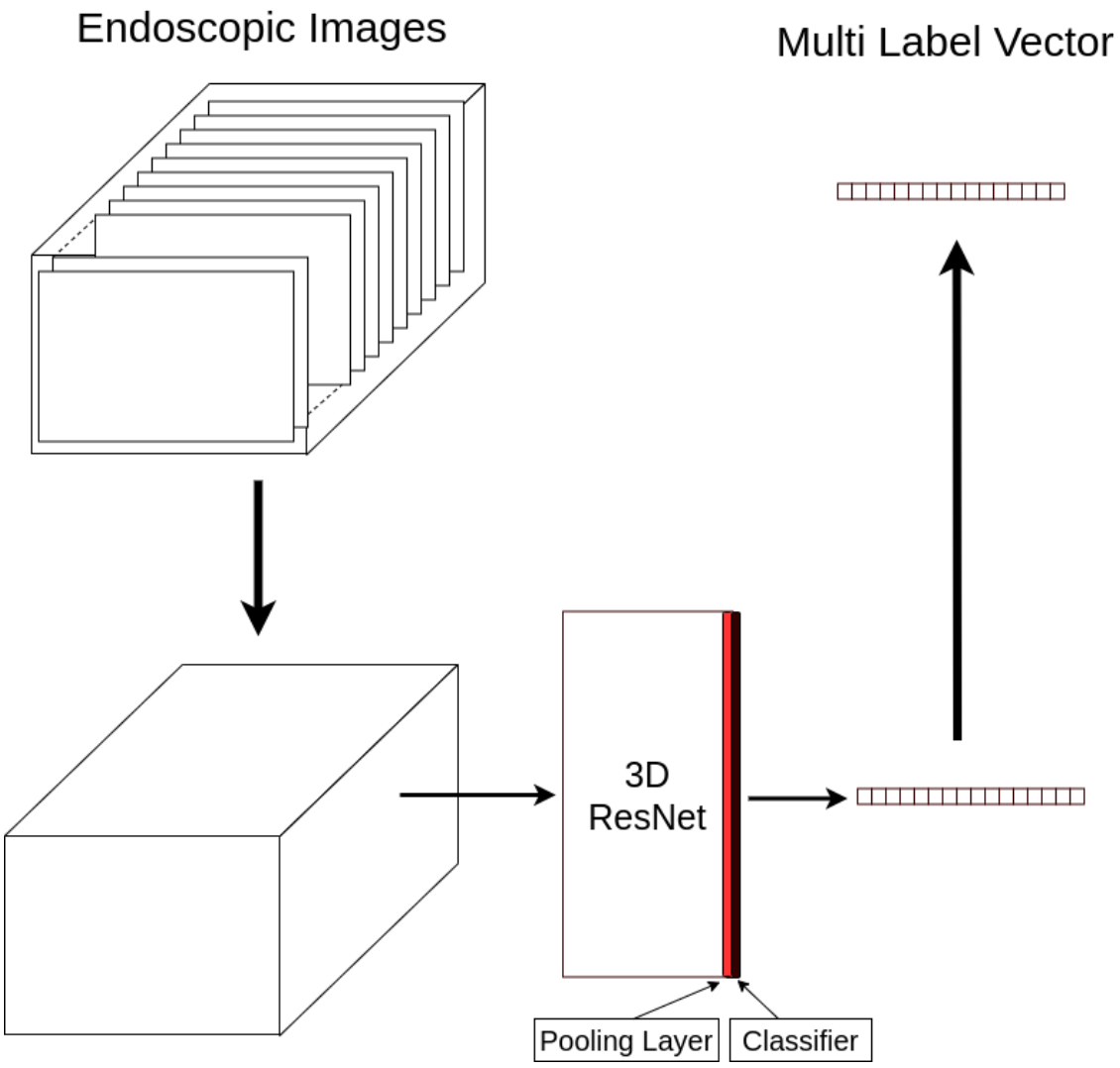


図 3.4: 3D-ResNet を使った手法

第 4 章

評価実験

4.1　実験条件の設定

4.1.1　モデルの損失関数

マルチラベル予測の学習のためバイナリ交差エントロピーを用いた。

4.1.2　モデルの最適化手法

確率的勾配降下法 [30] を拡張した Adaptive Moment Estimation (Adam) [31] を使用した。これは勾配の大きさと更新量によって学習率を変化させていく方法で、様々なタスクで高性能を記録している。

4.1.3　モデルの学習回数とバッチサイズ

モデルの学習回数はすべて 25 にした。これは訓練時における損失関数の推移から判断した。またバッチサイズは DenseNet121 で 50、DenseNet161 で 24、3D-ResNet で 3 とした。これは使用した計算機のメモリの容量によって決めた。

4.2　評価指標

評価指標は正解率、完全一致正解率、再現率、適合率、F1-Score の 5 つとした。以下本文中の各データとは、画像を個別に入力する手法では各画像、画像

を患者ごとにまとめて入力する手法では画像を時間軸で連結した 3 次元データを示す。

4.2.1 正解率 (Acc)

各データにおける正解率 $=$ $\frac{\text{正解したラベルの数}}{\text{マルチラベルにおける全てのラベルの数}}$ を計算し、これをすべてのデータにおいて平均を計算した。

4.2.2 完全一致正解率 (AllAcc)

$AllAcc = \frac{\text{マルチラベルにおける全てのラベルで一致したデータの数}}{\text{全てのデータの数}}$

4.2.3 適合率 (Precision)

$Precision = \frac{\text{真陽性}}{\text{真陽性+偽陽性}}$

全てのデータにおける適合率を計算するために、各データにおける混合行列を集計し、その後適合率を計算した。

4.2.4 再現率 (Recall)

$Recall = \frac{\text{真陽性}}{\text{真陽性+偽陰性}}$

全てのデータにおける再現率を計算するために、各データにおける混合行列を集計し、その後再現率を計算した。

4.2.5 F1-Score

適合率と再現率がトレードオフの関係であるため、2 つの指標を総合的に判断するために F1-Score を用いた。

$F1\text{-}Score = \frac{Precision * Recall}{(Precision + Recall)/2}$

4.2.6 評価指標とモデルの性能の関係性

- 正解率

  正解率は指標としてモデルの性能をあまり評価できない。マルチラベル予測の際にこの指標を用いると、マルチラベルのクラス数が多いほど真陰性の割合が多くなり、実際の予測がほとんど行われていなくても高い数値が出るからである。

- 完全一致正解率

  完全一致正解率は指標としてモデルの性能をあまり評価できない。マルチラベル予測の際にこの指標を用いると、マルチラベルのクラス数が多いほど全てを一致させることが困難になり、ほぼ全てのクラスで正解しているもとと全く正解していないものを区別できない。

- 適合率

  適合率はマルチラベル予測の指標として一般的に用いられる。適合率は陽性であると予測したものの中で、実際に陽性であるものの割合である。これは本用途においては病変があると予測したものの中で、実際に病変があったものの割合となっており、誤検知の少なさの指標と言える。

- 再現率

  再現率はマルチラベル予測の指標として一般的に用いられる。再現率は実際に陽性であるものの中で、陽性であると予測できたものの割合である。これは本用途においては病変があるデータの中で、病変があると予測できたものの割合となっており、見落としの少なさの指標と言える。

- F1-Score

  F1-Score はモデルの性能を最も表していると言える。本実験ではこの数値が高いものを良いモデルとして評価する。

4.3 DenseNet を用いた内視鏡画像からのマルチラベル予測

4.3.1 最も性能の良いモデルの選定

実験概要

1. DenseNet121-分類器拡張なし (モデル 1)
2. DenseNet121-分類器拡張あり (モデル 2)
3. DenseNet161-分類器拡張なし (モデル 3)
4. DenseNet161-分類器拡張あり (モデル 4)

以上の 4 モデルで実験を行い、性能を比較した。実験の流れは以下のように行った。各画像をマルチラベルと関連付けしたデータセットを用いた。画像をモデルに入力し出力されたマルチラベルと教師データのマルチラベルから損失を計算し最適化を行った。実験の出力結果は、学習過程での損失と F1-Score の推移と、検証データを用いた際の各評価指標の値となる。

実験結果

モデル 1 の結果を図 4.1 と表 4.1 に示す。図 4.1 は学習過程での損失と F1-Score の推移を示している。表 4.1 は検証データを用いた際の予測の結果を示す。この表の Total は図 3.2 における全てのラベルにおける結果を、Lesion と Label は図 3.2 のラベルの 1 番目と 2 番目以降に分けて計算した結果を示している。

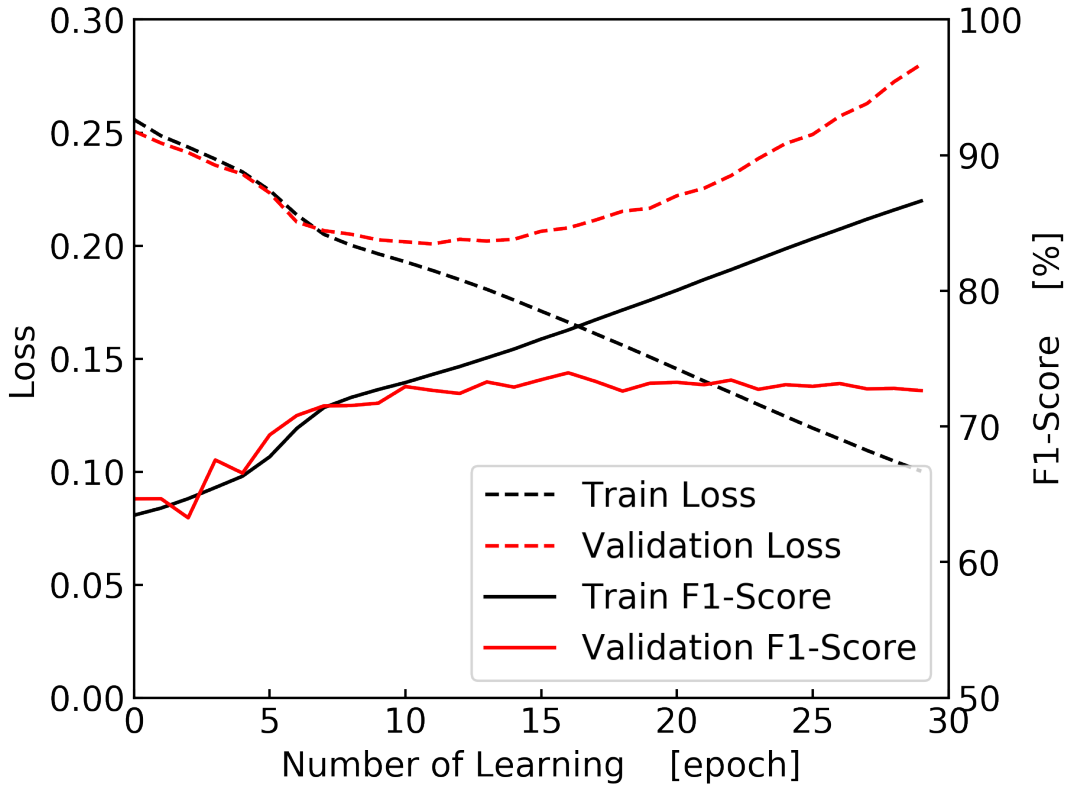


図 4.1: モデル 1 の学習過程

表 4.1: モデル 1 の検証結果

Total	Acc (%)	92.4
	AllAcc (%)	32.7
	F1-Score (%)	74.0
	Precision (%)	88.3
	Recall (%)	63.6
Lesion	Acc (%)	97.1
	F1-Score (%)	98.5
	Precision (%)	98.2
	Recall (%)	98.8
Label	Acc (%)	92.1
	AllAcc (%)	33.9
	F1-Score (%)	54.5
	Precision (%)	77.2
	Recall (%)	42.1

モデル 2 の結果を図 4.2 と表 4.2 に示す。図 4.2 は学習過程での損失と F1-Score の推移を示している。表 4.2 は検証データを用いた際の予測の結果を示す。この表の Total は図 3.2 における全てのラベルにおける結果を、Lesion と Label は図 3.2 のラベルの 1 番目と 2 番目以降に分けて計算した結果を示している。

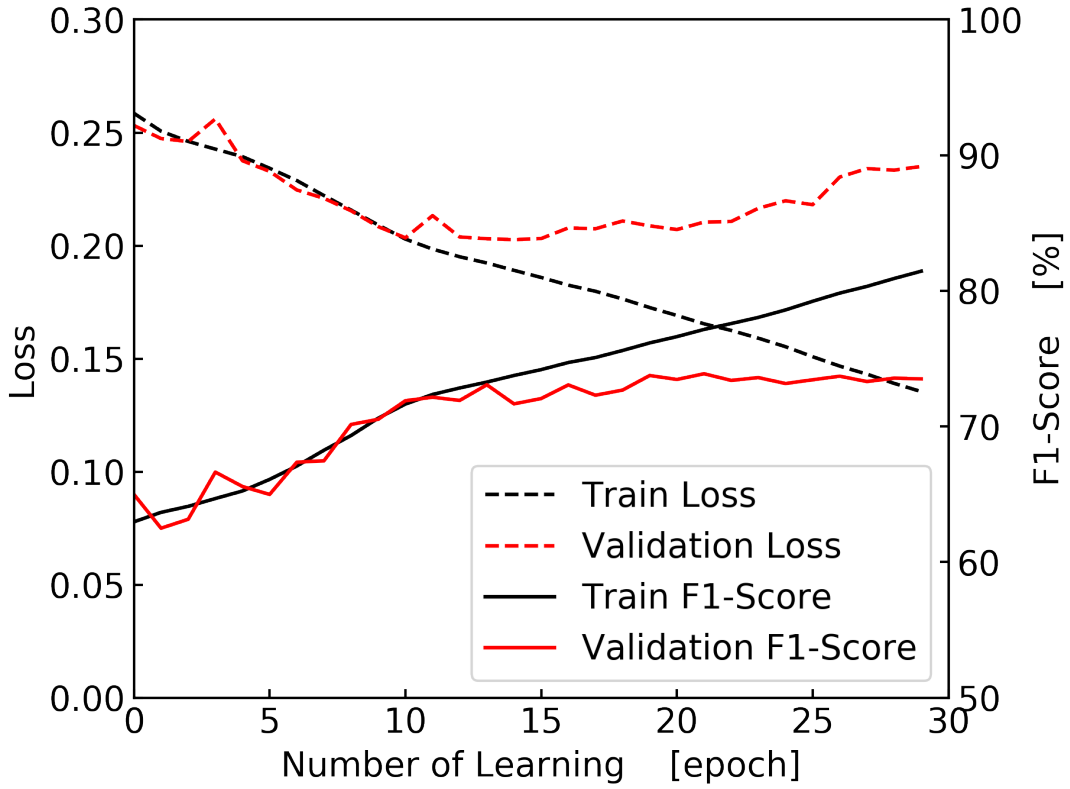


図 4.2: モデル 2 の学習過程

表 4.2: モデル 2 の検証結果

Total	Acc (%)	92.4
	AllAcc (%)	32.8
	F1-Score (%)	73.9
	Precision (%)	88.8
	Recall (%)	63.2
Lesion	Acc (%)	97.4
	F1-Score (%)	98.7
	Precision (%)	98.0
	Recall (%)	99.4
Label	Acc (%)	92.1
	AllAcc (%)	34.1
	F1-Score (%)	53.9
	Precision (%)	78.1
	Recall (%)	41.2

モデル 3 の結果を図 4.3 と表 4.3 に示す。図 4.3 は学習過程での損失と F1-Score の推移を示している。表 4.3 は検証データを用いた際の予測の結果を示す。この表の Total は図 3.2 における全てのラベルにおける結果を、Lesion と Label は図 3.2 のラベルの 1 番目と 2 番目以降に分けて計算した結果を示している。

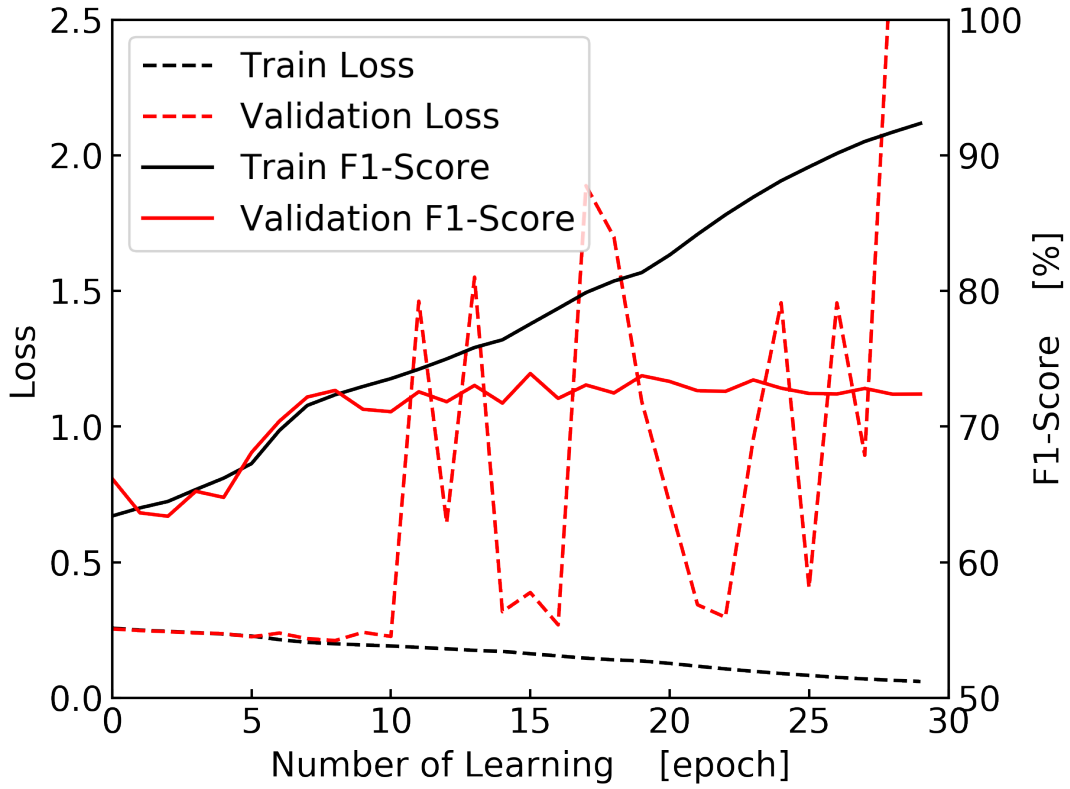


図 4.3: モデル 3 の学習過程

表 4.3: モデル 3 の検証結果

Total	Acc (%)	92.4
	AllAcc (%)	32.3
	F1-Score (%)	73.9
	Precision (%)	87.8
	Recall (%)	63.8
Lesion	Acc (%)	97.3
	F1-Score (%)	98.6
	Precision (%)	97.8
	Recall (%)	99.4
Label	Acc (%)	92.0
	AllAcc (%)	33.7
	F1-Score (%)	54.3
	Precision (%)	76.6
	Recall (%)	42.1

モデル 4 の結果を図 4.4 と表 4.4 に示す。図 4.4 は学習過程での損失と F1-Score の推移を示している。表 4.4 は検証データを用いた際の予測の結果を示す。この表の Total は図 3.2 における全てのラベルにおける結果を、Lesion と Label は図 3.2 のラベルの 1 番目と 2 番目以降に分けて計算した結果を示している。

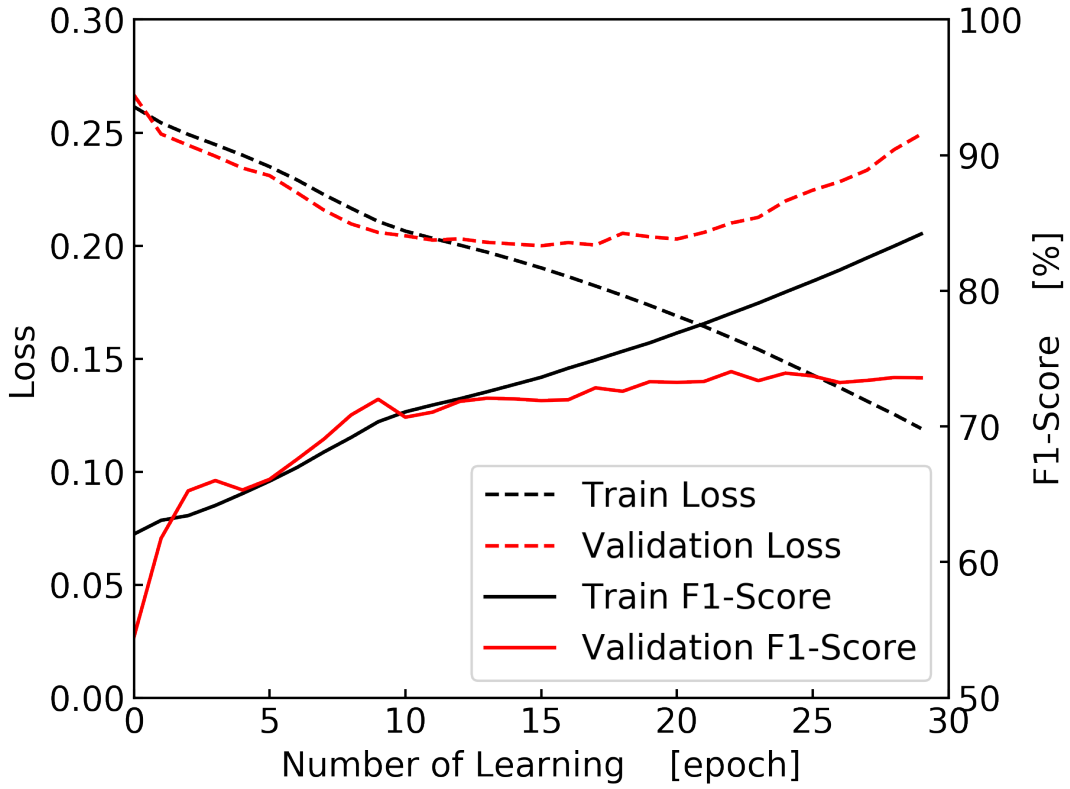


図 4.4: モデル 4 の学習過程

表 4.4: モデル 4 の検証結果

Total	Acc (%)	92.5
	AllAcc (%)	33.4
	F1-Score (%)	74.0
	Precision (%)	88.6
	Recall (%)	63.6
Lesion	Acc (%)	98.0
	F1-Score (%)	98.9
	Precision (%)	98.4
	Recall (%)	99.5
Label	Acc (%)	92.1
	AllAcc (%)	34.3
	F1-Score (%)	54.2
	Precision (%)	77.3
	Recall (%)	41.7

4.3.2 しきい値を変化させた際の適合率と再現率の変化の確認

実験概要

4.3.1 の実験で、マルチラベル全体での F1-Score が最も高かったモデル 1 を本実験で用いる。本実験では、モデルが出力したマルチラベルの各ラベルを二値化する際のしきい値を変化させた。しきい値は 0.1 から 0.9 まで 0.1 刻みで変化させ、その値における適合率、再現率、F1-Score を出力した。

図 4.5 はテスト推論において、モデルが出力したマルチラベルの値を二値化する際のしきい値を変化させた際の、適合率、再現率、F1-Score の変化を示している。

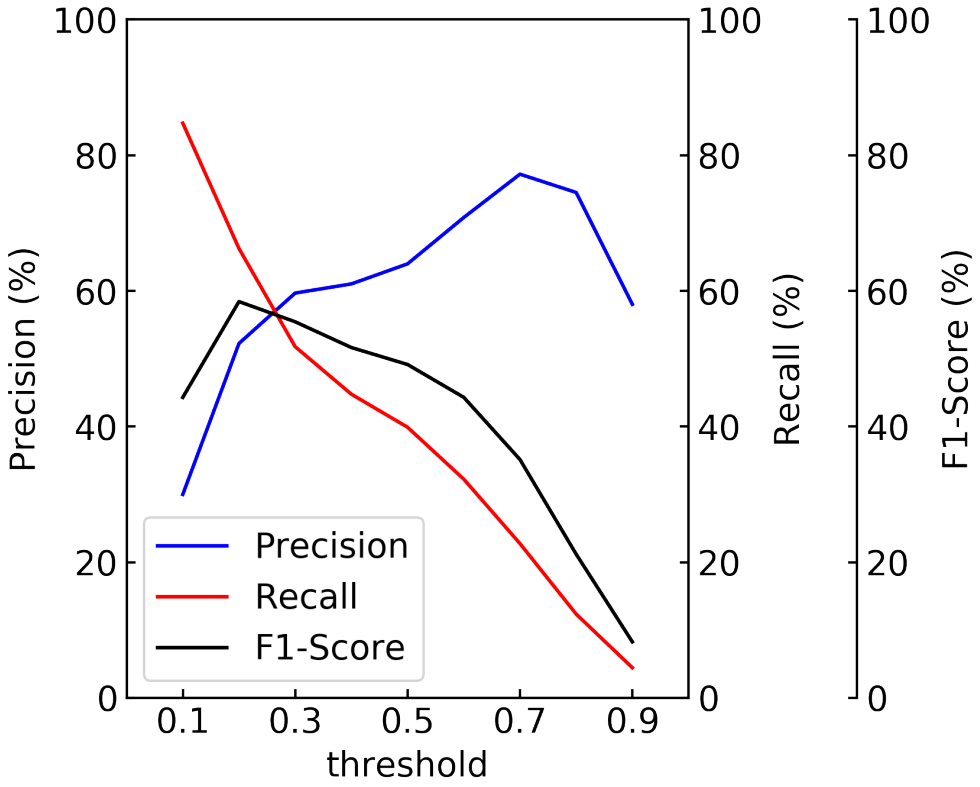


図 4.5: しきい値を変化させた際の適合率と再現率の変化

4.4 3D-ResNet を用いた内視鏡画像からのマルチラベル予測

4.4.1 最も性能の良いモデルの選定

実験概要

1. 3D-ResNet-AveragePool-分類器拡張なし (モデル 1)
2. 3D-ResNet-AveragePool-分類器拡張あり (モデル 2)
3. 3D-ResNet-MaxPool-分類器拡張なし (モデル 3)
4. 3D-ResNet-MaxPool-分類器拡張あり (モデル 4)

以上の 4 モデルで実験を行い、性能を比較した。実験の流れは以下のように行った。各患者ごとに画像を時間軸で連結した三次元データとマルチラベルと関連付けしたデータセットを用いた。三次元データをモデルに入力し出力されたマルチラベルと教師データのマルチラベルから損失を計算し最適化を行った。実験の出力結果は、学習過程での損失と F1-Score の推移と、検証データを用いた際の各評価指標の値となる。

実験結果

モデル 1 の結果を図 4.6 と表 4.5 に示す。図 4.6 は学習過程での損失と F1-Score の推移を示している。表 4.5 は検証データを用いた際の予測の結果を示す。この表の Total は図 3.2 における全てのラベルにおける結果を、Lesion と Label は図 3.2 のラベルの 1 番目と 2 番目以降に分けて計算した結果を示している。

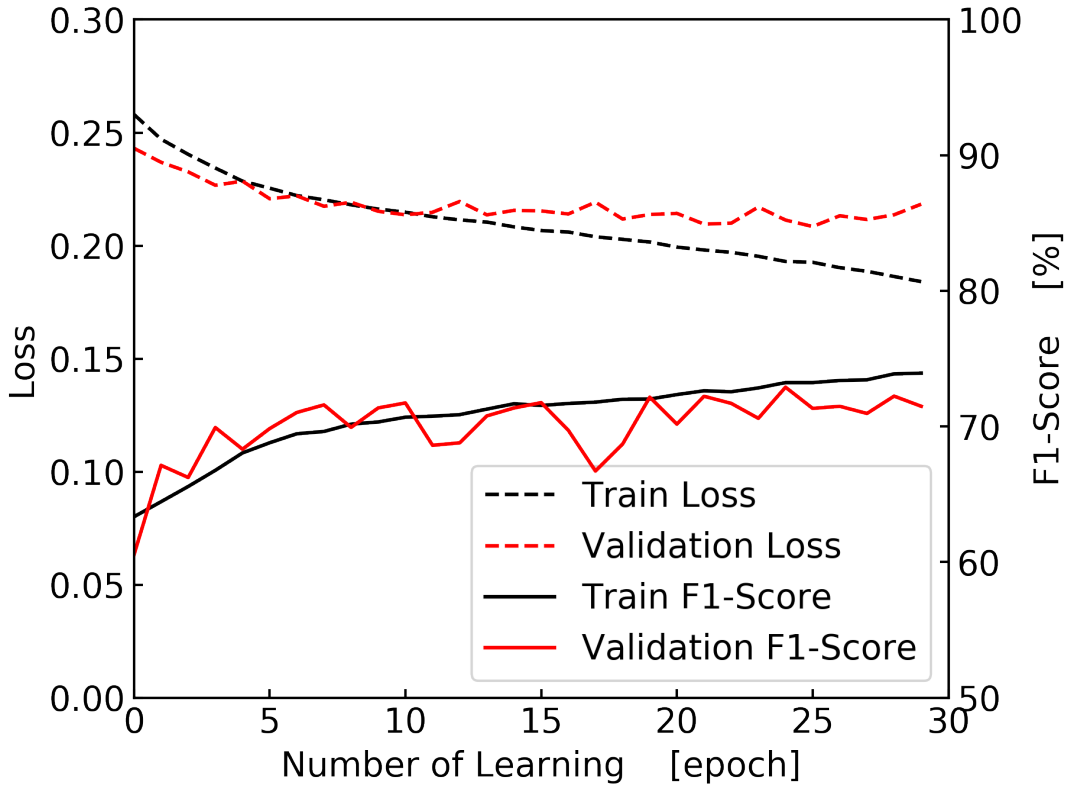


図 4.6: モデル 1 の学習過程

表 4.5: モデル 1 の検証結果

Total	Acc (%)	92.2
	AllAcc (%)	28.5
	F1-Score (%)	72.9
	Precision (%)	86.5
	Recall (%)	63.0
Lesion	Acc (%)	98.8
	F1-Score (%)	99.4
	Precision (%)	98.9
	Recall (%)	99.9
Label	Acc (%)	91.7
	AllAcc (%)	28.9
	F1-Score (%)	50.8
	Precision (%)	71.8
	Recall (%)	39.4

モデル 2 の結果を図 4.7 と表 4.6 に示す。図 4.7 は学習過程での損失と F1-Score の推移を示している。表 4.6 は検証データを用いた際の予測の結果を示す。この表の Total は図 3.2 における全てのラベルにおける結果を、Lesion と Label は図 3.2 のラベルの 1 番目と 2 番目以降に分けて計算した結果を示している。

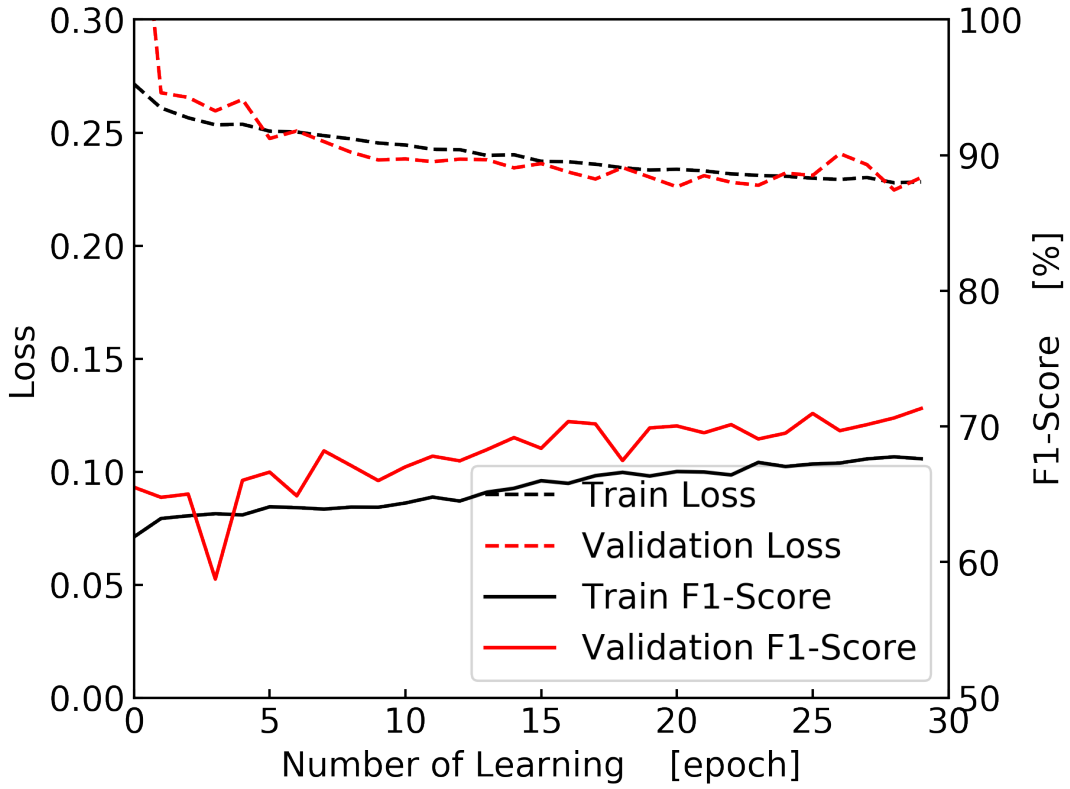


図 4.7: モデル 2 の学習過程

表 4.6: モデル 2 の検証結果

Total	Acc (%)	91.7
	AllAcc (%)	23.3
	F1-Score (%)	71.3
	Precision (%)	85.7
	Recall (%)	61.1
Lesion	Acc (%)	98.8
	F1-Score (%)	99.4
	Precision (%)	99.0
	Recall (%)	99.8
Label	Acc (%)	91.2
	AllAcc (%)	23.5
	F1-Score (%)	48.3
	Precision (%)	69.9
	Recall (%)	36.9

モデル 3 の結果を図 4.8 と表 4.7 に示す。図 4.8 は学習過程での損失と F1-Score の推移を示している。表 4.7 は検証データを用いた際の予測の結果を示す。この表の Total は図 3.2 における全てのラベルにおける結果を、Lesion と Label は図 3.2 のラベルの 1 番目と 2 番目以降に分けて計算した結果を示している。

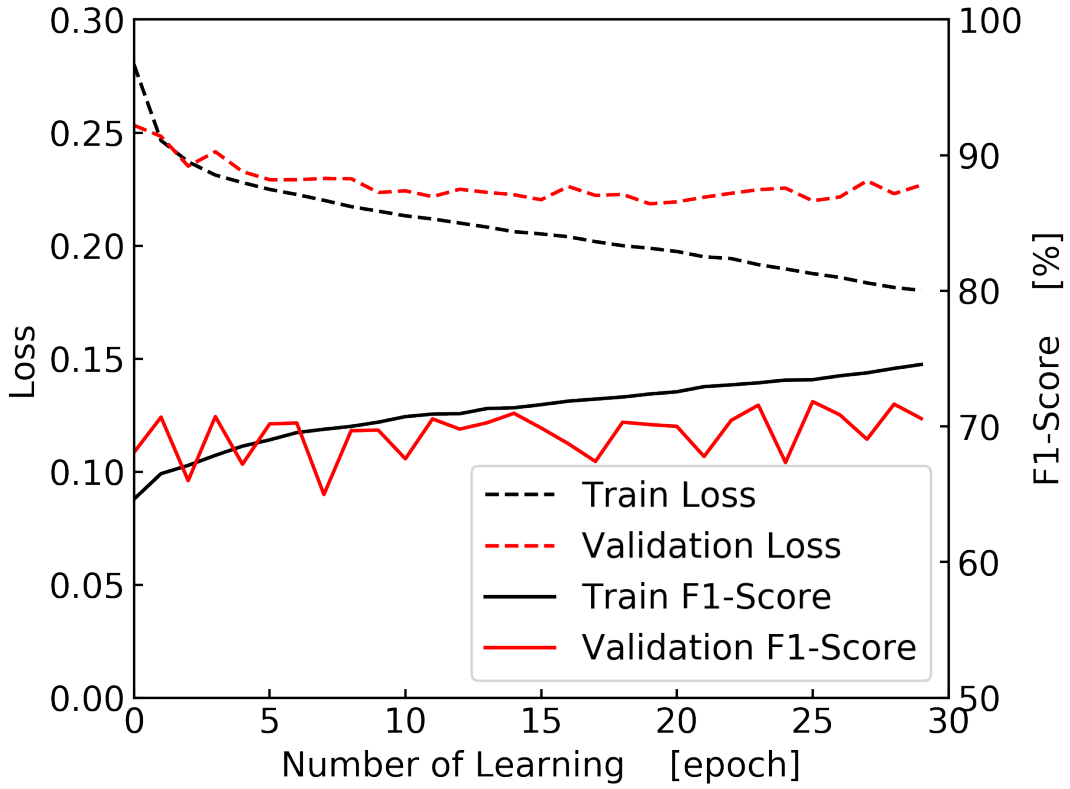


図 4.8: モデル 3 の学習過程

表 4.7: モデル 3 の検証結果

Total	Acc (%)	92.0
	AllAcc (%)	26.3
	F1-Score (%)	71.8
	Precision (%)	88.7
	Recall (%)	60.4
Lesion	Acc (%)	99.0
	F1-Score (%)	99.5
	Precision (%)	99.0
	Recall (%)	99.9
Label	Acc (%)	91.5
	AllAcc (%)	27.0
	F1-Score (%)	48.4
	Precision (%)	75.0
	Recall (%)	35.7

モデル 4 の結果を図 4.9 と表 4.8 に示す。図 4.9 は学習過程での損失と F1-Score の推移を示している。表 4.8 は検証データを用いた際の予測の結果を示す。この表の Total は図 3.2 における全てのラベルにおける結果を、Lesion と Label は図 3.2 のラベルの 1 番目と 2 番目以降に分けて計算した結果を示している。

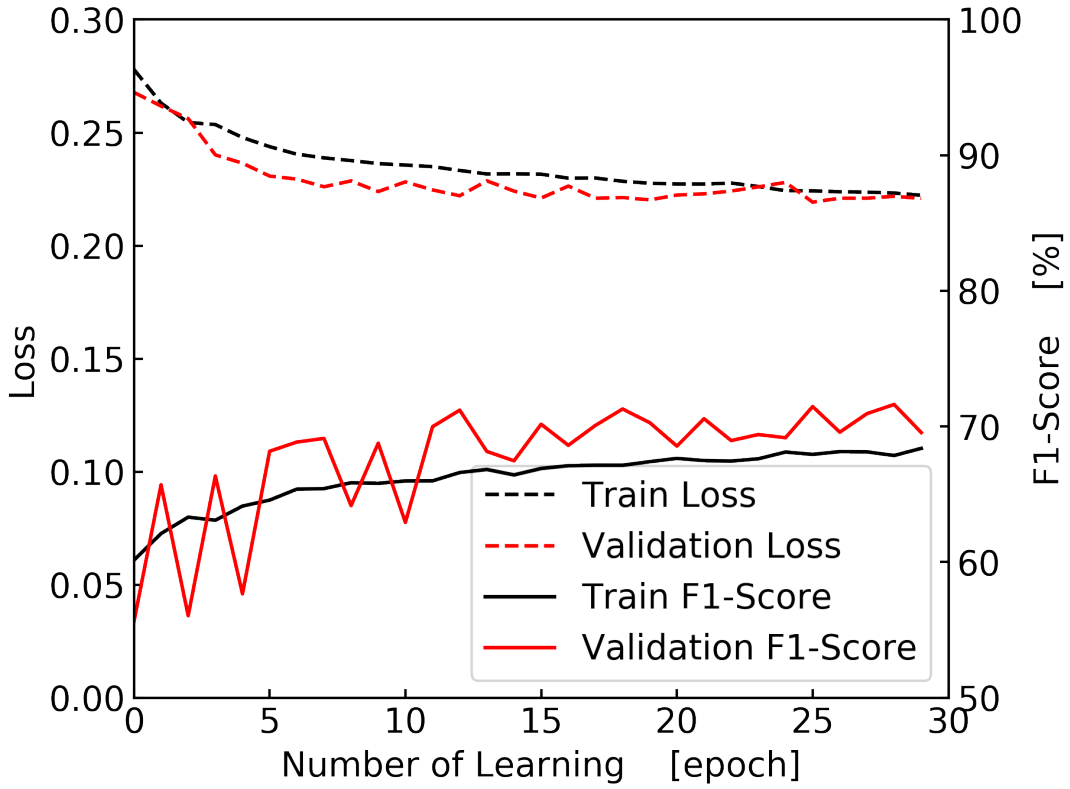


図 4.9: モデル 4 の学習過程

表 4.8: モデル 4 の検証結果

Total	Acc (%)	91.8
	AllAcc (%)	26.5
	F1-Score (%)	71.6
	Precision (%)	87.0
	Recall (%)	60.9
Lesion	Acc (%)	98.2
	F1-Score (%)	99.1
	Precision (%)	98.9
	Recall (%)	99.3
Label	Acc (%)	91.4
	AllAcc (%)	26.9
	F1-Score (%)	48.7
	Precision (%)	72.2
	Recall (%)	36.8

4.4.2 しきい値を変化させた際の適合率と再現率の変化の確認

実験概要

4.4.1 の実験で、マルチラベル全体での F1-Score が最も高かったモデル 1 を本実験で用いる。本実験では、モデルが出力したマルチラベルの各ラベルを二値化する際のしきい値を変化させた。しきい値は 0.1 から 0.9 まで 0.1 刻みで変化させ、その値における適合率、再現率、F1-Score を出力した。

図 4.10 はテスト推論において、モデルが出力したマルチラベルの値を二値化する際のしきい値を変化させた際の、適合率、再現率、F1-Score の変化を示している。

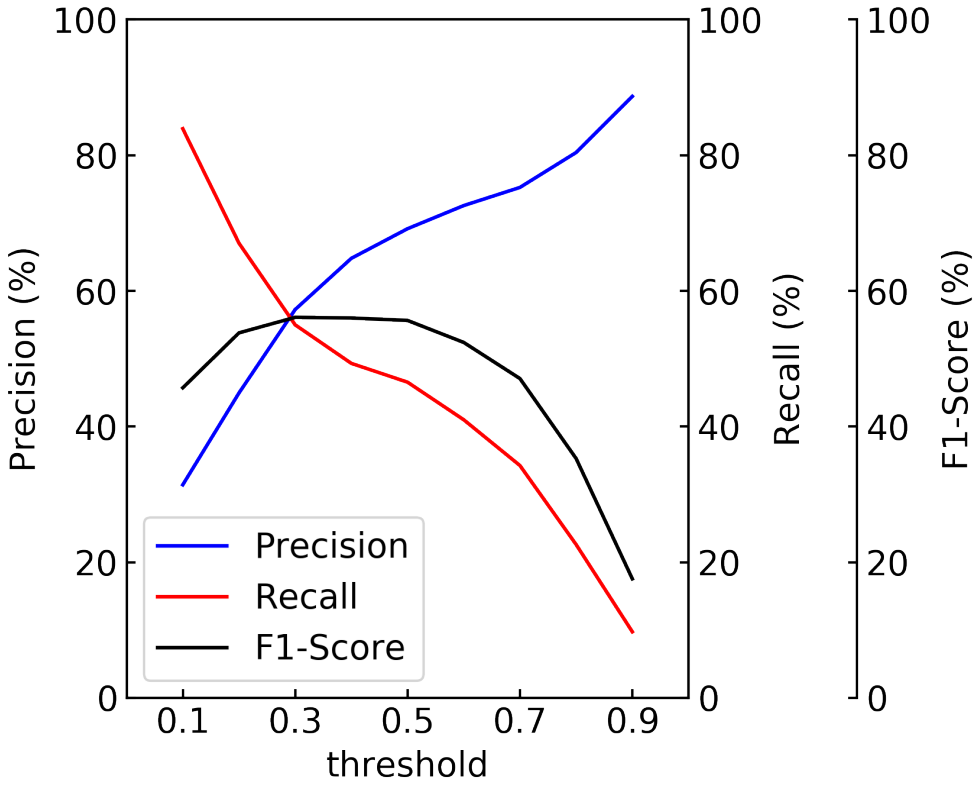


図 4.10: しきい値を変化させた際の適合率と再現率の変化

第 5 章

考察

5.1 各実験における考察

5.1.1 DenseNet を用いた内視鏡画像からのマルチラベル予測における考察

学習過程を見ると、損失がエポックが進むごとに減っていて、かつ Acc と AllAcc の両方が上昇していることから、学習がきちんと進んでいることがわかる。また訓練損失が減少を続けているのに対して、検証損失が学習途中から上昇に転じている。これはデータに関して過学習を起こしているが、DenseNet121 から DenseNet161 に増やしたのに関わらず変化がないことから、これはモデルの大きさによって特徴表現が狭まっているわけではない。マルチラベル学習でバイナリ交差エントロピーを用いていることから、モデルの学習が進んで予測されるラベルが増えるに従って損失が増えやすくなっていると考えられる。また各画像をマルチラベルと関連付けて学習するこの手法では、病変がある患者において、病変が全く見られない部位の画像も病変の存在を示すマルチラベルと学習するために学習が安定しずらいと考えられる。しかし Acc や F1-Score が良くなっていることから、学習は問題なく進んでいると見なすことができる。

検証結果を見ると、AllAcc が 30%を超えていて、F1-Score も 70%を超えており、モデルがきちんと学習されていることがわかる。しかしこの結果には複数の問題がある。AllAcc に関しては、マルチラベルの中でもそもそもラベルが真になっている部分の少ない教師データの際に予測ができているだけの可能性が

高い。F1-Score もマルチラベルの一つ目の病変が 1 つでもあるかどうかのラベルの推論結果が良いために、全体の結果が良くなっていると考えられる。このためマルチラベルの検証結果を一つ目のラベルと二つ目以降のラベルに分離した結果を提示した。この結果を見ると、一つ目のラベルは Precision と Recall ともに 98%を超えていて、病変があるかどうかの判定は高い精度を持っていることがわかる。また二つ目以降のラベルでも Precision が 77%、Recall が 42%、F1-Score が 54%と判定がある程度はできていることがわかる。

検証結果では推論時のマルチラベルの二値化の際のしきい値が 0.5 に設定されているが、テストの際にはしきち値を 0.1 から 0.9 まで 0.1 ずつ変化させて Precision と Recall の推移を見た。Precision と Recall の値の性質上、両者はトレードオフの関係にあり、しきち値が低いと Recall が上がり、高いと Precision が上がる。この実験の結果、しきち値が 0.3 のときに両者がバランス良く、F1-Score が一番高かった。またしきい値が 0.1 のときに Recall が 85%と一番高く、しきい値が 0.7 のときに Precision が 77%と一番高くなった。

この結果を用いて本研究での提案システムとしては、しきい値が 0.7 のときの結果を予測の範囲は狭いが信頼度の高い予測として、しきい値が 0.1 のときの結果を信頼度は低いが広範囲の予測として用いる。これにより、医師の診断時の判断の支援も行える。

5.1.2 3D-ResNet を用いた内視鏡画像からのマルチラベル予測における考察

3D-ResNet を用いた実験では DenseNet での場合と同様に、損失の推移や Acc と AllAcc の値から学習が問題なく行われていると見なすことができる。また DenseNet の場合とは異なり、学習回数が進んでも訓練損失と検証損失の差が開いておらず、過学習が抑えられていると。これは各患者ごとに画像を時間軸で連結した三次元データを用いていることから、全く関係ない特徴量とマルチラベルが学習されることが減り、学習が安定したことが考えられる。しかし AllAcc や F1-Score が DenseNet と比べてわずかに低くなっている。これは 3D-CNN を

用いたことにより、特徴抽出部分の最終層での特徴の合算の際に特徴が多少消失してしまったためであると考えられる。

検証結果は DenseNet の場合と同様に、一つ目のラベルは Precision と Recall ともに 98%を超えていて、病変があるかどうかの判定は高い精度を持っていることがわかる。また二つ目以降のラベルでも Precision が 71%、Recall が 39%、F1-Score が 50%と判定がある程度はできていることがわかる。しかし検証結果も学習過程での見られたのと同じ傾向があり、DenseNet よりもわずかに結果が低くなっている。

推論時のマルチラベルの二値化の際のしきい値の推移を見ると、しきい値が 0.2 のときに Precision と Recall のバランスが良く、F1-Score が一番高かった。またしきい値が 0.1 のときに Recall が 83%と一番高く、しきい値が 0.9 のときに Precision が 88%と一番高かった。しかし、しきい値が 0.9 のときは Recall が 9%、F1-Score も 17%と非常に低くなっている。このためしきい値が高すぎて予測ラベルが少なすぎる状態になっていると考えられるため、しきい値が 0.7 のときが Precision が 75%かつ Recall が 34%と信頼度の高い予測として有用だとみなすことができる。これにより、その他の結果と同様に 3D-ResNet のときは DenseNet のときよりも少し精度が低いという結果になった。

5.2 全体における考察

画像を個別に入力する手法と画像を患者ごとにまとめて入力する手法の両方で非常に有用な結果が出たと言える。画像を患者ごとにまとめて入力する手法のほうが、関係ない特徴とラベルの学習がなくなるために学習は安定する。それに対して画像を個別に入力する手法の方は、学習は不安定化する。しかし予測の際も 1 枚ずつ予測し、その結果を患者ごとに集約するために、間違った予測が弱まり、正しい予測が重ね合わさることで、患者単位の予測では高い精度を出すことに成功した。どちらの手法のモデルでもしきち値の変更で Precision が 70%前後の予測を信頼度の高い予測として、Recall が 80%前後の予測を信頼度は低いが可能性としてはありえる予測として提供することができ、診断の際

の判断の支援と見逃しの防止の両方を満たすシステムとして利用することができる。

第 6 章

結論

6.1　結論

本論文では、内視鏡画像からの簡易カルテ生成システムとしての深層学習モデルの学習を行った。既存研究では食道がんのみの予測かつ手動でのすべての画像へのラベル付けが必要となっていたのに対して、本研究では医者が実際に作成した簡易カルテから学習データとなるマルチラベルを自動で生成した。さらに食道がんのみに関わらず、頻出病名 15 個を含むマルチラベルを学習させることに成功した。既存研究では再現率に最適したモデルになっていたのに対し、本研究では適合率と再現率がともに高くなるように F1-Score に最適化されたモデルとなっている。そのため病変の見逃しと誤検知の両方が少ないモデルであるといえる。また推論時には、モデルで出力したマルチラベルを二値化する際のしきい値を変えた出力を複数予測として提供することにより、信頼度の異なる段階的な予測を可能にした。これにより医師の診断の際の支援として大きく役立つモデルになったといえる。今後の展望としては 2 つ挙げられる。一つ目はデータセットの複雑化である。データセット作成の際には病名ラベルだけでなく階層化した付加情報もマルチラベルかできるようにしたが、学習が煩雑で進まなくなったために病名ラベルのみを使用した。階層化した付加情報も学習が可能になるとより詳細な簡易カルテの生成が可能になる。二つ目は学習モデルの改良である。今回は個別に画像とマルチラベルを学習させる手法と患者ごとに画像をまとめる手法を提案したが、画像特徴を個別にとりつつ系列データを扱うモデルで患者ごとに特徴をまとめてマルチラベルと学習させることが可

能になれば、より精度の高い予測が可能になると考えられる。

謝辞

本研究を行うにあたり、指導教官の萩原将文教授から終始熱心なご指導を承りました。ここに感謝の意を表します。また研究室の方々には様々な相談をさせて頂き、特に同期の方々には研究を通じて活発な議論にお付き合い頂きましたことを感謝致します。さらに慶応義塾大学医学部消化器内科との共同研究であったため、種本俊先生と筋野智久先生には多大なご協力を頂き大変感謝いたします。

参考文献

[1] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," In *Advances in Neural Information Processing Systems 2*, pp. 396–404. Morgan-Kaufmann, 1990.

[2] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *CoRR*, Vol. abs/1404.5997, , 2014.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 09 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[5] M. JORDAN, "Attractor dynamics and parallelism in a connectionist sequential machine," *Proceedings of the Ninth Annual conference of the Cognitive Science Society*, pp. 531–546, 1986.

[6] J. Chung, Çaglar Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *ArXiv*, Vol. abs/1412.3555, , 2014.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," In I. Guyon,

U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017.

[9] D. Bahdanau, J. Chorowski, P. Serdyuk, h. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105. Curran Associates Inc., 2012.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

[12] D. Kingma and M. Welling, "Auto-encoding variational bayes," *ICLR*, 12 2013.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," In *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. Curran Associates, Inc., 2014.

[14] M.-T. Luong, H. Pham, and C. Manning, "Effective approaches to attention-based neural machine translation,", 2015.

[15] V. Bulitko and K. Doucet, "Anxious learning in real-time heuristic search," In *IEEE Symposium on Computational Intelligence and Games, CIG*, pp. 1–4, 08 2018.

[16] A. Lusci, G. Pollastri, and P. Baldi, "Deep architectures and deep learning in chemoinformatics: The prediction of aqueous solubility for drug-like molecules," *Journal of chemical information and modeling*, Vol. 53, , 06 2013.

[17] D. Elton, Z. Boukouvalas, M. Fuge, and P. Chung, "Deep learning for molecular design - a review of the state of the art," *Molecular Systems Design & Engineering*, No. 4, pp. 667–988, 2019.

[18] D. Kelley, J. Snoek, and J. Rinn, "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks.," *Genome Res.*, p. 028399, 10 2016.

[19] R. Singh and S. Srivastava, "Stock prediction using deep learning," *Multimedia Tools and Applications*, pp. 1–16, 11 2016.

[20] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, Vol. 542, pp. 115–118, 2017.

[21] A. Hosny, C. Parmar, J. Quackenbush, L. Schwartz, and H. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, Vol. 18, pp. 500–510, 05 2018.

[22] T. Hirasawa, K. Aoyama, T. Tanimoto, S. Ishihara, S. Shichijo, T. Ozawa, T. Ohnishi, M. Fujishiro, K. Matsuo, J. Fujisaki, and T. Tada, "Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images," *Gastric Cancer*, Vol. 21, pp. 1–8, 01 2018.

[23] Y. Horie, T. Yoshio, K. Aoyama, S. Yoshimizu, Y. Horiuchi, A. Ishiyama, T. Hirasawa, T. Tuchida, T. Ozawa, S. Ishihara, Y. Kumagai, M. Fujishiro, I. Maetani, J. Fujisaki, and T. Tada, "The diagnostic outcomes

of esophageal cancer by artificial intelligence using convolutional neural networks," *Gastrointestinal Endoscopy*, Vol. 89, , 08 2018.

[24] O. Hosokawa, M. Hattori, K. Douden, H. Hayashi, K. Ohta, and Y. Kaizaki, "Difference in accuracy between gastroscopy and colonoscopy for detection of cancer," *Hepato-gastroenterology*, Vol. 54, pp. 442–4, 04 2007.

[25] Wei Liu and Dragomir Anguelov and Dumitru Erhan and Christian Szegedy and Scott Reed, and Cheng-Yang Fu, and Alexander C. Berg, "Ssd: Single shot multibox detector,", 2016, To appear.

[26] 土屋 雅稔, 宇津呂 武仁, 松吉 俊, 佐藤理史 , 中川 聖一, "日本語複合辞用例データベースの作成と分析," 情報処理学会論文誌, Vol. 47, No. 6, pp. 1728–1741, jun 2006.

[27] "Pytorch," `https://pytorch.org/`.

[28] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[29] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[30] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, Vol. 23, , 09 1952.

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

付録 A

他のモデルにおける実験

A.1　実験 4.3.2

実験 4.3.1 の他のモデルにおける実験を行った。

A.1.1 モデル２の結果

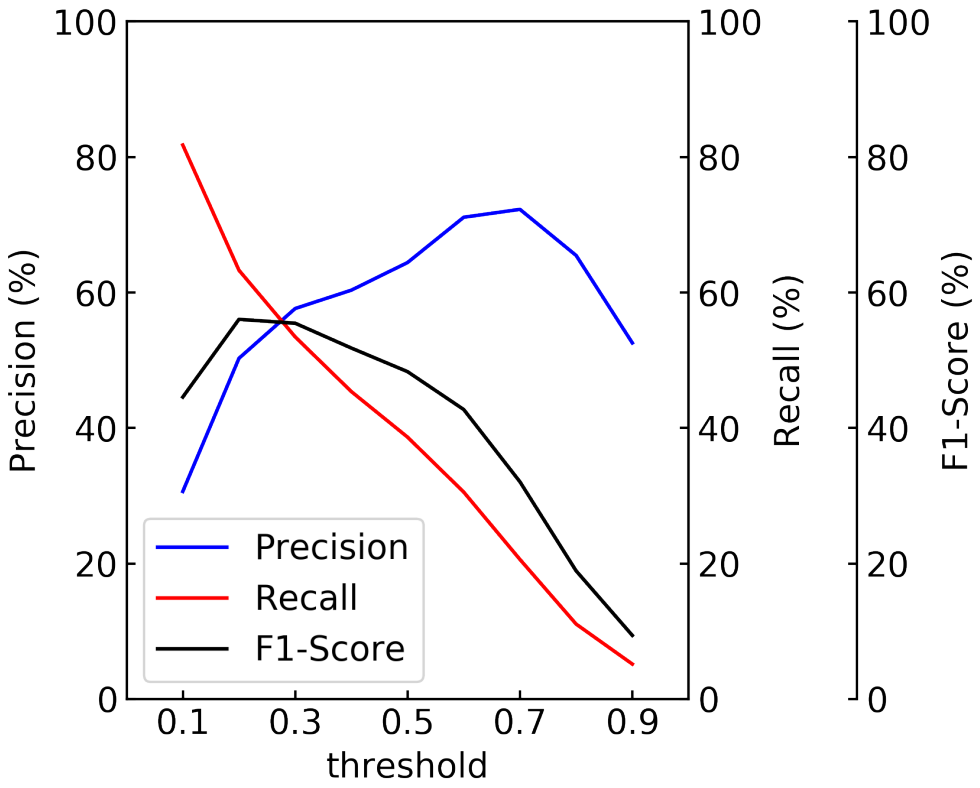


図 A.1: しきい値を変化させた際の適合率と再現率の変化

A.1.2 モデル３の結果

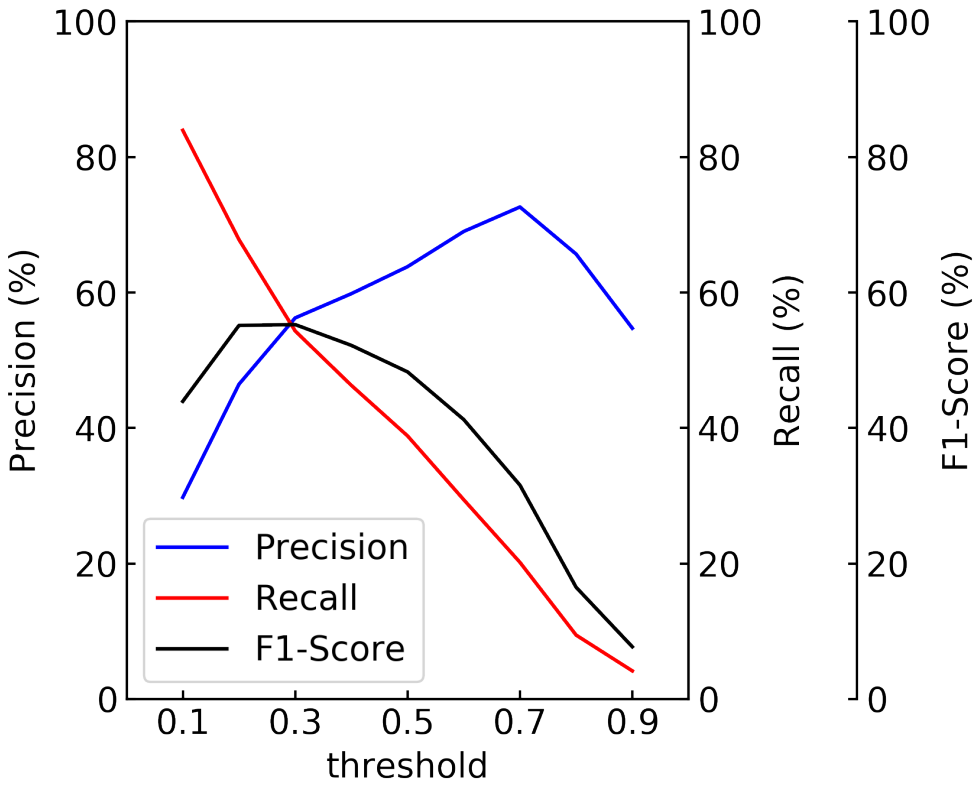


図 A.2: しきい値を変化させた際の適合率と再現率の変化

A.1.3 モデル４の結果

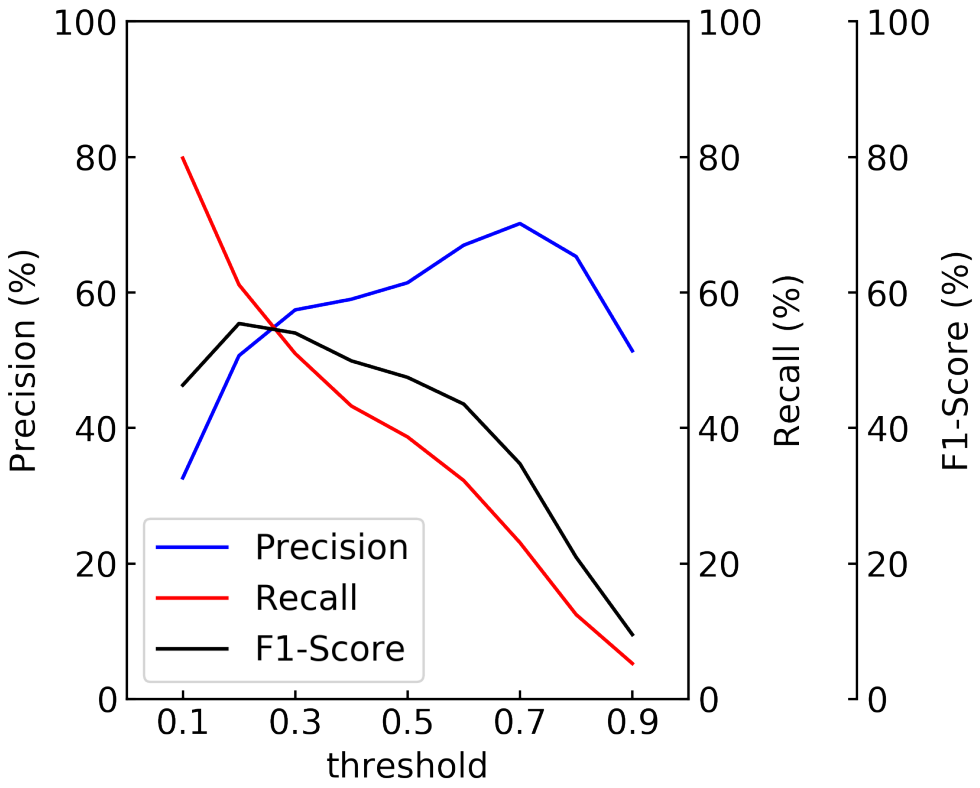


図 A.3: しきい値を変化させた際の適合率と再現率の変化

A.2 実験 4.4.2

実験 4.4.1 の他のモデルにおける実験を行った。

A.2.1 モデル２の結果

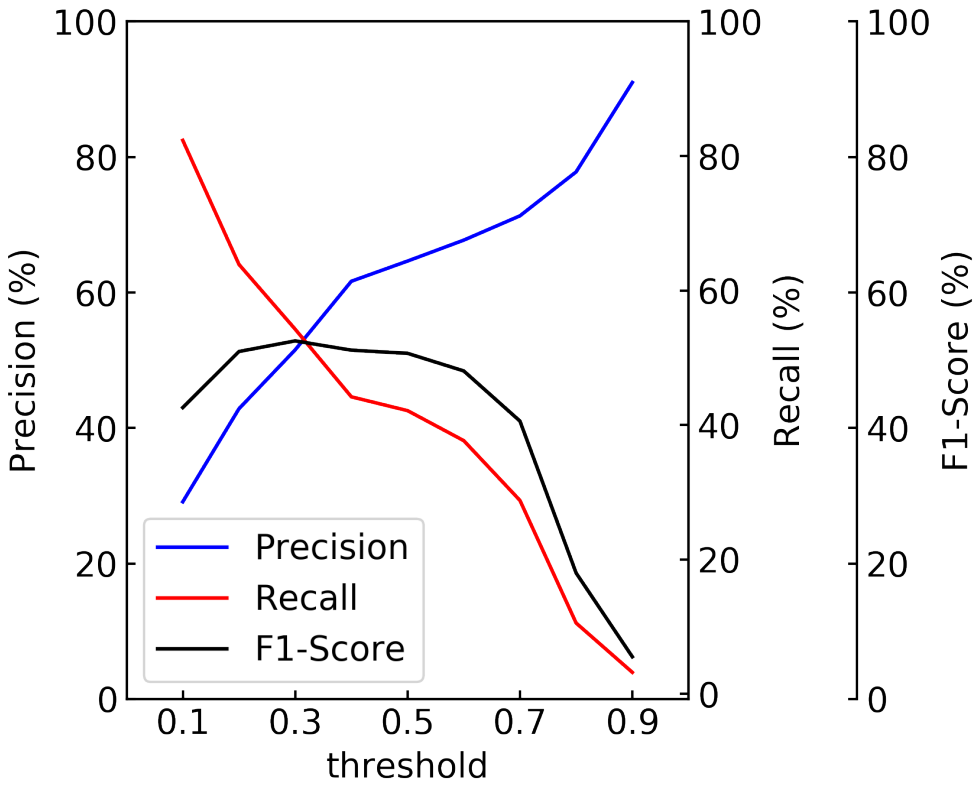


図 A.4: しきい値を変化させた際の適合率と再現率の変化

A.2.2 モデル３の結果

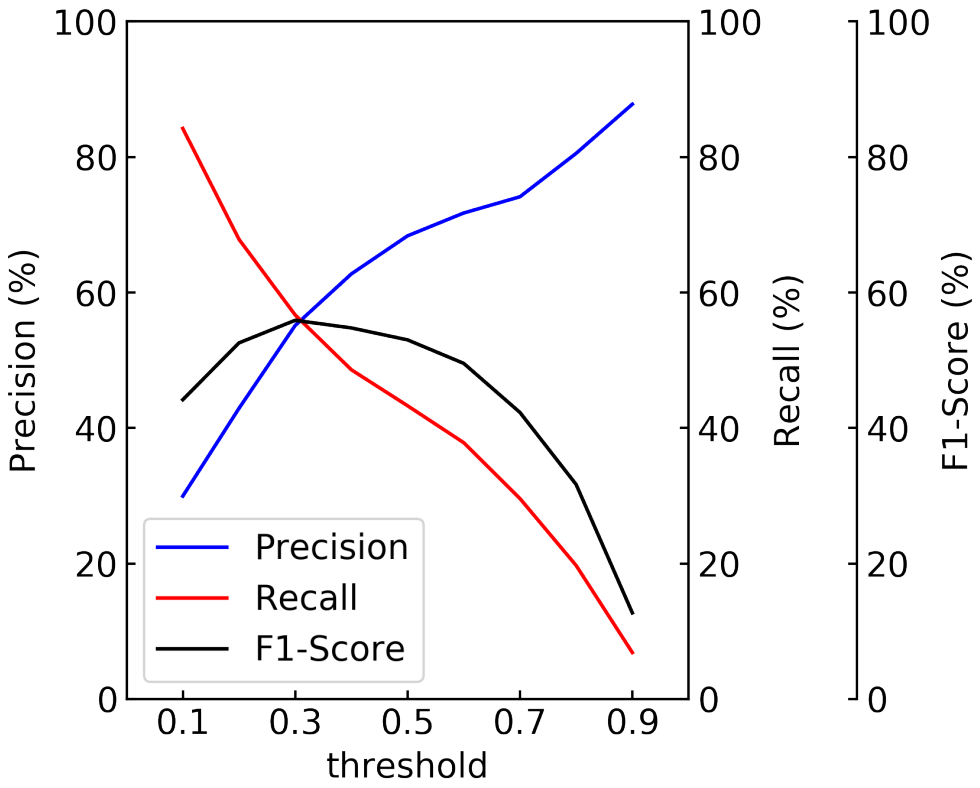


図 A.5: しきい値を変化させた際の適合率と再現率の変化

A.2.3 モデル４の結果

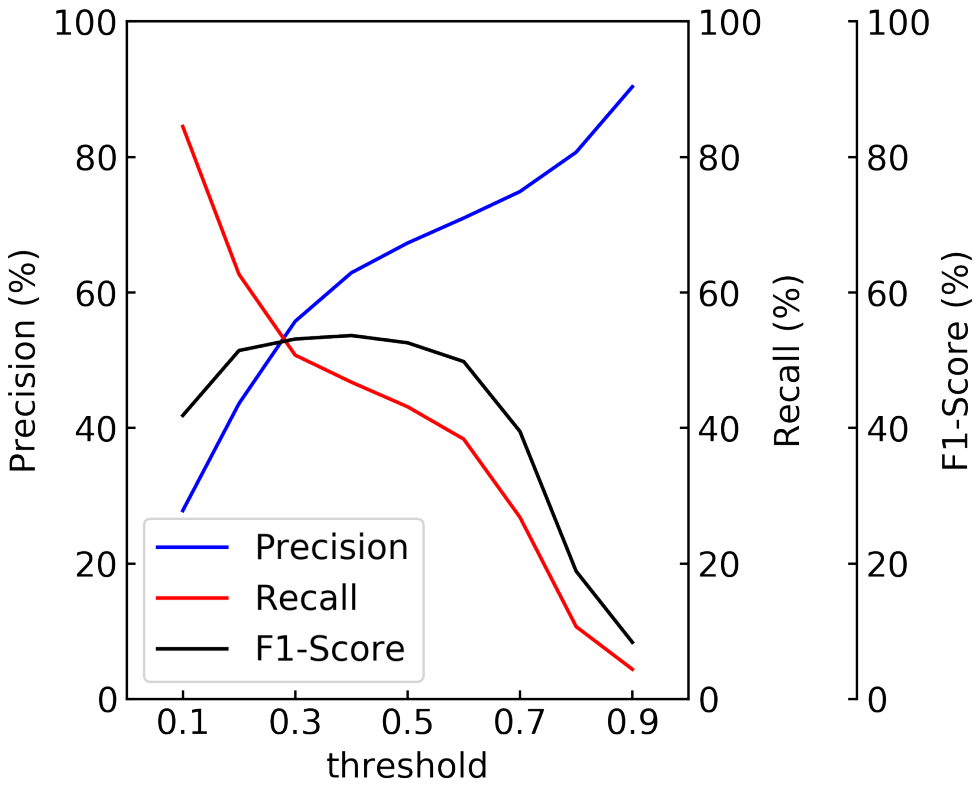


図 A.6: しきい値を変化させた際の適合率と再現率の変化