令和 7 年度

修士学位論文

LLM を用いた胸部 X 線画像および所見データに対するラベル付与時のノイズ軽減手法

A Noise Reduction Method for Label Annotation in Chest X-ray Images and Clinical Reports Using Large Language Models

有馬伊織

指導教員　吉田真一

2026 年 2 月 27 日

高知工科大学大学院 工学研究科 基盤工学専攻

情報学コース

要 旨

LLM を用いた胸部 X 線画像および所見データに対するラベル付与時のノイズ軽減手法

有馬伊織

本研究は，胸部 X 線画像診断支援モデルの学習における放射線読影レポート由来のラベルノイズ軽減を目的とし，大規模言語モデル（LLM）を用いた高品質な教師データの構築手法を提案した．既存の MeSH ベースのラベル抽出手法は，レポート内の否定表現や複雑な文脈を誤認し，ノイズを混入させる課題がある．本研究では，LLM による再判定において、存在確率 0.9 以上の高確信度な判定のみを採用するフィルタリングを適用した．その結果，データセット全体でラベル分布の適正化が行われ，特に曖昧な所見を含む「Other Finding」ラベルは、既存手法の 2,323 件から 258 件へと約 89%削減され，ラベルノイズの除去に成功した．このノイズ軽減がモデル学習に与える影響を検証するため，Vision Transformer（ViT）を用いた分類モデルを構築し，差が見られた Cardiomegaly（心拡大）と Calcinosis(石灰化) を分析した．Grad-CAM を用いた解析の結果，特定の症例では予測確信度が向上するとともに，注視領域が分散した肺野から解剖学的に妥当な部位へと集約されることを確認した．以上の成果は，LLM による精密なラベルクレンジングが，画像モデルの本質的な特徴学習を促進し，医療 AI の信頼性と説明責任を支える基盤技術として有効であることを示している．

キーワード　胸部 X 線画像，大規模言語モデル，Grad-CAM，ViT

Abstract

A Noise Reduction Method for Label Annotation in Chest X-ray Images and Clinical Reports Using Large Language Models

Arima Iori

This study aims to reduce label noise derived from radiology reading reports in the training of chest X-ray image diagnosis support models. We propose a method for constructing high-quality supervisory data using large language models (LLMs). Conventional MeSH-based label extraction methods suffer from label noise caused by misinterpretation of negation expressions and complex clinical contexts within reports. To address this issue, we applied an LLM-based re-evaluation process and introduced a filtering strategy that retains only predictions with high confidence scores (existence probability 0.9). As a result, the overall label distribution of the dataset was refined, and a substantial reduction in label noise was achieved. In particular, the ambiguous " Other Finding " label was reduced by approximately 89%, from 2,323 instances in the conventional method to 258 instances after filtering. To evaluate the impact of this noise reduction on model training, we constructed a classification model based on a Vision Transformer (ViT) and analyzed two representative labels, Cardiomegaly and Calcinosis, for which notable differences were observed. Grad-CAM analysis revealed that, in certain cases, prediction confidence increased and the model ' s attention shifted from diffusely distributed lung regions to anatomically plausible areas. These results demonstrate that precise label refinement using LLMs facilitates more meaningful fea-

ture learning in image-based models and serves as an effective foundational technology for enhancing the reliability and explainability of medical AI systems.

key words Chest X-ray images, large language model, Grad-CAM, ViT

目次

図目次

表目次

第 1 章

序論

胸部 X 線画像診断は，呼吸器や循環器疾患のスクリーニングにおいて極めて重要な役割を果たしている．近年，深層学習を用いた自動診断モデルの研究が盛んであるが，大規模な画像データセットに対し，専門医が全ての症例に正確な教師ラベルを付与することはコストの観点から困難である．そのため，放射線科医が作成した読影レポート（報告文）から，自動テキストマイニングツールを用いて付与された MeSH ラベルを教師データとして利用することが標準となっている [1]．しかし，既存の自動抽出ツールには，診断モデルの精度を阻害するラベルノイズの問題が存在している．従来の単語ベースの抽出アルゴリズムは，文章の医学的な文脈を解釈する能力が不十分であり，例えば，心拡大は認められない（No cardiomegaly）といった否定表現や，胸水の疑い（Suspected pleural effusion）といった不確定な推測表現を，単語の出現のみを根拠に一律で陽性と判定する傾向がある．このようなラベルノイズは，画像モデルに対して疾患が存在しない領域を疾患部位として学習させるという誤った信号を送り，モデルの識別精度と説明性を低下させる要因となる．近年では，大規模言語モデル（LLM）の高い自然言語理解能力を活用し，読影レポートの文脈を考慮したラベル生成を行う研究も報告されている [2][3][4][5][7]．これらの研究では，LLM を用いることで否定表現や推測表現を適切に解釈できる可能性が示唆されている．しかしながら，LLM によるラベル付与の信頼性評価や，ラベルの確信度を考慮した教師データ選別，さらに画像モデルの学習過程への影響まで踏み込んで検証した研究は十分ではない．本研究では，自然言語理解能力を有する大規模言語モデル（LLM）である GPT-4o mini を活用し，読影レポートの文脈を考慮した再ラベリング手法を提案する．特に，LLM が出力する判定確率に対して確信度 0.9 以上という閾値を設けることで，曖昧な判定を排除した教師データ

セットを構築する．さらに，ラベルの置換だけではなく，ラベルノイズが画像分類モデルの学習過程における疾患特徴の抽出能力に与える影響を，Grad-CAM[6] によって検証を行う．

第 2 章

関連技術

2.1 Mesh ラベル

MeSH（Medical Subject Headings）は，米国国立医学図書館（NLM）によって管理・維持されている，医学用語の階層構造を持つ語彙集である．医療現場や学術論文において，医師や研究者によって異なる表現がなされる医学概念を，単一の標準的な用語（MeSH ターム）に統合・整理することを目的としている [11]．例えば，心臓が大きくなっている，心陰影の拡大，Cardiomegaly といった多様な記述は，すべて MeSH タームである Cardiomegaly（心拡大）という一つのタグに紐付けられる．このように表記揺れを吸収し，情報の検索性やデータセットの構造化を向上させるための共通言語として機能している．本研究で対象とする IU データセットなどの大規模画像データセットにおいて，各症例に対して疾患ラベルを付与する際，すべてのレポートを医師が精査することは時間的・コスト的に困難である．そのため，読影レポート内に含まれる語彙を MeSH 辞書と照合し，特定のキーワードが検出された場合にその疾患ラベルを自動的に付与する手法が用いられてきた．これにより，構造化されていない自由記述のレポートから，機械学習モデルが学習可能な形式の教師データを生成することが可能となっている．しかし，キーワードの有無に依存する従来の MeSH 抽出手法には，画像分類モデルの精度を低下させるラベルノイズの原因が存在する．第一に，否定表現の誤認である．例えば，レポートに心拡大は認められないと記述されている場合，文章全体としては陰性（疾患なし）を意味するが，従来のシステムは心拡大という単語の存在のみに反応し，陽性ラベルを付与してしまう．第二に，不確定な表現の処理である．心拡大の疑いがある（suggestive of cardiomegaly）や，心拡大を否定できないといった，確定

的ではない医学的推論が含まれる場合，これらを陽性として扱うか，陰性として扱うかの判断が難しく，結果として不正確なラベルが混入する要因となる．図 2.1 に MeSH ラベルの概要図を載せる．

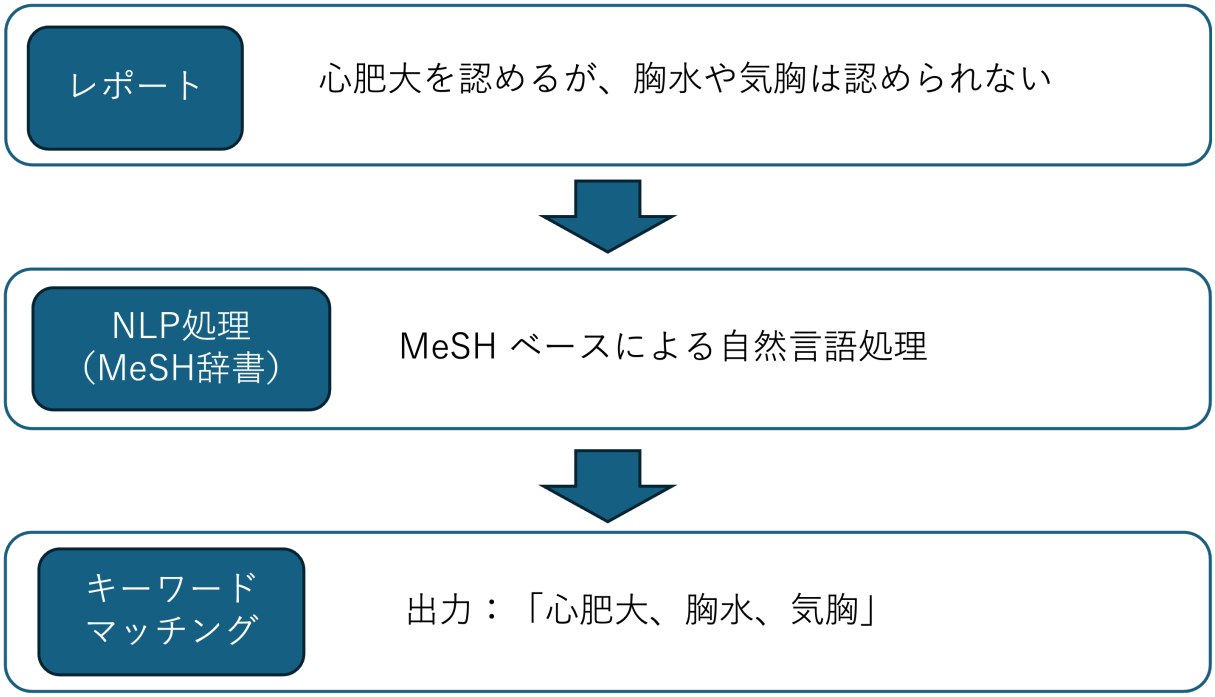


図 2.1: MeSH ラベル概要図

第 3 章

提案手法

このシステムは，放射線画像（本研究では胸部 X 線画像を用いている）とその読影レポートが入力データとして与えられ，可能性のある疾患情報を出力する診断支援システムの構築を想定している．しかし，このようなシステムを構築するためには，ラベルが正しく与えられている必要があるが，読影レポートのみでラベルがないデータも数多く存在する．それらを医師が全てラベル付けすることが現実的でないため，読影レポートから自動的に疾患ラベルを付与するシステムが提案されてきた．MeSH タグからラベルを抽出するものもその一つであるが，MeSH タグのような単語ベースの抽出手法には，否定表現や不確定な文脈を正しく解釈できず，ラベルノイズを混入させるという課題がある．そこで本研究では，大規模言語モデル（LLM）の文脈理解能力を活用し，従来の自動抽出手法におけるノイズを軽減した診断支援システムを提案する．また，学習したモデルに対して可視化手法である Grad-CAM を用いて説明性についても検討する．図 3.1 にシステムの概要図について示す．

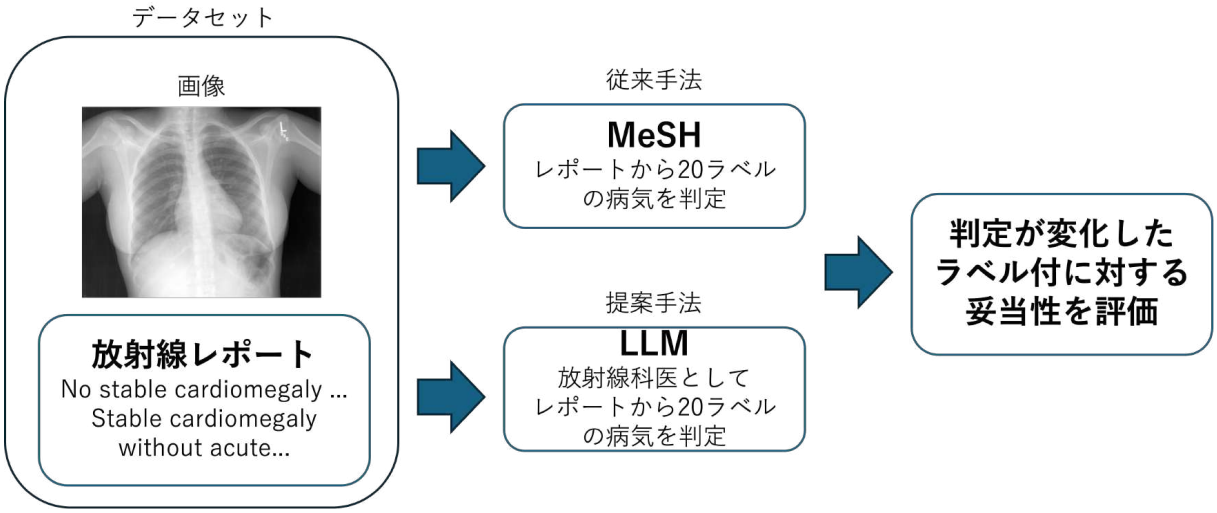


図 3.1: システム概要図

3.1　LLM による再ラベリングとコンテキスト理解

既存の MeSH ラベルが抱える否定表現の誤認や，文脈無視の問題を解決するため，推論能力を持つ GPT-4o mini を再ラベリングエンジンとして採用した． 具体的には，放射線報告文を入力とし，対象となる 20 の疾患ラベルそれぞれについて，医学的文脈に基づいた陽性・陰性の判定を指示するプロンプトを設計した．これにより，従来の単語マッチングでは困難であった,「心拡大の兆候はないが，軽度の胸水が認められる」といった，同一文章内での肯定・否定の混在を正確に識別することが可能となった．図 3.2 が実際の指示プロンプトである.LLM に対して放射線科の専門医という役割を与えることによって，医療用語の優先的な解釈や、診断レポート特有の定型句や省略表現に対する感度を高める [12]．また，数千件のデータをデータセットへ変換するために JSON 形式で指定し，20 ラベル以外の他の出力の制限を行なった．最後に，単語の有無だけではなく、否定・肯定・疑いに基づいたルー

ルを設けることで，従来の MeSH 手法では難しい意味の重み付けを可能とした [13].

```
あなたは胸部 X 線の放射線診断 AI です。
入力されたレポートを読み、以下の 20 項目について
「存在確率（0.0〜1.0)」を推定してください。

【非常に重要】
- 出力は JSON のみ。
- 下記の 20 ラベル以外を出力してはいけません。
- ラベル名は 1 文字たりとも変更してはいけません。
【出力すべき 20 ラベル】
Cardiomegaly, Scoliosis, Bone Fractures,
Pleural Effusion, Pleural Thickening,
Pneumothorax, Hernia, Calcinosis, Emphysema,
Pneumonia,Edema, Atelectasis, Cicatrix,
Opacity, Lesion, Airspace Disease,Hypoinflation,
Medical Device, Other Finding, Normal

【確率のルール】
- 明確に否定されている場合（no...） → 0.01〜0.05
- 明確に存在する → 0.8〜1.0
- 曖昧（possible, may represent） → 0.3〜0.6
- 記載なし → 0.0
```

図 3.2: 再ラベリングに使用した指示プロンプト

3.2　確信度スコアに基づくラベルフィルタリング

LLM の判定結果の信頼性をさらに担保するため，各判定に対する確信度を用いたフィルタリングを導入した．LLM に対し，各疾患の存在確率を 0.0 から 1.0 の範囲で算出させ，本研究では確信度 0.9 以上の判定のみを最終的な教師ラベルとして採用した．この閾値設定は，判定が曖昧な症例を学習データから除外することで，教師データセットの純度を高めることを目的としている．

3.3　ラベルノイズクレンジングデータの学習への適用

ノイズが除去されたラベルを用い，画像分類モデルとして Vision Transformer (ViT) を学習させた． 学習プロセスでは，既存の MeSH ラベルで学習したモデルと，本提案手法でラベルクレンジングしたラベルで学習したモデルを同一のアーキテクチャ・ハイパーパラメータで構築した．これにより，モデルの性能差や識別根拠の変化が，純粋にラベルの質に起因するものであることを比較検証できる設計とした．

3.4　Grad-CAM による識別根拠の定性評価

ラベルノイズの軽減がモデルの挙動に与える影響を視覚化するため，Grad-CAM を用いた解析を行った．具体的には，各ラベルにおいて MeSH と LLM で判定が一致した症例，および LLM と不一致の症例を抽出し，モデルが画像上のどの領域（解剖学的特徴）を根拠に予測を行ったかを精査した．

第 4 章

実験設定

4.1　データセット

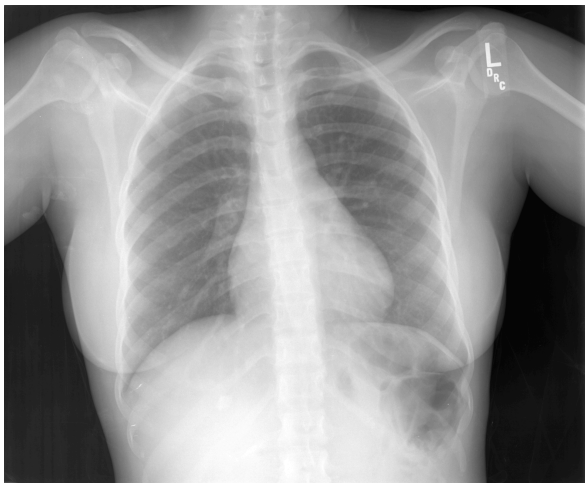
本研究では，インディアナ大学が公開しているオープンデータセット（IU-dataset）を使用した [10]．本データセットは，胸部 X 線画像（正面および側面）と，それに対応する放射線科医による読影レポートで構成されている．使用した画像の代表例を図 4.1 に示す．対象疾患として Cardiomegaly（心拡大），Scoliosis（側弯症），Bone Fractures（骨折），Pleural Effusion（胸水），Pleural Thickening（胸膜肥厚），Pneumothorax（気胸），Hernia（ヘルニア），Calcinosis（石灰化），Emphysema（肺気腫），Pneumonia（肺炎），Edema（浮腫），Atelectasis（無気肺），Cicatrix（瘢痕），Opacity（陰影），Lesion（病変），Airspace Disease（肺胞性病変），Hypoinflation（低肺膨張），Medical Device（医療機器），Other Finding（その他の所見），Normal（正常）の計 20 のラベルを分析対象とした．既存ラベルとしてデータセットに付随する MeSH（Medical Subject Headings）タームを比較対象の初期ラベルとして使用しており，各ラベルの陽性症例数の内訳は表 4.1 の通りである．IU-dataset から抽出した正面画像および読影レポートのペア，計 3,689 件を解析対象とした．モデルの学習および評価の妥当性を担保するため，全症例を学習（Train），検証（Validation），テスト（Test）の 3 セットに分割した．分割比率は 7 : 1.5 : 1.5 とした．学習データ 2,582 件，検証データ 553 件とし，Grad-CAM を用いた識別根拠の解析は，学習に一切使用していない独立したテストデータ 554 件に対して実施した．表 4.2 にデータセットの内訳を示す．

表 4.1: データセットにおける各ラベルの陽性数

ラベル	陽性数
Cardiomegaly	326
Scoliosis	88
Bone Fractures	83
Pleural Effusion	136
Pleural Thickening	46
Pneumothorax	25
Hernia	44
Calcinosis	282
Emphysema	113
Pneumonia	37
Edema	41
Atelectasis	294
Cicatrix	186
Opacity	408
Lesion	103
Airspace Disease	123
Hypoinflation	264
Medical Device	153
Other Finding	258
Normal	1837

表 4.2: データセットの分割と構成（分割比 7:1.5:1.5）

データセット区分	分割比率 (%)	症例数 (Images/Reports)
学習 (Train)	70%	2582
検証 (Validation)	15%	553
テスト (Test)	15%	554
合計	100%	3,689



正面像

側面像

図 4.1: 使用した胸部 X 線画像

4.2　学習モデルのアーキテクチャとハイパーパラメータ

本研究では，画像分類のモデルとして，医療画像解析において高い汎用性と性能を示す Vision Transformer (ViT-base/16) を採用した [8][9]．ViT は画像をパッチ単位で処理し，Self-Attention 機構を用いることで，従来の畳み込みニューラルネットワーク（CNN）よりも大域的な特徴抽出を可能とする．入力画像は 224×224 ピクセルにリサイズし，ImageNet-1k で事前学習された重みを初期値として転移学習を行った．また，医療画像特有のマルチラベル分類に対応するため，出力層にはシグモイド関数を適用した．比較の公平性を担保するため，既存の MeSH ラベルを用いた Baseline モデルと，提案手法によるラベルクレンジングを用いたモデルにおいて，最適化アルゴリズムや学習率等の全てのハイパーパラメータを共通化している．モデルの詳細および学習時のパラメータ設定を表 4.3 に示す．

表 4.3: 学習モデルの仕様およびハイパーパラメータの設定

項目	設定値 / 内容
Base Model	Vision Transformer (ViT-base/16)
Input Resolution	224×224 pixels
Pre-trained weights	ImageNet-1k
Optimizer	Adam
Learning Rate	1×10^{-5}
Batch Size	8
Loss Function	Binary Cross-Entropy Loss
Activation (Output)	Sigmoid (Multi-label)

4.3　評価指標および可視化手法

モデルの性能と識別根拠の妥当性を検証するため，表 4.4 に示す 3 つの観点から分析を実施した．まず, LLM によるラベルクレンジングの効果を定量化するため, MeSH ラベルとの陽性件数の推移を比較した (ラベル純度評価). 次に, モデルが予測を行う際の自信度を評

価するため, テストデータに対する予測確信度を計測した．最後に, Grad-CAM を用いた識別根拠の可視化を行った．具体的には, モデルが判定に寄与したと判断した画像領域をヒートマップ化し, 放射線科医の解剖学的知見に基づいて, その妥当性を改善・悪化・どちらでもないの 3 段階で定性的に評価した．心拡大（Cardiomegaly）症例においては，特に心臓境界部に注意が集まっているか，石灰化（Carcinosis）症例においては肺野領域に注意が集まっているのか評価した．図 4.4 に Grad-CAM における評価の流れを示す．

表 4.4: 評価指標と解析アプローチの概要

評価項目	**具体的な解析手法**
ラベル純度評価	MeSH と LLM 間での疾患別陽性数の変化を解析。
予測確信度解析	予測確率の変動による識別精度の質的評価。
識別根拠の可視化	Grad-CAM による注視領域の特定と解剖学的妥当性の評価。

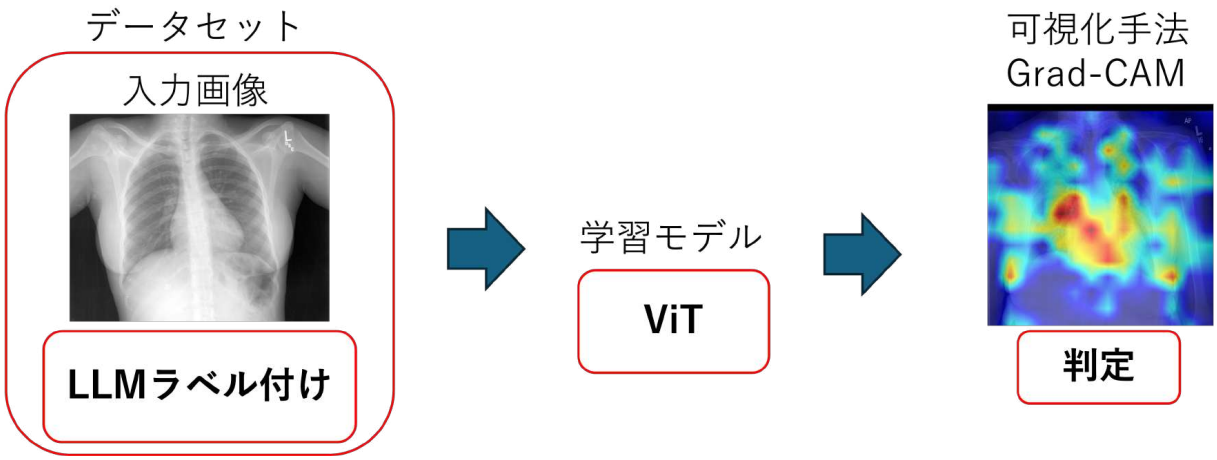


図 4.2: Grad-CAM 評価

第 5 章

実験結果

5.1　読影レポート解析による判定精度の評価

本節では, 既存の MeSH ラベルと提案する LLM ラベルの間で判定が乖離した症例を抽出し, その要因を解析した結果を述べる．解析の結果, LLM は医学的文脈を理解することで, 既存手法における否定表現の誤認および，複雑な診断表現の見落としを改善していることが確認された.

5.1.1　既存手法における偽陽性の除去

表 5.1 に, MeSH が陽性と誤判定した症例の解析結果を示す．既存の MeSH ラベルでは, Cardiomegaly や Bone fracture 等の症例において, 文中に「 no stable...」や「no definite...」といった明確な否定語が含まれているにもかかわらず，疾患名そのものの出現に反応して陽性と判定されていた．これは，単語の出現頻度や位置に依存する従来のテキストマイニング手法における典型的な偽陽性のパターンである．対して, 提案手法である LLM は, 文頭の否定語が文全体の意味を規定していることを正確に解釈し, 疾患を陰性と判定した．また，Medical Device の例に見られるように，モニターのリード線（Leads）などの臨床所見とは無関係な単語に対する誤反応も排除されており，ラベルノイズが減少した.

表 5.1: 否定表現・ノイズの誤解による偽陽性の除去

対象疾患	読影レポートの抜粋（重要箇所）	MeSH	LLM
Cardiomegaly	No stable **cardiomegaly**, ... Stable **cardiomegaly** without acute...	1	0
Bone Fracture	No displaced rib **fractures** are visualized.	1	0
Pleural Thickening	...Stable bilateral apical **pleural capping**.	1	0
Pleural Effusion	...no definite **pleural effusion** seen.	1	0
Opacity	...may represent scarring alternatively small **pleural effusions**.	1	0
Medical Device	...overlying external cardiac monitor **leads**.	1	0
Lesion	**Negative** for acute cardiopulmonary disease.	1	0
Hypoinflation	...limited assessment of heart size due to **obscured** heart.	1	0
Cicatrix	...Mild hypoinflation **without acute disease**.	1	0
Other Finding	...dextroscoliosis of the thoracic spine.	1	0

5.1.2　既存手法における偽陰性の救済

表 5.2 に，MeSH が陰性と誤判定した症例の解析結果を示す．Pneumonia や Airspace Disease の症例では,「suggestive of（〜を示唆する）」や「consistent with（〜と矛盾しない）」といった，臨床的に重要ながら断定を避けた表現に対し，MeSH が反応せず陰性と判定していた．一方，LLM はこれらの医学的ニュアンスを解釈し，実質的な陽性所見としてラベルを付与することに成功した．さらに，Emphysema の症例において COPD という記述から関連疾患を同定するなど，医学的知識に基づいた柔軟なラベリングが行われており，従来手法では難しい偽陰性の改善がされた．

表 5.2: 複雑な文脈・示唆的表現の解釈による偽陰性の救済

対象疾患	読影レポートの抜粋（重要箇所）	MeSH	LLM
Pneumonia	Right lower lobe infiltrate, **suggestive of pneumonia**	0	1
Calcinosis	Multiple **calcified granulomas** in the bilateral...	0	1
Atelectasis	...small left basilar **scar**. (COPD に伴う無気肺所見)	0	1
Airspace Disease	Patchy bilateral opacities, ...**consistent with pneumonia**.	0	1
Normal	**No** acute cardiopulmonary **abnormalities**. (正常として定義)	0	1
Emphysema	**COPD** with no acute findings. (疾患概念の包含)	0	1

5.2　既存 MeSH ラベルと提案手法によるラベルクレンジングの定量的比較

5.2.1　ラベルクレンジングによるノイズ除去数の遷移

GPT-4o mini による再ラベリングの結果，多くの疾患カテゴリでノイズ除去およびラベルの適正化が確認された．全ラベル数 6429 件に対し，ノイズクレンジング数が 2286 件であり，36%のラベルノイズを削減した．表 5.3 における Other Finding では，1,800 件を超えるノイズが除去され，Normal においても 385 件の見逃しが補完されるなど，データセット全体のラベルノイズが減少しました．また，多くの項目で修正成功率が 100%またはそれに近い数値を示しており，GPT-4o mini が医療報告文の否定表現や文脈を正確に理解できていることが示されました．

表 5.3: IU-Xray 全 20 ラベルにおける MeSH ラベルと LLM ラベルの比較および修正成功率の評価

疾患ラベル	MeSH 数 (A)	LLM 数 (B)	差分 (B-A)	修正成功率	推計ノイズ除去数
Other Finding	2,323	258	-2,065	88.0%	1,816 件
Normal	1,348	1,837	+489	77.0%	385 件
Opacity	409	408	-1	100%	1 件
Cardiomegaly	319	326	+7	100%	9 件
Atelectasis	295	294	-1	100%	1 件
Calcinosis	280	282	+2	83.3%	5 件
Hypoinflation	265	264	-1	100%	1 件
Cicatrix	192	186	-6	83.3%	5 件
Medical Device	153	153	0	96.2%	25 件
Pleural Effusion	138	136	-2	100%	2 件
Airspace Disease	122	123	+1	100%	1 件
Lesion	115	103	-12	100%	12 件
Emphysema	101	113	+12	100%	14 件
Scoliosis	88	88	0	-	0 件
Bone Fractures	83	83	0	100%	2 件
Pleural Thickening	52	46	-6	100%	6 件
Pneumothorax	25	25	0	-	0 件
Hernia	44	44	0	-	0 件
Edema	41	41	0	-	0 件
Pneumonia	36	37	+1	100%	1 件

5.2.2　ViT による AUC 比較

LLM によるクレンジングを行なったラベルと，MeSH ラベルとの画像診断モデルでの性能比較を行った．表 5.4 に 20 回分のモデル全体平均 AUC を示した．結果として，MeSH ラベル LLM ラベルどちらも 0.94 と同じ平均 AUC となった．しかし，ラベル別平均 AUC では，図 5.1 に 20 ラベルの平均 AUC を示しており，Cardiomegaly,Normal,Lesion におい

て 2%の向上が見られ，Other Finding においては 9%の低下が見られた．

表 5.4: 20 回分モデル全体平均 AUC

ラベル	平均 AUC
MeSH	0.94(94%)
LLM	0.94(94%)

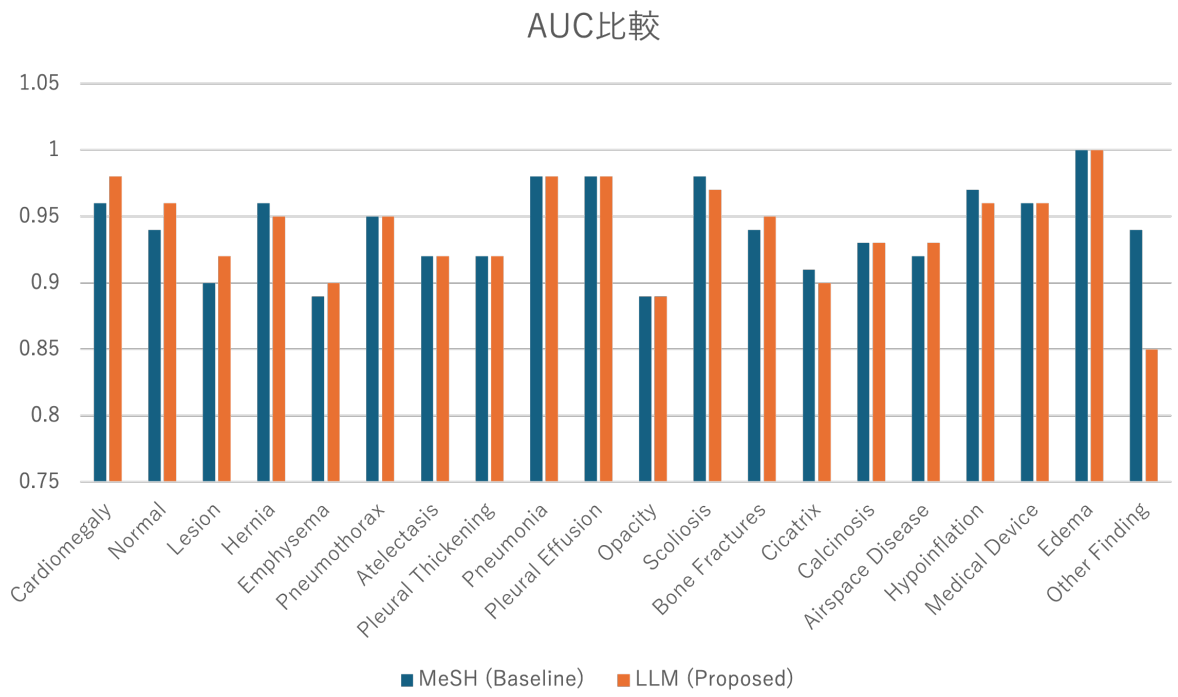


図 5.1: AUC 比較

5.3　可視化による識別根拠の妥当性評価

LLM によるクレンジングを行なったラベルと，MeSH ラベルで学習した ViT を用いてテスト画像に Grad-CAM を適用した．その際に，判定が一致、不一致のものを改善・悪化・どちらでもないで評価を行なった．以下に具体的な判定基準について示す．

- 改善：モデルが疾患の特徴を捉えるようになった状態

- 悪化：注目領域が特徴から遠ざかった状態
- どちらでもない：判定に有意な変化が見られない状態

5.3.1　Grad-CAM における各ラベルごとの評価

図 5.2 にはラベル数の多い症例についての評価を行なっており，Normal，Other Finding においては，改善した数が悪化した数より多くなっていることが確認できた．次に図 5.3 では，ラベル数が少ない症例の評価を行なっており，評価が MeSH ラベルより悪くなっている症例もあるが，ほとんどのラベルにおいて改善していることが確認できた．

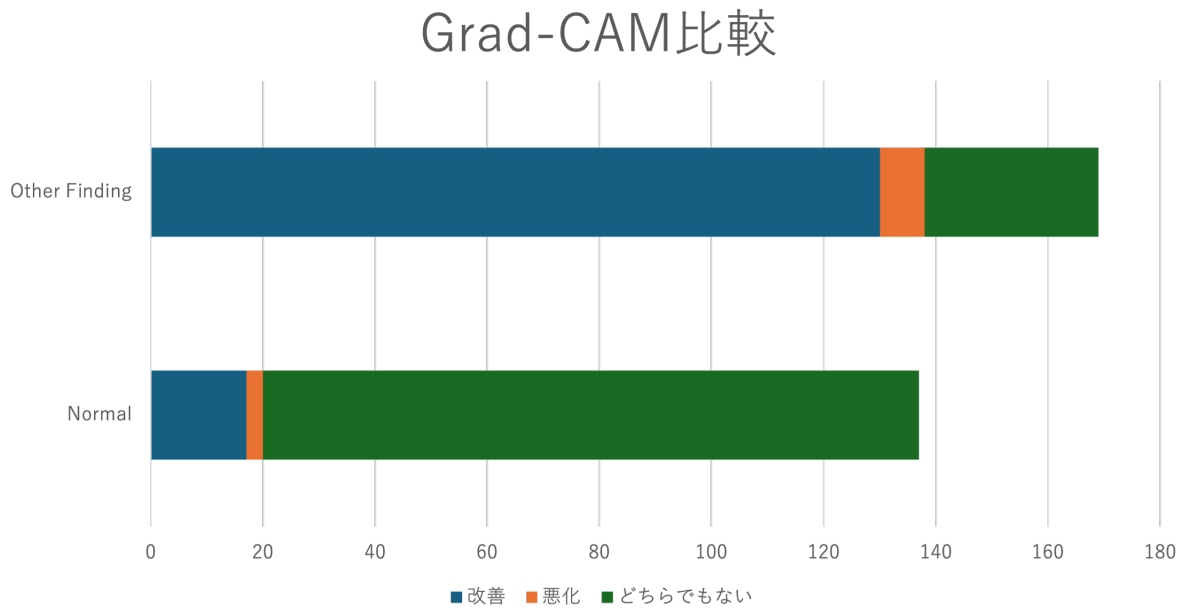


図 5.2: ラベル数の多い症例評価

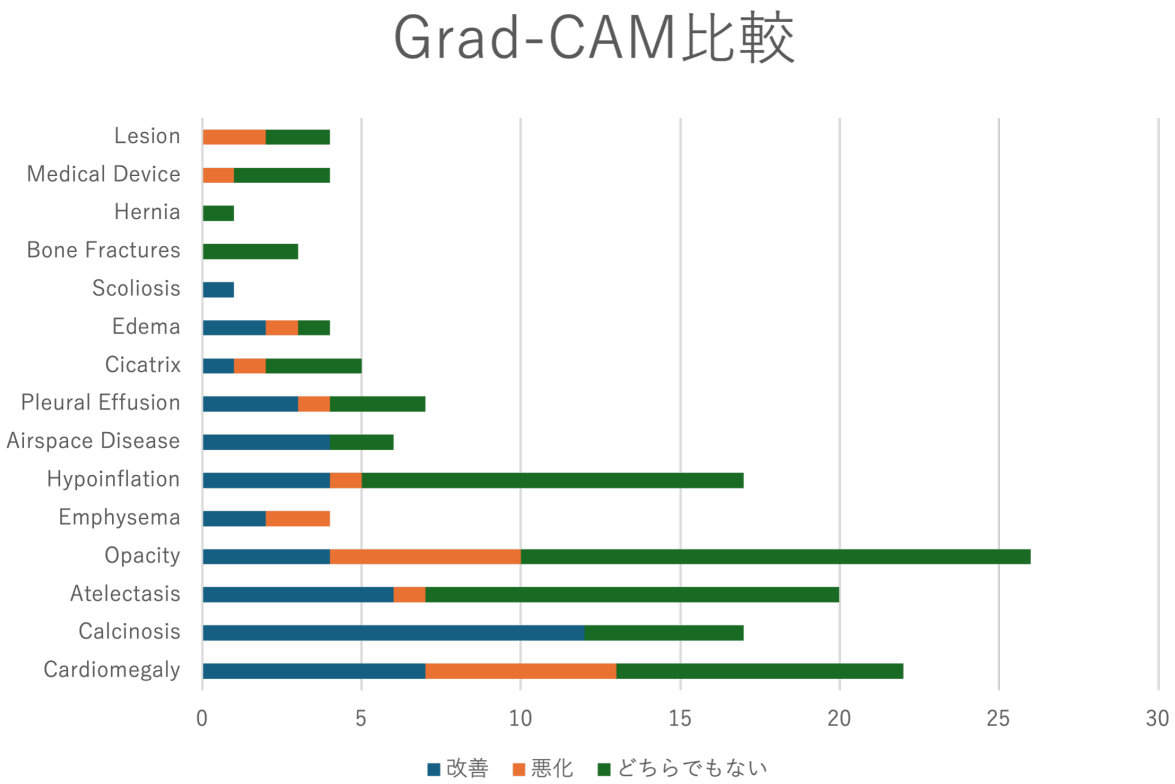


図 5.3: ラベル数が少ない症例評価

可視化結果の定性的評価として，特に改善が顕著であった心拡大（Cardiomegaly）および石灰化（Calcinosis）について詳細な分析を行う．

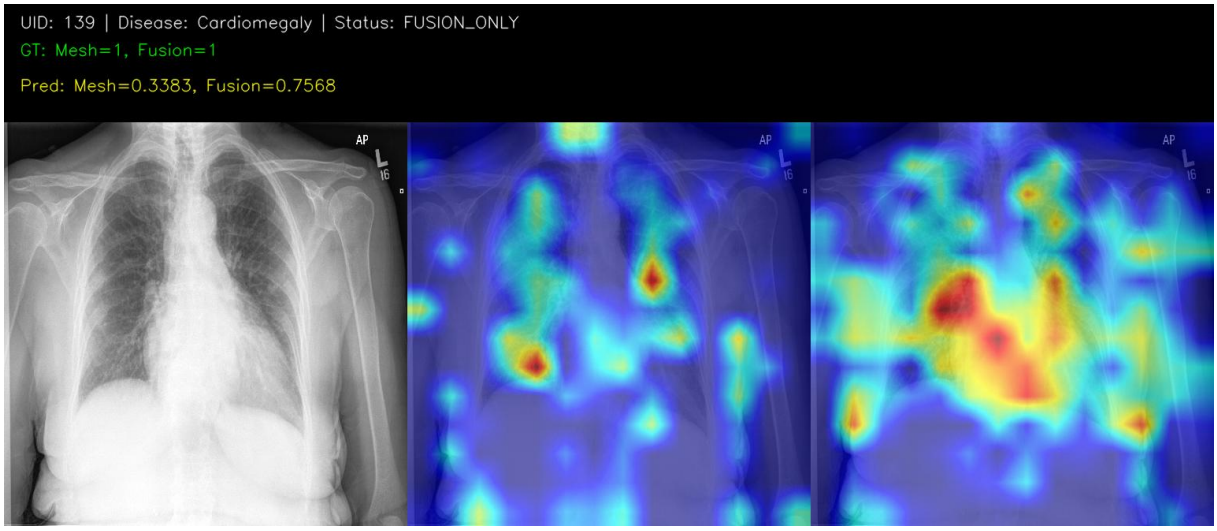
5.3.2　心拡大（Cardiomegaly）における識別根拠の変化

表 5.5 に示す通り，Cardiomegaly 症例では 22 症例中 7 症例で注視領域の明確な改善が確認された．図 5.4（UID:139 および 177）に見られるように，Baseline モデルでは注視点が肺野や肩関節付近に分散していたのに対し，提案手法では心臓境界部（心尖部および右心縁）へと集約されている．これは，LLM が否定表現や他疾患との混同を排除したことで，モデルが「心胸比の拡大」という疾患の本質的な特徴を学習できた結果と考えられる．一方で，6 症例で「悪化」と評価された要因については，図 5.5（UID:227 および 569）で見られる．UID:227 では，心拡大に加えて「pulmonary interstitial edema（間質性肺水腫）」に伴

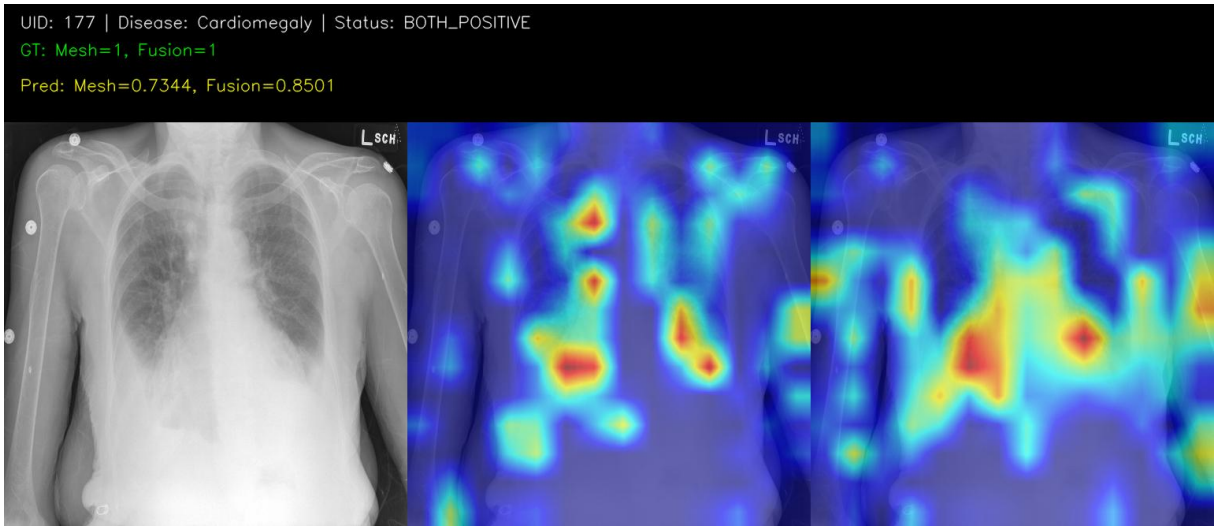
う両側肺底部の透過性低下（opacities）が重畳しており，心境界の視認性を著しく低下させていた．また，UID:569 は「Extensive post-op changes（広範な術後変化）」や「Surgical clips」といった人工物が存在し，さらに「Right pleural densities」などの複雑な術後所見が混在していた症例である．これらの分析結果は，LLM によるテキストレベルのラベルクレンジングが成功しても，画像上に強いノイズ（病変の重畳や術後変化）が存在する場合，画像モデルが疾患特有の幾何学的特徴を同定することが困難になるという，画像認識上の根本的な課題を明示している．

表 5.5: Cardiomegaly 可視化結果の定量的内訳

判定パターン	改善	どちらでもない	悪化	合計
一致	6	9	5	20
不一致	1	0	1	2
合計	7	9	6	22

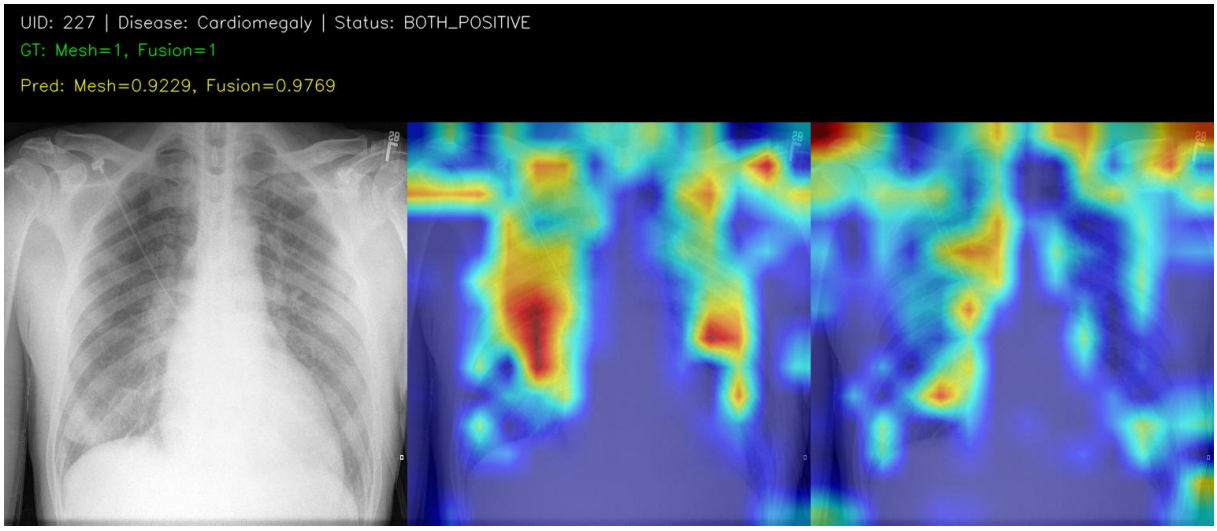


(a) 改善例：注視領域が心臓境界部へ集約し、確信度が向上した症例

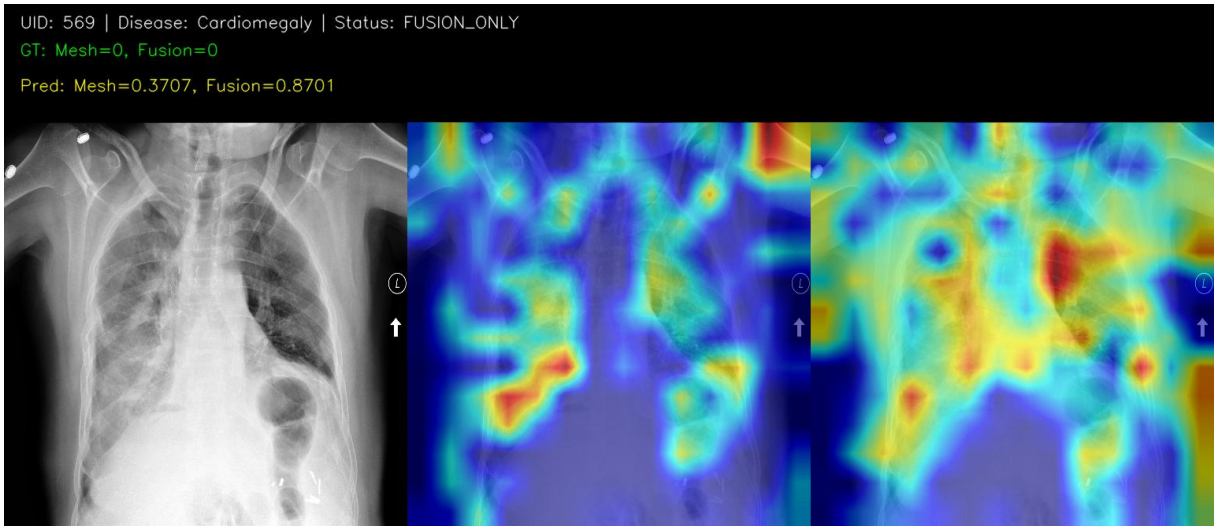


(b) 改善例：注視領域が心臓境界部へ集約し、確信度が向上した症例

図 5.4: Cardiomegaly 症例における Grad-CAM の比較（左：元画像、中：MeSH 学習モデル、右：提案手法学習モデル



(a) 悪化例：画像のコントラストや他疾患の重畳により識別が困難となった症例



(b) 悪化例：画像のコントラストや他疾患の重畳により識別が困難となった症例

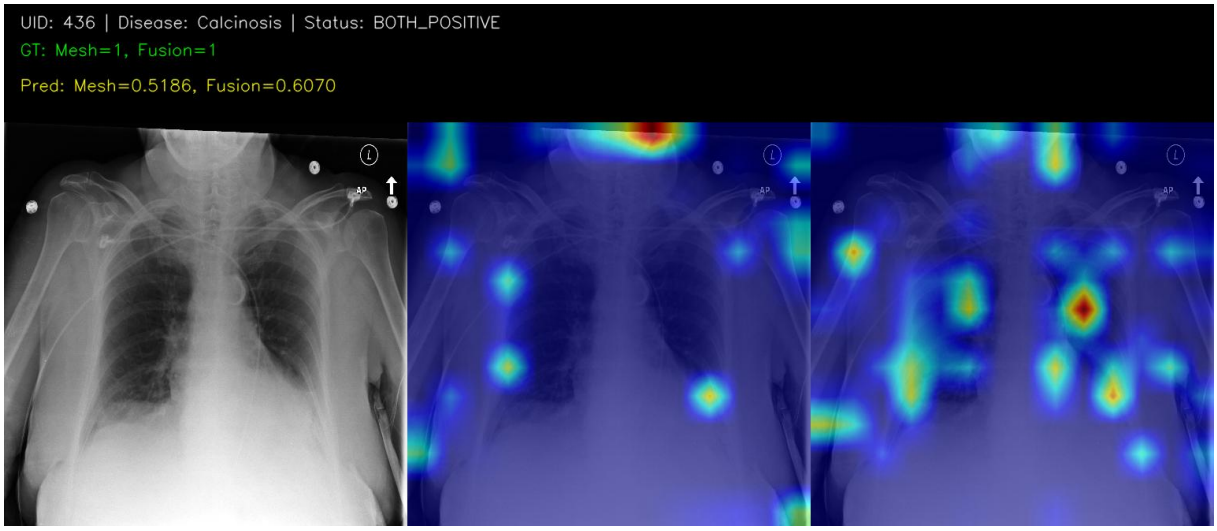
図 5.5: Cardiomegaly 症例における Grad-CAM の比較（左：元画像、中：MeSH 学習モデル、右：提案手法学習モデル

5.3.3　石灰化（Calcinosis）における識別根拠の改善

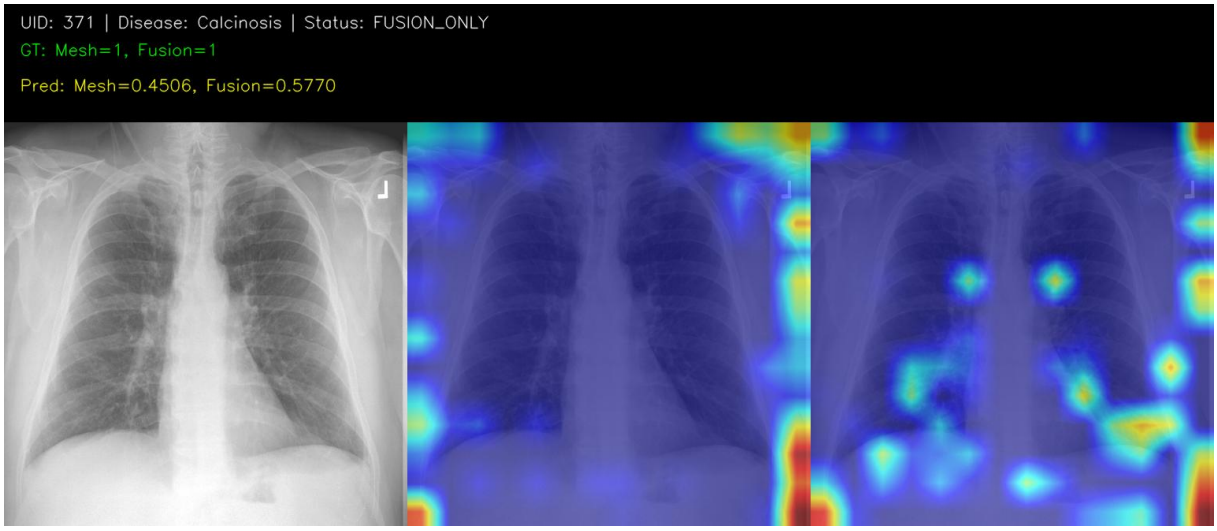
表 5.6 に示す通り，Calcinosis においては 17 症例中 12 症例で改善が認められ，悪化は 0 件であった．図 5.6（UID:436 および 371）の比較すると，MeSH モデルでは画像全体を漠然と注視していたのに対し，提案手法では肺野内の結節状の石灰化病変を正確に捉えている．Calcinosis は MeSH ラベルにおいて偽陰性（見落とし）が多い疾患であったが（表 5.2 参照），LLM がレポート文脈からこれらの微細な所見を正しくラベル化したことが，モデルの空間的な認識精度を向上させたと推察される．

表 5.6: Calcinosis 可視化結果の定量的内訳

判定パターン	改善	維持	悪化	合計
一致	9	2	0	11
不一致	3	3	0	6
合計	12	5	0	17



(a) 改善例：注視領域が肺へ移行し、確信度が向上した症例



(b) 改善例：注視領域が肺へ移行し、確信度が向上した症例

図 5.6: Calcinosis 症例における Grad-CAM の比較（左：元画像、中：MeSH 学習モデル、右：提案手法学習モデル

第 6 章

考察

6.1　ラベルクレンジングによる識別能力向上のメカニズム

LLM を用いた読影レポートの文脈解析により，従来の MeSH ラベルで発生していたノイズを修正するプロセスが，画像分類モデル（ViT）の識別根拠（Grad-CAM）の妥当性を劇的に向上させた要因について，以下の 3 点から考察する．第 1 に，否定表現の解釈による偽陽性の排除が，モデルの学習における誤った相関を断ち切ったと考えられる． 第 5.2 節の表 5.1 に示した通り，MeSH ラベルでは否定語を無視して疾患名を陽性としていたため，モデルは疾患が存在しない肺野を疾患の根拠として学習していた可能性がある．図 5.4 の Baseline モデルにおいて，心拡大症例の注視点が肺野全域に分散していたのは，このようなラベルノイズにより，疾患と解剖学的構造の正しい対応関係を構築できなかったためと推察される．LLM によるラベルクレンジングは，モデルが真の病変部位（心臓境界部など）のみを注視するための正しい教育データとして機能したと言える．第 2 に，MeSH ラベルと LLM ラベルでの AUC 比較では心臓や肺の広範囲に病変が見られるものが性能が向上する可能性があるが，局所的であり、細かい部分に関しては性能が落ちる可能性を考えられる．Other Finding ラベルにおいては、1800 件ほどノイズを除去したためデータ数の減少による性能低下が起こったと考える．第 3 に，示唆的表現の救済による学習効率の向上である．Calcinosis（石灰化）において高い改善率（表 5.6）が見られた．MeSH が見落としていた微細な所見を，LLM が「suggestive of」や関連語から拾い上げたことで，モデルはこれまで正常と定義されていた画像の中から微小な病変の特徴を効率的に抽出できるようになった．これが，Grad-CAM における集約性と，予測確信度の向上に寄与した．

6.2　現状の限界と今後の課題

一方で，Cardiomegaly 症例において一部で認められた悪化例（UID:227 等）は，本手法の限界と今後の課題を示唆している．これらの症例では，他疾患（胸水や浸潤影）の重畳により解剖学的境界が物理的に不明瞭であった．これは，テキスト情報のノイズ除去だけでは解決できない画像自体の複雑性に起因する限界である． 今後は，単一のラベルクレンジングだけでなく，LLM が抽出した所見の位置情報（右下肺野、心臓周囲など）を画像モデルの Attention メカニズムに直接関与するする，より高度なマルチモーダル学習アルゴリズムの構築が期待される．

第 7 章

結論

本研究では，大規模言語モデル（LLM）を用いた胸部 X 線読影レポートの文脈解析によるラベルノイズ軽減手法を提案し，その有効性を定量的および定性的に検証した．従来の MeSH ベースのラベリング手法では，否定表現の誤認や複雑な医学的表現の解釈不足に起因する深刻なラベルノイズが課題であった．これに対し，LLM（GPT-4o mini）を活用することで，文脈に即した高精度なラベル付与を実現した．その結果，Other Finding（その他所見）に分類されていたノイズを約 89%削減することに成功し，データセットのラベルノイズを削減することができた．さらに，ラベルノイズが除去されたデータセットを用いて学習した画像分類モデルの識別根拠を Grad-CAM により解析した結果，心拡大 (Caldiomegaly) や石灰化 (Calcinosis) といった疾患において，注視領域が解剖学的に妥当な部位へ集約されることを確認した．特に，Baseline モデルで肺野に分散していた注視点が疾患特有の境界部へと修正されたことは，ラベルの質的向上がモデルの「本質的な特徴学習」に直結することを示唆している．本研究の成果は，膨大な未構造データが存在する医療分野において，LLM が胸部 X 線読影レポートのラベル付として機能し，信頼性の高い医療診断支援 AI の構築に寄与できる可能性を実証したものである．

謝辞

本研究を進める上でご指導をしてくださった吉田真一教授に心より感謝申し上げます．研究の方向性に思い悩んだ際や，思うような結果が出ずに不安に陥った際，吉田先生の助言のお陰で，修士論文の完成まで辿り着くことができました．また， ご多忙中にもかかわらず副査を引き受けてくださった妻鳥貴彦准教授と敷田幹文教授に感謝致します．

参考文献

[1] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, “When Radiology Report Generation Meets Knowledge Graph,” AAAI, vol. 34, no. 07, pp. 12910–12917, Apr. 2020, doi: 10.1609/aaai.v34i07.6989.

[2] K. Singhal et al., “Large language models encode clinical knowledge,” Nature, vol. 620, no. 7972, pp. 172–180, Aug. 2023, doi: 10.1038/s41586-023-06291-2.

[3] J. Zhou et al., “Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4,” Nat Commun, vol. 15, no. 1, p. 5649, Jul. 2024, doi: 10.1038/s41467-024-50043-3.

[4] S. M. Santomartino, J. R. Zech, K. Hall, J. Jeudy, V. Parekh, and P. H. Yi, “Evaluating the Performance and Bias of Natural Language Processing Tools in Labeling Chest Radiograph Reports,” Radiology, vol. 313, no. 1, p. e232746, Oct. 2024, doi: 10.1148/radiol.232746.

[5] G. Leonardi, L. Portinale, and A. Santomauro, “Enhancing Medical Image Report Generation through Standard Language Models: Leveraging the Power of LLMs in Healthcare”.

[6] A. Alqutayfi et al., “Explainable Disease Classification: Exploring Grad-CAM Analysis of CNNs and ViTs,” JAIT, vol. 16, no. 2, pp. 264–273, 2025, doi: 10.12720/jait.16.2.264-273.

[7] T. B. Brown et al., “Language Models are Few-Shot Learners,” Jul. 22, 2020, arXiv: arXiv:2005.14165. doi: 10.48550/arXiv.2005.14165.

[8] A. Vaswani et al., “Attention Is All You Need,” Aug. 02, 2023, arXiv: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[9] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for

Image Recognition at Scale," Jun. 03, 2021, arXiv: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[10] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the Amer- ican Medical Informatics Association, 23(2):304–310, 07 2015. ISSN 1067-5027. doi: 10.1093/jamia/ocv080.

[11] Yuan, J.; Liao, H.; Luo, R.; and Luo, J. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. arXiv preprint arXiv:1907.09085.

[12] S. Hendawi, T. Kanan, M. Elbes, A. Mughaid, and S. AlZu'bi, "Automated Prompt Engineering Pipelines: Fine-Tuning Llms for Enhanced Response Accuracy," 2024, SSRN. doi: 10.2139/ssrn.5004054.

[13] W. Xie, G. Gwizdz, and D. Feng, "Prompting a Weighting Mechanism into LLM-as-a-Judge in Two-Step: A Case Study," Feb. 19, 2025, arXiv: arXiv:2502.13396. doi: 10.48550/arXiv.2502.13396.

付録 A

表 A.1: 17 疾患ラベル別 Grad-CAM 識別根拠の定性評価結果（網羅版）

疾患カテゴリ	疾患ラベル	ラベル一致状況	改善	維持	悪化	合計
心臓	**Cardiomegaly**	一致	6	9	5	20
		不一致	1	0	1	2
肺野	**Emphysema**	一致	0	2	0	2
		不一致	2	0	0	2
	Edema	一致	2	1	1	4
	Atelectasis	一致	5	9	1	15
		不一致	1	4	0	5
	Cicatrix	一致	1	3	1	5
	Opacity	一致	3	14	6	23
		不一致	1	1	0	2
	Lesion	不一致	0	1	1	2
	Airspace Disease	一致	4	2	0	6
	Hypoinflation	一致	2	11	1	14
		不一致	2	1	0	3
胸膜	**Pleural Effusion**	一致	3	2	1	6
		不一致	0	1	0	1
骨	**Scoliosis**	一致	1	0	0	1
	Bone Fractures	一致	0	1	0	1
		不一致	0	2	0	2
横隔膜・腹部	**Hernia**	不一致	0	1	0	1
医療機器	**Medical Device**	一致	0	1	0	1
		不一致	0	2	1	3
その他	**Normal**	一致	12	83	0	95
		不一致	5	34	3	42
	Other Finding	一致	5	2	0	7
		不一致	133	29	8	170

表 A.2: 20 疾患ラベル別 AUC の比較結果

疾患ラベル	MeSH (Baseline)	LLM (提案手法)	変化
心拡大 (Cardiomegaly)	0.96	**0.98**	↑
石灰化 (Calcinosis)	0.93	0.93	—
胸水 (Pleural Effusion)	0.98	0.98	—
ヘルニア (Hernia)	**0.96**	0.95	↓
肺気腫 (Emphysema)	0.89	**0.90**	↑
気胸 (Pneumothorax)	0.95	0.95	—
無気肺 (Atelectasis)	0.92	0.92	—
胸膜肥厚 (Pleural Thickening)	0.92	0.92	—
肺炎 (Pneumonia)	0.98	0.98	—
正常 (Normal)	0.94	**0.96**	↑
不透明度 (Opacity)	0.89	0.89	—
脊柱側弯症 (Scoliosis)	**0.98**	0.97	↓
骨折 (Bone Fractures)	0.94	**0.95**	↑
瘢痕 (Cicatrix)	**0.91**	0.90	↓
病変 (Lesion)	0.90	**0.92**	↑
気腔病変 (Airspace disease)	0.92	**0.93**	↑
低膨張 (Hypoinflation)	**0.97**	0.96	↓
医療器具 (Medical Device)	0.96	0.96	—
浮腫 (Edema)	1.0	1.0	—
その他 (Other Finding)	0.94	0.85	↓
全体平均 (Macro AUC)	**0.94**	**0.94**	

付録 B