

[研究助成 (A)]

病理医が信頼を寄せる人協調型 AI 診断システムの開発

2241017

研究代表者
(助成金受領者)香川大学
医学部歯科口腔外科学講座

准教授

助 川 信太郎

共同研究者

岐阜大学
工学部電気電子・情報工学科情報コース

教授

原 武 史

[研究の目的]

病理診断は、病理医が顕微鏡下で組織構造や細胞の形態を詳細に観察し、所見を総合的に判断する形態学的診断であり、極めて高度な専門的技術と経験を要する分野である。中でも、口腔内は常に外界からの物理的・化学的刺激を受ける環境下にあり、炎症や再生、変性といった組織反応が多彩に起こるため、得られる病理組織像も複雑かつ多様であることが知られている。したがって、これらの病理組織像に対して正確な診断を行うには、長年にわたる豊富な臨床経験と体系的な教育を通じた熟練した診断能力が不可欠である。

しかしながら、本邦においては病理医の人材不足が長年にわたって深刻な問題となっており、特に高度な専門性が求められる診断分野においては、限られた病理医に過度な負担が集中しているのが現状である。こうした背景のもと、近年では人工知能（AI）技術の応用によって、病理医の業務を支援しようとする試みが進められてきたが、実用化には未だ至っていない。その主な要因として、病理診断が単なる画像認識を超え、医師の経験則や臨床文脈を踏まえた複合的な判断によってなされる点が挙げられる。すなわち、AIが提示する診断に対して、病理医がどの程度信頼を寄せ、どのような条件下でAIの出力を参考にするのかといった、人とAIの協働関係の構築が十分に検討されてこなかつ

たことが、実装の障壁となっている。

さらに、AIによる診断支援の有効性は客観的な性能評価だけでなく、医療現場での受容性や実際の意思決定に与える影響の理解が不可欠である。本研究では、病理診断支援に資するAIシステムの開発を行うとともに、病理医がAIとどのように協調し、適切に信頼を置くかという「人とAIの協働における信頼形成のプロセス」を明らかにすることを目的としている。これにより、単なる診断精度の向上にとどまらず、病理医がAIを道具として使いこなしながら、自らの専門性を維持・強化できる診断体制の構築に寄与することが期待される。また、AIに対する過信や不信といった極端な依存の回避を促し、医療現場における持続可能かつ実践的な協働モデルの確立に貢献する。

また、医師によるAI診断への過信や不信の両極端な依存傾向を避け、バランスの取れた判断を導くための枠組みを提示することで、医療分野における人協調型AIの新たな実装モデルとして広く応用可能な診断支援システムの確立を目指す。

[研究の内容、成果]

本研究における人協調型AI病理診断システム開発の技術的課題は、①病理専門医と同等の安定した診断パフォーマンスを発揮できる深層学習モデルの構築、②病理医がAI診断結果に

対して過信や不信に陥ることなく適切な距離感をもって協調的に判断を行うための信頼校正手法の構築である。具体的には、専門医による正確な診断ラベリングを用いた高品質な病理画像データセットの整備とモデル訓練を実施するとともに、AI 補助診断に対する病理医の信頼度を動的に推定し、過剰または過小な依存を検出・調整する信頼校正 AI の設計・実装を進める。これにより、人と AI の協働による診断支援の質的向上に資する診断インターフェースの開発に向けて、評価と検証を実施した。

1) データセットの作成

本研究では、香川大学医学部附属病院歯・顎・口腔外科において治療された口腔癌症例を対象とし、両施設の倫理委員会の承認を得た上で取得したデータを用いた。症例ごとの病理標本は、ガラススライドからバーチャルスライド (Whole Slide Image: WSI) に変換して使用した。スライドのスキャンには、Ventana DP 200 スライドスキャナー (ロシュ社, スイス・バーゼル) を用い、WSI を作成した。得られた WSI は、QuPath ソフトウェア (<https://qupath.github.io> バージョン 0.4.1) を用いてタイリング処理を行い、病理診断に最適とされる設定で画像を切り出した。切り出し画像は倍率 1.0 とし、1 辺あたり 768 ピクセルの正方形領域を 192 ピクセルずつ重複させてセグメント化し、最終的に 256 × 256 ピクセルの PNG (Portable Network Graphics) 形式の画像データを作成した。

また、WSI の評価およびアノテーションは、日本口腔病理学会認定の専門資格を有し、15 年以上の実務経験を有する口腔病理専門医によって実施された。これにより、高精度かつ信頼性の高いラベリングデータセットの構築がおこなった (図 1)。

2) 病理診断 AI における深層学習モデルの構築

本研究では、病理診断支援を目的とした AI モデルとして、深層学習 (Deep Learning) を

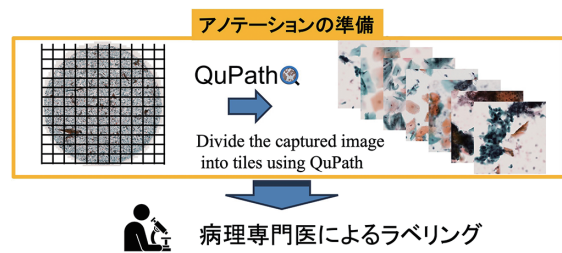


図1 専門家によるラベリング

用いた構築を行った。モデルアーキテクチャとしては、画像認識において高い性能を示す畳み込みニューラルネットワーク (Convolutional Neural Network: CNN) を VGG16 ならびに ResNet50 採用し、各種パフォーマンス指標に基づき検証を行った。最適化アルゴリズムについては、一般的な SGD (確率的勾配降下法) に加えて、より効率的な学習が可能なモメンタム付き SGD (SGDM) および、近年提唱され精度改善が期待されるシャープネス認識最小化 (Sharpness-Aware Minimization: SAM) を比較対象として選定した。SGDM のモメンタム係数は 0.9 に設定した。SAM においては、近傍サイズ (ρ) の最適化を目的として、グリッドサーチにより {0.01, 0.025, 0.05, 0.1, 0.2, 0.5} の各値について性能を検証した。さらに、学習率の調整に関しては、学習率スケジューラの有無による影響を評価した。学習率スケジューラでは学習率減衰を使用し初期学習率は 0.01 に設定し、学習過程における性能変化を詳細に解析した。

これらの検証を通じて、病理画像に対する診断支援 AI として最適な学習条件およびモデルを検証した。

これらの検証を行い、深層学習のベストモデルは ResNet50、最適化アルゴリズム: SAM (近傍サイズ 0.025)、学習率スケジューラありが最も高いパフォーマンスを示した (表 1)。

3) AI 診断が病理診断に与える影響の検証

本研究では、病理医と AI による協調的意思決定が、より高い診断満足度とパフォーマンス

表 1

Optimizer	Learning rate	Accuracy	Precision	Recall	F1 score	AUC
		SD	SD	SD	SD	SD
		95%CI	95%CI	95%CI	95%CI	95%CI
SGD	w/o Scheduler	0.8879	0.8518	0.6501	0.6937	0.9020
		0.0016	0.0099	0.0089	0.0095	0.0075
		0.8873–0.8885	0.8482–0.8553	0.6469–0.6533	0.6903–0.6971	0.8993–0.9047
	with Scheduler	0.8970	0.8049	0.7581	0.7780	0.9098
		0.0026	0.0066	0.0064	0.0057	0.0030
		0.8961–0.8979	0.8026–0.8073	0.7558–0.7604	0.7759–0.7800	0.9087–0.9108
SAM	w/o Scheduler	0.8870	0.8529	0.6438	0.6872	0.8766
		0.0004	0.0024	0.0017	0.0019	0.0069
		0.8868–0.8871	0.8521–0.8538	0.6432–0.6444	0.6875–0.6879	0.8741–0.8791
	with Scheduler	0.9016	0.8534	0.7139	0.7578	0.9328
		0.0018	0.0072	0.0124	0.0098	0.0016
		0.9010–0.9023	0.8509–0.8560	0.7094–0.7183	0.7543–0.7614	0.9322–0.9334

SD: standard deviation, 95% CI: 95% confidence interval, AUC: Area under the ROC curve

を達成するためには、単に AI の診断精度を向上させるだけでは不十分であると考える。医療人が AI の出力をどのように受け取り、どの程度信頼し、最終的な診断にどのような影響を受けるのかを明らかにすることが、本研究の重要なテーマである。そのため本研究では、AI 診断結果が病理医の診断意思決定に与える影響度の定量的検証を行った。検証には、前段で開発・選定された最適深層学習モデル（ベストモデル）による診断出力を使用した。

検証プロセスは以下の手順で実施した。

病理医に対し、AI の出力を提示する前に対象画像の事前診断を独立に実施。次に、同一画像に対する AI の診断結果および診断確信度（信頼度スコア）を提示した上で、改めて診断施行し、AI 出力が診断に与える影響を評価した（図 2）。

影響度の定量化には、連続信頼法（Continuous Confidence Rating）を採用した。これは、各病理医が画像ごとにそれぞれの診断ラベルについて、0～100 の視覚アナログスケールを用いて信頼度スコアを付与する方法である。得られたスコアは SoftMax 関数を用いて正規化し、3 分類の出力合計が 1.0（100%）になるよう変換を行った。

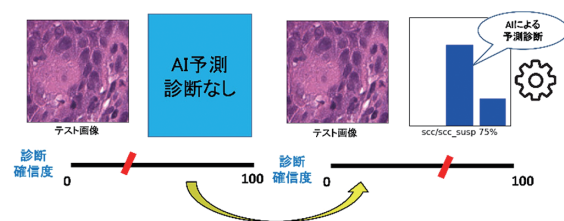


図 2 AI 補助診断に対する確信度の変化の分析
AI を参考にしたどの画像でより効果的かを解析

AI の支援によって診断がどのように変化したか、また信頼度にどのような変容が生じたかについては、ROC 曲線（Receiver Operating Characteristic Curve）および AUC（Area Under the Curve）を用いて定量評価を実施した。また AI 診断の有無の影響度を効果量にて評価をした。効果量は Cohen によって提案された指標を Sawilloski によって提案された基準に基づいて判断した。

結果を図 3 に示す。AI 診断の有無によるマクロ AUC およびミクロ AUC の平均値の間には、いずれも統計的に有意な差が認められた（ $p = 0.031$ ）。このことから、AI の提示が病理医の診断性能に影響を及ぼしていることが明らかとなった。さらに、AI 診断の導入による効果量はマクロ AUC においては 1.46、ミクロ AUC においては 2.04 という結果が得られた。

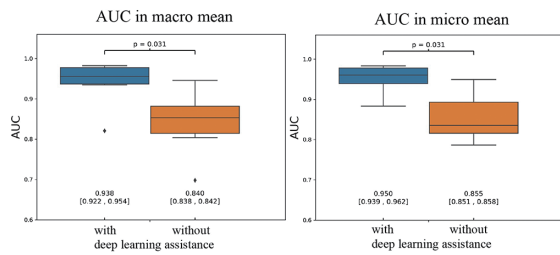


図3 AI診断の有無によるAUC

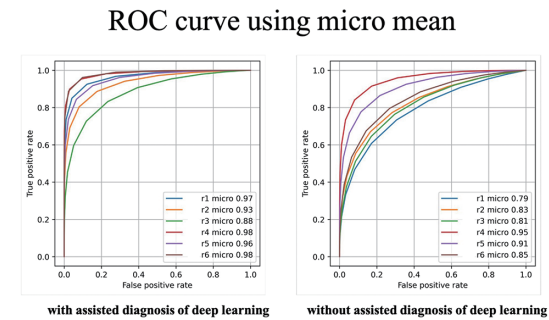
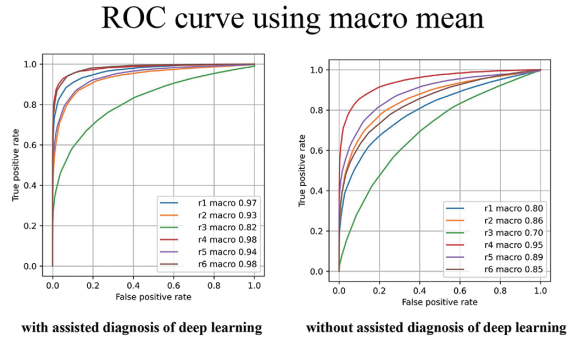


図4 AI診断の有無による病理医別の診断結果

これらの値は、いずれも一般的な解釈において「非常に大きい」効果 (very large effect) に相当し、AI支援が診断精度の向上に強く寄与していることを示している。以上の結果より、AI診断を使用することは病理医の診断能力を統計的に有意かつ実質的に向上させることが示された。図4では、病理医別の診断結果をAI診断ありとなしであるもののAUCを示す。口腔病理医 #1は、AI診断を使用した場合に最も効果的でした。AI診断なしの場合、マクロ平均値は0.80 (95%CI: 0.795 ~ 0.810), ミクロ平均値は0.79 (95%CI: 0.776 ~ 0.791) であった。AI診断を使用した場合、マクロ平均値は0.97 (95%信頼区間: 0.960 ~ 0.966), ミクロ平均値は0.97 (95%信頼区間: 0.958 ~ 0.965)

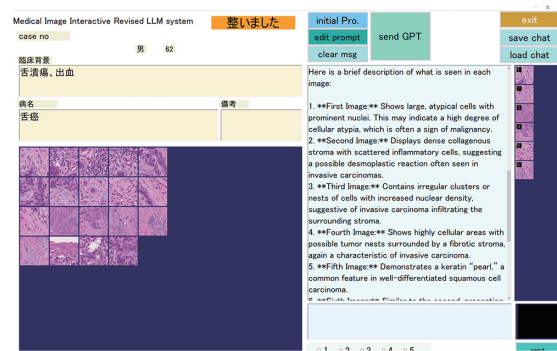


図5 視覚一言語モデルを用いた所見文自動生成システム

であった。AI診断効果では、個人差があることが明らかになった。

4) 視覚一言語モデルを用いた所見文自動生成と病理医に与える影響

これまでの検討結果から、病理医が診断を行う際には単に正しい診断名が提示されるだけでは不十分であり、AIがどのような根拠に基づいて診断に至ったのかという「診断過程の理解」や「納得感」が診断精度や判断行動に影響を及ぼす可能性が示唆された。病理医は診断において、最終的な結果のみならず、形態学的特徴、所見の位置づけ、鑑別過程といった説明可能性を重視する傾向がある。したがって、AIが生成した結果に対して診断者が理解可能な根拠が提示されることは、AIの臨床導入における受容性や信頼性に寄与すると考えられる。

そこで本研究では、診断理由を視覚的かつ言語的に補完する手法として、視覚一言語モデル (Visual-Language Model: VLM) を用いた所見文自動生成システムを構築し、その有用性を検証した (図5)。

本システムでは、AIが病理画像から抽出した特徴に基づいて診断名と所見文を生成し、病理医へ提示することで、AIの判断根拠を明示し、診断行動における納得度および判断過程への影響を明らかにすることを目的とした。さらに、本研究では実際の臨床利用を想定し、生成された所見文の内容理解のしやすさや診断支援としての受容性にも着目し、診断行為における

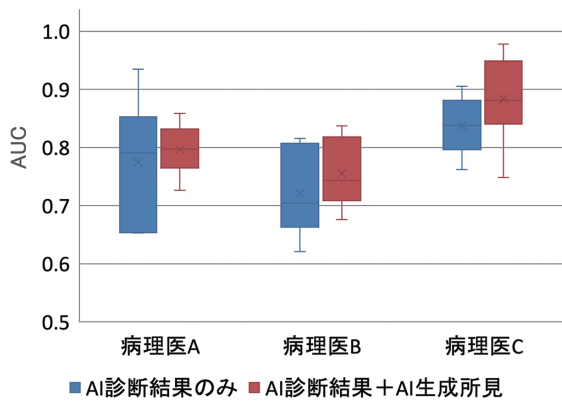


図6 意思決定影響評価研究

AI 情報提示の効果を検討した。

対象症例は口腔細胞診画像 30 例とし、診断経験を有する口腔病理医 3 名により評価を行った。診断条件は 2 段階で設定し、まず「細胞診画像 + AI による病理診断名のみ」を提示して診断を行わせ、次に同一症例に対して「細胞診画像 + AI 診断名 + AI 生成所見文」を提示し、再度診断を実施した。回答時間は制限せず、自由な判断環境とした。両条件間で診断の変化を比較することで、生成 AI 所見が病理医の診断行動および認知過程に与える影響を評価した。

評価指標として AUC を用い、診断性能を定量的に比較した。さらに、リサンプリング手法としてジャックナイフ法を適用し、条件間の差異について Z 検定により統計解析を行った。

その結果、AI 生成所見文の有無による診断結果の差異は認められず、統計学的にも有意差は示されなかった ($p > 0.05$) (図 6)。

[まとめと今後の研究]

本研究では、病理医と AI が協調して診断を行う「人協調型 AI 病理診断支援システム」の構築を目指し、高精度な病理画像データセットの作成、深層学習による最適モデルの開発、そ

して AI 診断が病理医の判断に与える影響の検証を実施した。AI の導入により診断精度が有意に向上し、視覚一言語モデルによる所見文提示が病理医の納得感にも寄与する可能性があるが、現時点では被験者の数が少なく統計学的な有意差は示されなかった。AI 診断の活用が、病理診断の質と効率を向上させるためにはより有効な提示方法などを考慮することが必要であると考えられた。

今後の課題としては、まず病理医ごとの診断傾向や経験年数に応じて、AI 出力と診断行動の関係性を最適化する仕組み、すなわち個別化された AI 信頼校正 (trust calibration) が必要となる。また、本研究で用いた視覚一言語モデルが生成する所見文については、医学的妥当性や用語選択の正確性、表現の一貫性を含め、医療文書としての品質評価を体系的に行う必要がある。さらに、実臨床において AI 提示情報をどの段階で、どの形式で、どの程度提示すべきかといった運用設計も重要な検討項目である。

加えて、AI 支援診断における診断責任や説明責任の所在を整理する法的・倫理的枠組みの整備も不可欠である。特に、AI の所見が診断者の判断に影響を与えた場合の責任構造、AI による診断根拠説明 (explainability) の必然性、トレーサビリティ確保に向けた標準化など、多面的な検討が求められる。

[成果の発表、論文など]

論文

Shintaro Sukegawa, Takeshi Hara, et al. Training high-performance deep learning classifier for diagnosis in oral cytology using diverse annotations. Sci Rep. 2024 Jul 30; 14(1): 17591. doi: 10.1038/s41598-024-67879-w.

国内シンポジウムでの発表

第 63 回日本生体医工学会大会・学術大会 生体医工学における信頼される AI の実現に向けて～口腔外科医に寄り添う AI の姿～ (2024 年 5 月 24 日)