学位論文（博士）

表在性非乳頭部十二指腸上皮性腫瘍の組織学的悪性度評価における AI 分類器の開発と検証

氏名　山本　一太

所属　山口大学大学院医学系研究科

医学専攻　消化器内科学講座

令和８年５月

目　次

１．要旨・・・・・・・・・・・・・・・・・・・・・３

２．研究の背景・・・・・・・・・・・・・・・・・・５

３．方法・・・・・・・・・・・・・・・・・・・・・７

４．結果・・・・・・・・・・・・・・・・・・・・１２

５．考察・・・・・・・・・・・・・・・・・・・・１６

６．結語・・・・・・・・・・・・・・・・・・・・２０

７．謝辞・・・・・・・・・・・・・・・・・・・・２１

８．参考文献・・・・・・・・・・・・・・・・・・２２

１．要旨

【背景・目的】

表在性非乳頭部十二指腸上皮性腫瘍(SNADET)は比較的稀な腫瘍であるが、近年発見率や治療数は増加傾向にある。適切な治療方針を決定し過大な侵襲を避けるためには、内視鏡画像による正確な術前診断が不可欠である。しかし、現在、標準化された診断方法は確立されていない。本研究では、低異型度と高異型度の腫瘍を識別するための深層学習を用いた人工知能（AI）分類器を開発し、その診断能を評価した。

【対象・方法】

山口大学の SNADET 85 症例（低異型度 41 例、高異型度 44 例、白色光画像 317 枚）を学習用データセットとして AI 分類器の訓練を行った。検証は、3 施設から収集した 69 病変（低異型度 31 例、高異型度 38 例、白色光画像 229 枚）のテスト用データセットを用いて実施した。また、比較対象として、3 名の内視鏡専門医がスコアリングシステムを用いて同テストセットを評価した。

【結果】

AI 分類器の診断能は、正診率 78.3%、感度 92.1%、特異度 61.3%、陽

性的中率（PPV） 74.5%、陰性的中率（NPV） 86.4%、F1 measure（感度と陽性的中率の調和平均）は 0.824 であった。一方、3 名の内視鏡医による診断は、正診率 58.0%、感度 50.0%、特異度 67.7%、PPV 65.5%、NPV 52.5%、F1 measure 0.567 であった。AI の診断能は、正診率、感度において、内視鏡医よりも有意に高かった（$p<0.05$）。

【結論】

高異型度腫瘍の鑑別において、AI 分類器は内視鏡医と比較して高い正診率と感度を示した。これにより、治療方針の決定を支援するツールとしての有用性が示唆された。

２．研究の背景

非乳頭部十二指腸表在性腫瘍（SNADET）は、十二指腸の粘膜または粘膜下に存在する散発性の腫瘍であり、ファーター乳頭から発生しないものと定義される。SNADET は前癌病変あるいは早期癌と考えられており、治療的介入を要する病変とされている [1]。近年、日本における SNADET の発見率は上昇しており [2]、内視鏡治療の進歩に伴って現在は内視鏡的切除が第一選択の治療法として推奨されている [3]。

SNADET の診療において、適切な治療方針の決定には、内視鏡による正確な組織学的異型度の評価が不可欠である。低異型度の腫瘍と診断された病変は、コールドスネアポリペクトミー（CSP）のような低侵襲な内視鏡的治療[4]、あるいは経過観察[5,6]が選択できる可能性があるとされている。一方で、高異型度の腫瘍は、内視鏡的粘膜切除術（EMR）や内視鏡的粘膜下層剥離術（ESD）などの電気焼灼を用いた手技で切除されるべきである [7]。

組織学的な異型度を予測するために、これまでに複数の内視鏡画像における診断基準が提唱されているが [8,9]、現在まで標準化された診断基準は存在しない。また、生検による組織学的診断は、内視鏡診断と比較して診断精度が低く、更に線維化を引き起こして内視鏡的切除を困難にする可能性がある[9,10]。したがって SNADET の切除に際しては、生検を伴わない正確な内視鏡診断が望ましいとされる [11,12]。

近年、人工知能（Artificial Intelligence：AI）の発展が著しく、消化管腫瘍の診断においても AI 分類器が応用されている。十二指腸病変に関しては、Inoue らが AI によって感度 94.7%、特異度 87.4%で SNADET を検出したと報告している [13]。しかし、内視鏡画像から組織学的異型度を推定する AI に関する研究は未だ存在しない。本研究では、SNADET における低異型度腫瘍と高異型度腫瘍を鑑別するための、深層学習を用いた AI 分類器を開発し、その診断能を評価した。

３．方法

本研究は、山口大学医学部附属病院（H2022-148）、周東総合病院、小倉記念病院、国立病院機構関門医療センター、および徳山工業高等専門学校の倫理審査委員会の承認を得て実施された。

＜対象＞

2008 年 6 月から 2024 年 3 月の間に山口大学医学部附属病院で内視鏡的切除が施行された SNADET を訓練症例とした。各病変は、治療後の切除標本に基づき、低異型度または高異型度と病理組織学的に診断された。画像の質が低いもの（ピントのずれ、血液や粘液の付着、ハレーションなど）および類似の画像（ほぼ同一の距離、画角で撮影されたと判断されるもの）は除外した。過学習を防ぐため、1 症例あたりの画像数は最大 6 枚に制限した。画像はオリンパス社製の内視鏡（GIF-H260、GIF-Q260J、GIF-H260Z、GIF-H290、GIF-H290Z、または GIF-EZ1500）を用いて撮影された。テスト症例は、2009 年 9 月から 2023 年 2 月の間に協力 3 施設（関門医療センター、小倉記念病院、周東総合病院）で内視鏡的に切除された SNADET とした。これらもオリンパス社製の内視鏡（GIF-H260、GIF-Q260J、GIF-H260Z、GIF-H290、GIF-H290Z、または GIF-XZ1200）を用いて撮影された。なお、家族性大腸腺腫症に関連する症例はいずれのセットからも除外した。

すべての病変は、各施設の病理専門医によって診断された。診断は内視

鏡治療により得られた切除標本に基づき、消化管上皮性腫瘍に対する改訂ウィーン分類（VCL）[14] に従って分類し、カテゴリー3（低異型度腺腫）、カテゴリー4.1（高異型度腺腫）、カテゴリー4.4（粘膜内癌）、カテゴリー5（粘膜下層浸潤癌）のいずれかへ分類された。さらに、胃型形質を有する腫瘍は腸型腫瘍よりも悪性度が高いことが報告されている [15,16]ことから、本研究ではVCLカテゴリー3の腸型病変を「低異型度腫瘍」と定義し、胃型、混合型、およびVCLカテゴリー4/5の病変を「高異型度腫瘍」と定義した。

＜AI分類器の開発＞

AI 分類器は徳山工業高等専門学校との協力により開発された。設計には、1,400 万枚を超える画像の大規模データセットである ImageNet で事前にトレーニングされたEfficientNet B3を学習に用いた [17,18]。日本消化器内視鏡学会の専門医2名が、画像上の腫瘍が存在する領域にマーキングを行った。マーキングされた領域に外接する正方形で切り抜き、マーキング外を黒色で塗りつぶし、300 × 300 ピクセルにリサイズした（図1）。AI 分類器の構築には、NVIDIA GeForce RTX 3070 と AMD Ryzen Threadripper 3960Xを用いた。

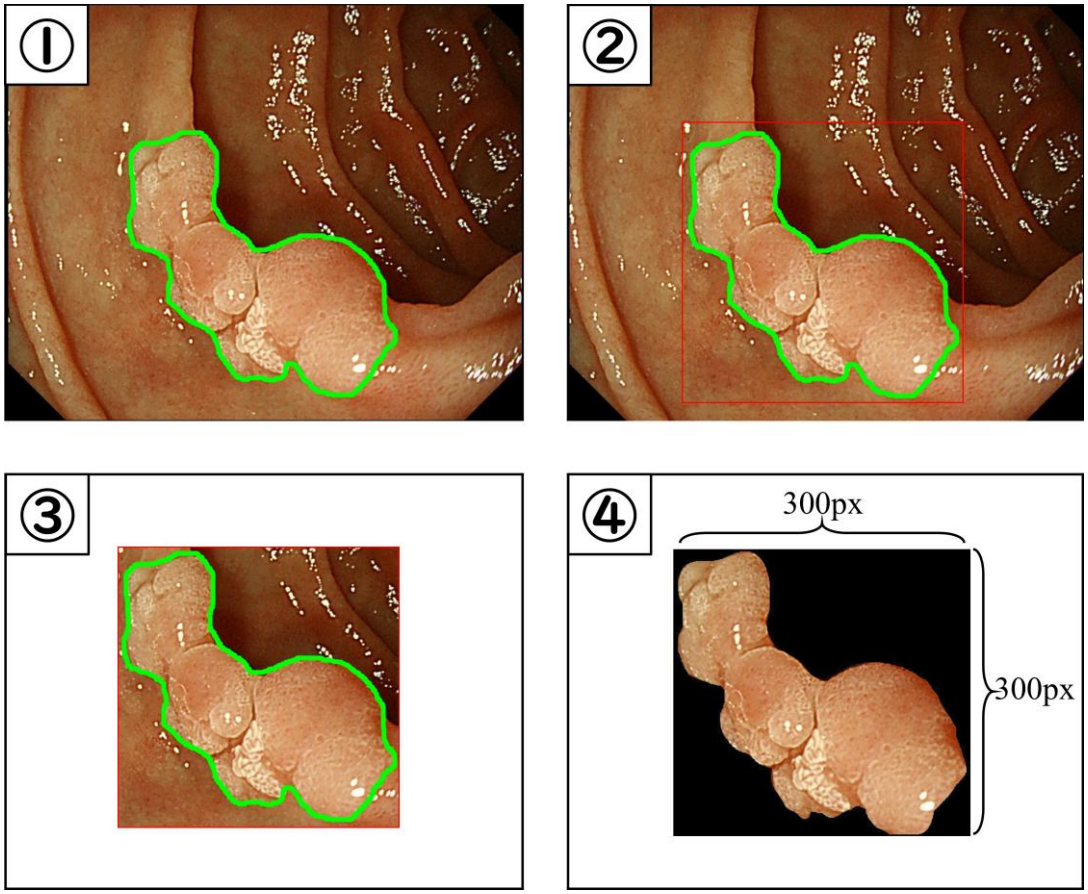


図１：訓練時の画像処理

＜Leave-one-out 法による診断能の推定＞

訓練用画像 317 枚を用い、leave-one-out 法 ［19］ により AI 分類器の低異型度/高異型度の鑑別能を推定した。この手法では、317 枚の画像を 316 枚の学習用画像と 1 枚の疑似テスト画像に分割する。316 枚の画像を用いて AI 分類器を設計し、1 枚の疑似テスト画像を診断する。この手順を 317 回繰り返し、内部妥当性を評価した。

＜AI 診断方法：単一画像解析と複数画像解析＞

従来のAI診断システムの多くは画像単位で診断を行うが、本研究では、より実際の臨床現場に即した、複数画像による解析を検討した。まずAIの診断に先立ち、各症例につき1枚の代表画像を選定した。これは、病変のマーキングを行った日本内視鏡学会専門医2名が、最も質が高く診断に適すると判断される画像を、事前に選択したものである。次に、全画像をAI分類器により低異型度または高異型度に分類した。その後、各症例の最終診断を決定するために2つの解析手法を行った。第1の手法は、代表画像に対するAIの予測をその症例の最終診断とするもので、これを「単一画像解析」と定義した。第2の手法は、1症例内の全画像における予測結果のうち、最も頻度の高いものを最終診断とする多数決方式であり、これを「複数画像解析」とした。なお、判定が同数の場合は、代表画像の結果を優先した。高異型度腫瘍を陽性と定義し、正診率、感度、特異度、陽性的中率、陰性的中率、およびF1 measureを算出した。F1 measureは感度と陽性的中率の調和平均であり、機械学習の分野でしばしば用いられる指標である。診断性能のバランスを評価する指標として、F1 measureも用いた。

＜内視鏡医の診断との比較＞

AIと医師の診断能を比較するため、3名の日本消化器内視鏡学会専門医によりテスト症例が診断された。病変径、色調、肉眼型、結節の有無の4

項目について、それぞれ 0〜2 点を割り当てる診断用のスコアリングシステム［8］を用い、テスト症例の 69 病変を診断した。このスコアリングシステムでは、合計 2 点以下は VCL カテゴリー3、3 点以上は VCL カテゴリー4 以上と分類される。3 名のスコアの中央値を内視鏡医全体における最終診断とした。AI の診断と同様に、正診率、感度、特異度、PPV、NPV、F1 measure を算出した。また、内視鏡医間の一致度を Fleiss のカッパ（κ）係数を用いて評価した。

＜統計解析＞

異なる AI モデル間、および AI と内視鏡医の間の診断能比較にはマクネマー検定（McNemar’s test）を用い、正診率、感度、特異度のそれぞれについて個別に検定を行った。

４．結果

＜臨床病理学的特徴＞

学習セットおよびテストセットの特徴を表 1、2 にまとめる。訓練症例は 80 名の患者から得られた 85 病変（低異型度 41 病変、高異型度 44 病変）とし、1 病変あたり最大 6 枚、合計 317 枚の白色光画像が用いられた。テスト症例では、66 名の患者から 69 病変（低異型度 31 病変、高異型度 38 病変）、計 229 枚の白色光画像が用いられた。

	Low-grade neoplasia (41 lesions, 136 images)	**High-grade neoplasia (44 lesions, 181 images)**
Median age, years (range)	66 (38-84)	69 (37-83)
Male/female, n (%)	30 (73%) / 11 (27%)	33 (75%) / 11 (25%)
Location of the lesion, n (%)		
Superior part / Descending and horizontal part	6 (15%) / 35 (85%)	12 (27%) / 32 (73%)
Median tumor diameter, mm (range)	5 (3-15)	13 (2-30)
Macroscopic type, n		
0-I/0-IIa/0-IIb/0-IIc	2 / 27 / 1 / 11	7 / 29 / 0 / 8
Histology, n		
VCL Category 3 / 4.1 / 4.4 / 5	41 / 0 / 0 / 0	0 / 11 / 33 / 0
Intestinal / mixed / Gastric type	41 / 0 / 0	38 / 2 / 4
Underwent biopsy, n (%)	32 (76%)	36 (82%)

表 1：訓練症例の臨床病理学的特徴（85 症例、317 枚）

	Low-grade neoplasia (31 lesions, 91 images)	High-grade neoplasia (38 lesions, 138 images)
Median age, years (range)	64 (47-87)	69 (46-92)
Male/female, n (%)	25 (81%) / 6 (19%)	28 (74%) / 10 (26%)
Location of the lesion, n (%)		
Superior part / Descending and horizontal part	5 (16%) / 26 (84%)	14 (37%) / 24 (63%)
Median tumor diameter, mm (range)	6 (2-20)	11 (4-25)
Macroscopic type, n		
0-I/0-IIa/0-IIb/0-IIc	8 / 17 / 1 / 5	15 / 20 / 0 / 3
Histology, n		
VCL Category 3 / 4.1 / 4.4 / 5	31 / 0 / 0 / 0	0 / 16 / 18 / 4
Intestinal / mixed / Gastric type	31 / 0 / 0	37 / 0 / 1
Underwent biopsy, n (%)	24 (77%)	34 (90%)

表 2：テスト症例の臨床病理学的特徴（69 症例、229 枚）

＜訓練症例における診断能＞

leave-one-out 法による訓練症例での成績を表 3 に示す。単一画像解析では、正診率 80.0%、感度 86.4%、特異度 73.2%、F1 measure 0.817 であった。複数画像解析では、正診率 82.4%、感度 84.1%、特異度 80.5%、F1 measure 0.831 であった。

		Accuracy	Sensitivity	Specificity	PPV	NPV	F1 measure
Training	**Single-image**	**80.0%**	**86.4%**	**73.2%**	**73.6%**	**83.3%**	**0.817**
	Multi-image	**82.4%**	**84.1%**	**80.5%**	**82.2%**	**82.5%**	**0.831**

表 3：訓練症例における成績（leave-one-out 法）

＜テスト症例における診断能＞

テスト症例における成績を表 4 に示す。単一画像解析では、正診率 75.4%、感度 86.8%、特異度 61.3%、F1 measure 0.795 であった。複数画像解析では、正診率 78.3%、感度 92.1%、特異度 61.3%、F1 measure 0.824 であった。

		Accuracy	Sensitivity	Specificity	PPV	NPV	F1 measure
Test	Single-image	75.4%	86.8%	61.3%	73.3%	79.2%	0.795
	Multi-image	78.3%	92.1%	61.3%	74.5%	86.4%	0.824

表 4：テスト症例における成績

＜内視鏡医の成績＞

スコアリングシステムを用いた内視鏡医の成績を表 5 に示す。結果は、正診率 58.0%、感度 50.0%、特異度 67.7%、F1 measure 0.567 であった。3 名の観察者間の Fleiss のカッパ係数は 0.166 であった。

	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 measure
Dr. A	59.4%	63.2%	54.8%	63.2%	54.8%	0.632
Dr. B	65.2%	55.3%	77.4%	75.0%	58.5%	0.636
Dr. C	58.0%	47.4%	71.0%	66.7%	52.3%	0.554
Median score	58.0%	50.0%	67.7%	65.5%	52.5%	0.567

表 5：内視鏡医の成績（スコアリングシステム）

＜診断能の比較＞

図 2 に、AI の単一画像解析と複数画像解析の比較結果を示す。正診率、感度、特異度のいずれにおいても有意差は認められなかった（p > 0.05）。図 3 には、AI（複数画像解析）と内視鏡医の比較結果を示す。AI の正診率は内視鏡医よりも有意に高かった（78.3% vs. 58.0%, p < 0.05）。同様に感度においても有意差が認められたが（92.1% vs. 50.0%, p < 0.05）、特異度については有意差は認められなかった（61.3% vs. 67.7%, p = 0.581）。

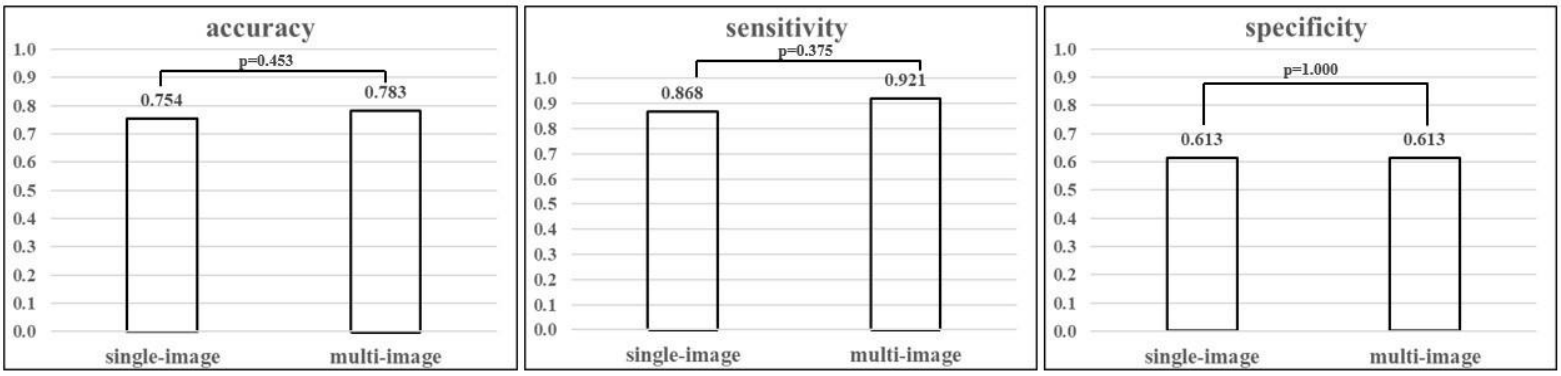


図 2：AI の単一画像解析と複数画像解析の比較

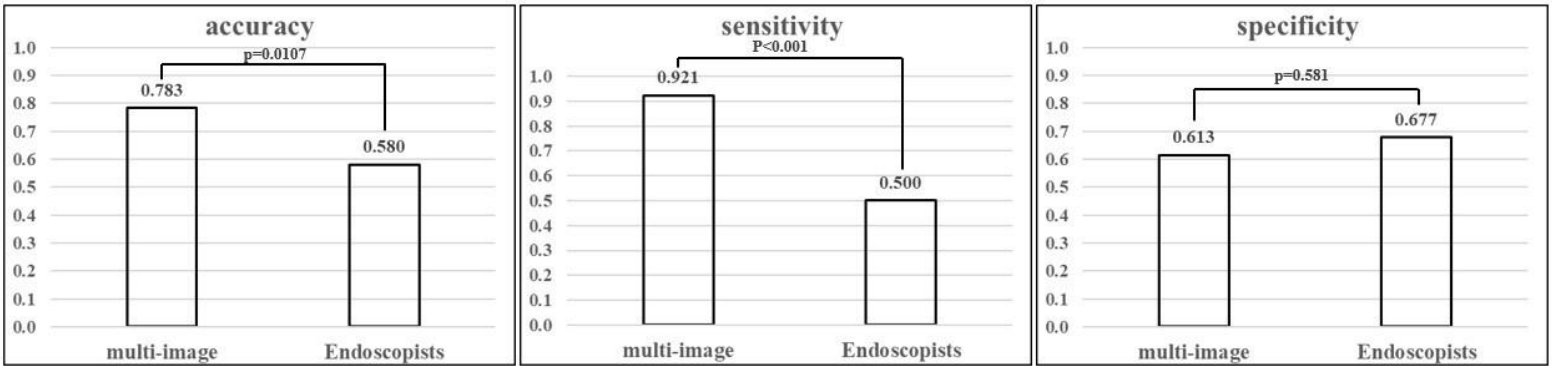


図 3：AI（複数画像解析）と内視鏡医との比較

５．考察

筆者らの知る限り、これは SNADET の内視鏡画像による組織学的異型度分類を AI を用いて評価した初の研究である。本研究で開発した AI 分類器は、訓練症例を取得した施設とは異なる 3 施設から収集されたテスト症例において、正診率 78.3%、感度 92.1%、特異度 61.3%、F1 measure 0.824 という良好な性能を示した。この結果は、モデルの訓練方法が適切であり、訓練用データの多様性が十分であったことを示唆している。解析の結果、単一画像解析と複数画像解析の間に統計的な有意差は認められなかったが、多数決に基づき複数の画像を用いて診断する方式がより適切であると判断した。この手法であれば、内視鏡医が代表画像を選択する際に生じる主観的な選択バイアスを軽減することが可能になる。加えて、内視鏡医は通常、1 症例につき複数の画像を参照して診断を下すため、この方式はより臨床に即した方法であると考えた。

日本において SNADET には内視鏡的切除が広く推奨されており [3]、治療方針の決定における正確な内視鏡診断の必要性は極めて高い。十二指腸の内視鏡手技は技術的に難易度が高く、偶発症のリスクが高いことが知られており、特に ESD や EMR などの電気的な焼灼を伴う手技は、穿孔や出血などの合併症を伴いやすい [20]。一方で、CSP は技術的に比較的簡便であり、偶発症の発生率も低い [20,21]。しかし、CSP では粘膜下層の完全な切除が難しく、切除断端の組織学的評価が困難になる場合がある。したが

って、CSP は術前に VCL カテゴリー3 に相当する低異型度腫瘍と診断された病変にのみ適応されるべきとされている［4]。また、10㎜ 未満の低異型度腺腫については切除を行わない選択肢も許容され得ることから［5,6]、信頼性の高い術前の異型度診断は、不要な侵襲を回避することが可能になる。一方、VCL カテゴリー4 または 5 に分類される病変は十分な切除が必要であり、電気的な焼灼を用いた手技が局所再発率を低下させることが示されている［7]。以上より、低異型度と高異型度の正確な内視鏡的鑑別は、低侵襲治療や経過観察の適切な選択を可能にし、臨床管理を最適化するために不可欠である。

本 AI 分類器は、内視鏡スコアリングシステムを用いた消化器内視鏡専門医の診断と比較し、より高い診断性能を示した。特に、内視鏡医間の一致度が低かった（κ = 0.166）ことから、AI がより高い精度且つ安定した診断を下すことができる可能性を示唆している。診断には生検が用いられる場合もあるが、先行研究では SNADET における生検の感度は内視鏡診断と比較しても低いこと（58.0%）が報告されており［22]、また生検後の線維化は内視鏡的切除を妨げる要因となり得るとされている［10]。現在、内視鏡治療の可能性がある病変についてはルーチンの生検は推奨されていない。本 AI 分類器を診断補助として用いることで、悪性度の高い腫瘍の見落としを防ぐ一助になると考える。

本 AI は白色光画像のみで学習されているが、他の画像モダリティを併

用することで診断能がさらに向上する可能性がある。現在の臨床においては、SNADET の診断には白色光画像に加えてインジゴカルミン散布像、狭帯域光観察（NBI）、および拡大 NBI を併用して診断することが一般的である。本 AI システムにおいても、複数のモダリティを統合することで診断精度がさらに改善される可能性があるため、マルチモーダルな AI の研究が今後の課題である。

本研究にはいくつかの limitation がある。第一に、学習画像の数が少なく、単一施設からのデータである点である。一般に、AI モデルは未知のテスト画像で性能を評価する際、訓練の場合と比較して性能が低下する傾向がある。しかし本 AI では、訓練症例とは完全に独立した 3 つの施設からのテスト症例を用いたにもかかわらず、性能の低下はごく軽度であった（複数画像解析において F1 measure は訓練時 0.831 に対し、テスト時 0.824）。しかしながら、今後のモデルでは、より高い汎用性を確保するために、より多様、多施設にわたる訓練画像を用いるべきである。第二に、比較的悪性度が高く予後不良であることが知られている胃型病変[15,16] が少数しか含まれていない点である。この点でも、さらなるデータ収集と検証が不可欠と考える。最後に、本研究は低異型度腫瘍と高異型度腫瘍の鑑別を目的としたため、本 AI は腫瘍と非腫瘍の識別能力を有さず、腫瘍の組織学的異型度の判定に焦点を当てて訓練、検証されている。SNADET の異型度の鑑別という臨床的課題に対する研究を主目的としたた

めであるが、非腫瘍症例も含めた診断の確立は今後の研究課題であると考える。

６．結語

本 AI 分類器は、SNADET の組織学的異型度分類において良好な診断能力を有していた。AI の診断が、臨床における最適な治療方針決定に寄与する可能性が示唆された。

７．謝辞

本研究の遂行にあたり、終始多大なご協力、ご指導を賜りました山口大学名誉教授 浜本 義彦先生、徳山工業高等専門学校 荻原 宏是先生に深く感謝申し上げます。指導教官として多大なご指導賜りました五嶋 敦史先生並びに西川 潤教授、高見 太郎教授にも深謝いたします。また、研究の遂行にあたり他施設から多大なご協力を賜りました清時 秀先生、白井 保之先生、柳井 秀雄先生にも感謝の意を表します。

８．参考文献

1. Sellner F. Investigations on the significance of the adenoma-carcinoma sequence in the small bowel. Cancer 1990; 66: 702 – 715.
2. Yoshida M, Yabuuchi Y, Kakushima N, et al. The incidence of non-ampullary duodenal cancer in Japan: The first analysis of a national cancer registry. J Gastroenterol Hepatol 2021; 36: 1216 – 1221.
3. Nakagawa K, Sho M, Fujishiro M, et al. Clinical practice guidelines for duodenal cancer 2021. J Gastroenterol 2022; 57: 927 – 941.
4. Miyazaki K, Nakayama A, Sasaki M, et al. Resectability of small duodenal tumors: a randomized controlled trial comparing underwater endoscopic mucosal resection and cold snare polypectomy. Am J Gastroenterol 2024; 119: 856 – 863.
5. Sakaguchi Y, Tsuji Y, Ushiku T, et al. The natural history of sporadic non-ampullary duodenal epithelial tumors: Can we wait and see? DEN Open 2021; 1: e9.
6. Okada K, Fujisaki J, Kasuga A, et al. Sporadic nonampullary duodenal adenoma in the natural history of duodenal cancer: a study of follow-up surveillance. Am J Gastroenterol 2011; 106: 357 – 364.
7. Wang H, Sidhu M, Gupta S, et al. Cold snare EMR for the removal of large duodenal adenomas. Gastrointest Endosc 2023; 97: 1100 – 1108.
8. Kakushima N, Yoshida M, Iwai T, et al. A simple endoscopic scoring system to differentiate between duodenal adenoma and carcinoma. Endosc Int Open 2017; 5: E763 – E768.
9. Muramoto T, Ohata K, Sakai E, et al. A new classification for the diagnosis of superficial non-ampullary duodenal epithelial tumors using endocytoscopy: A prospective study. J Gastroenterol Hepatol. 2021;36(11):3170-3176. doi:10.1111/jgh.15585
10. Kinoshita S, Nishizawa T, Ochiai Y, et al. Accuracy of biopsy for the preoperative diagnosis of superficial nonampullary duodenal adenocarcinoma. Gastrointest Endosc 2017; 86: 329 – 332.
11. Yamasaki Y, Uedo N. Stop taking routine biopsy specimens for the diagnose of a duodenal adenoma! Endosc Int Open 2021; 9: E470 – E471.
12. Ishii R, Ohata K, Sakai E, et al. Simple scoring system for the diagnosis of superficial non-ampullary duodenal epithelial tumors. Dig Endosc 2021; 33: 399 – 407.
13. Inoue S, Shichijo S, Aoyama K, et al. Application of convolutional neural networks for detection of superficial nonampullary duodenal epithelial tumors in esophagogastroduodenoscopic images. Clin Transl Gastroenterol 2020; 11: e00154.
14. Dixon MF. Gastrointestinal epithelial neoplasia: Vienna revisited. Gut 2002; 51: 130 – 131.
15. Ushiku T, Arnason T, Fukayama M, Lauwers GY. Extra-ampullary duodenal adenocarcinoma. Am J Surg Pathol 2014; 38: 1484 – 1493.

16. Toba T, Inoshita N, Kaise M, et al. Clinicopathological features of superficial non-ampurally duodenal epithelial tumor; gastric phenotype of histology correlates to higher malignant potency. J Gastroenterol 2018; 53: 64 – 70.
17. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning. 2019; DOI: 10.48550/arXiv.1905.11946
18. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. 2014; DOI: 10.48550/arXiv.1409.0575.
19. Lachenbruch, PA. Estimation of error rates in discriminant analysis. Ph.D. dissertation. Univ. of California, Los Angeles; 1965.
20. Kato M, Takeuchi Y, Hoteya S, et al. Outcomes of endoscopic resection for superficial duodenal tumors: 10 years' experience in 18 Japanese high volume centers. Endoscopy 2022; 54: 663 – 670.
21. Takizawa K, Kakushima N, Tanaka M, et al. Cold snare polypectomy for superficial non-ampullary duodenal epithelial tumor: a prospective clinical trial (pilot study). Surg Endosc 2022; 36: 5217 – 5223.
22. Goda K, Kikuchi D, Yamamoto Y, et al. Endoscopic diagnosis of superficial non-ampullary duodenal epithelial tumors in Japan: Multicenter case series. Dig Endosc 2014; 26 (Suppl 2): 23 – 29.