



Title	機械学習モデルを用いた脳腫瘍病理画像解析に関する研究
Author(s)	遠田, 建
Description	配架番号 : 2977
Degree Grantor	北海道大学
Degree Name	博士(医学)
Dissertation Number	甲第16884号
Issue Date	2026-03-25
DOI	https://doi.org/10.14943/doctoral.k16884
Doc URL	https://hdl.handle.net/2115/99729
Type	doctoral thesis
File Information	ENDA_Ken.pdf, 全文



学 位 論 文

機械学習モデルを用いた
脳腫瘍病理画像解析に関する研究
(Studies on Brain Tumor Histopathological Image
Analysis Using Machine Learning Models)

2026 年 3 月

北 海 道 大 学

遠田 建

学 位 論 文

機械学習モデルを用いた
脳腫瘍病理画像解析に関する研究
(Studies on Brain Tumor Histopathological Image
Analysis Using Machine Learning Models)

2026 年 3 月

北 海 道 大 学

遠田 建

目 次

発表論文目録および学会発表目録	1 頁
要旨	2 頁
略語表	5 頁
緒言	7 頁
方法	13 頁
結果	24 頁
考察	49 頁
結論	54 頁
謝辞	54 頁
利益相反	55 頁
引用文献	56 頁

発表論文目録および学会発表目録

本研究の一部は以下の雑誌に受理された。

Ken Enda, Yoshitaka Oda, Zen-ichi Tanei, Kenichi Satoh, Hiroaki Motegi, Terasaka Shunsuke, Shigeru Yamaguchi, Takahiro Ogawa, Wang Leil, Masumi Tsuda, Shinya Tanaka

Transfer Learning Strategies for Pathological Foundation Models: A Systematic Evaluation in Brain Tumor Classification

Pathology International, accepted

本研究の一部は以下の学会に発表した。

1. 遠田建、小田義崇、種井善一、王磊、津田真寿美、田中伸哉

AI を活用した脳腫瘍の病理組織画像の分類：グリオーマ、リンパ腫、転移性脳腫瘍の識別

第 6 回日本メディカル AI 学会学術集会、2024. 6. 21-22、名古屋

2. 遠田建、小田義崇、種井善一、王磊、津田真寿美、田中 伸哉

AI による悪性脳腫瘍分類：基盤モデルの転移学習戦略比較

第 114 回日本病理学会総会、2025. 4. 17-19、仙台

3. 遠田建、小田義崇、田中伸哉

病理基盤モデルの転移学習戦略：脳腫瘍分類における体系的評価

第 7 回メディカル AI 学会、2025. 6. 28-29、京都

4. 遠田建

Introducing Pathology Foundation Models: Through Brain Tumor Classification

The 4th Hokkaido University Medical AI Symposium、2025. 8. 16、札幌

要旨

【背景と目的】

悪性脳腫瘍は発生頻度は比較的低いものの、適切な治療選択のために正確な診断が極めて重要な疾患群である。主要な悪性脳腫瘍には神経膠腫、原発性中枢神経系リンパ腫（Primary Central Nervous System Lymphoma: PCNSL）、転移性脳腫瘍が含まれ、それぞれ治療方針が大きく異なるため、正確な鑑別診断が患者の予後に直接影響を与える。2021 年 WHO（World Health Organization）中枢神経系腫瘍分類第 5 版では、成人型びまん性神経膠腫を膠芽腫（IDH（Isocitrate Dehydrogenase）野生型）、星細胞腫（IDH 変異型）、乏突起膠腫（IDH 変異型かつ 1p/19q 共欠失）の 3 タイプに分類しているが、分子診断には時間を要するため、初期治療方針決定には形態学的診断が重要な役割を果たしている。

病理診断の標準化と客観化を目指し、近年では機械学習技術、特に病理基盤モデルの応用が期待されている。2023 年以降、DINOv2 アルゴリズムを基盤とした次世代の基盤モデル（UNI、Prov-GigaPath 等）が相次いで登場し、従来モデルを大幅に上回る性能を示している。しかし、これらの基盤モデルを実際の分類タスクに適用する際の最適な転移学習戦略（Fine-tuning (FT) と Linear Probing (LP)）については、特に脳腫瘍分類に特化した体系的な比較研究が限定的である。

本研究では、施設内データセットと外部データセット（EBRAINS）を用い、次世代基盤モデル（UNI、Prov-GigaPath）と従来型モデル（CTransPath、ImageNet 事前学習 ViT-L（Vision Transformer-Large）、ResNet-RS 50）について、FT と LP の適用戦略を比較検討し、実臨床環境における基盤モデルの最適な訓練戦略を明らかにすることを目的とした。

【方法】

北海道大学医学部腫瘍病理学教室で 2016 年から 2023 年に収集した 254 症例（膠芽腫 66 例、星細胞腫 29 例、乏突起膠腫 24 例、PCNSL 52 例、転移性腫瘍 81 例、背景組織 5 例）から施設内データセットを構築した。H&E（Hematoxylin and Eosin）染色された永久標本をデジタル化し、腫瘍領域から 512×512 ピクセル以上のパッチに分割、計 368,287 枚のパッチを作成した。外部検証用に EBRAINS データセットから 698 症例を収集した。

5 種類のエンコーダー (UNI、Prov-GigaPath、CTransPath、ViT-L、ResNet-RS 50) を比較し、3 つの訓練条件 (FT: エンコーダーと全結合層を同時に更新、LP: 事前学習済みエンコーダーを固定し全結合層のみ更新、Random Init: 事前学習なしでランダム初期化) を設定した。5 分割交差検証を用い、症例あたりの訓練パッチ数を 500、100、25、10 枚と変化させて訓練と評価を行った。評価指標として、マクロ平均 Recall、パッチ精度、Overall Accuracy を算出し、GradCAM (Gradient-weighted Class Activation Mapping) と UMAP (Uniform Manifold Approximation and Projection) による可視化解析も実施した。

【結果】

施設内データセット (500 パッチ/症例) では、基盤モデル UNI(LP)、UNI(FT)、Prov-GigaPath(LP) が全指標で堅牢な性能を示し、細粒度分類でマクロ平均 Recall 0.87-0.88、パッチ精度 0.93-0.94 を達成した。粗粒度分類では Prov-GigaPath(LP) が最高性能 (254 例中 251 例正解) を示した。

外部データセット (EBRAINS) では注目すべき結果が得られた。UNI(LP) が全指標で優れた性能を示し (細粒度 : マクロ平均 Recall 0.75 ± 0.01 、粗粒度 : 0.94 ± 0.01)、UNI(FT) を有意に上回った ($p < 0.01$)。この逆転現象は次世代基盤モデルに特有で、従来型モデル (ViT-L_{IN}、CTransPath) では依然として $FT \geq LP$ の関係が維持された。

訓練パッチ数の影響分析では、基盤モデルは 10 パッチ/症例という少量データでも堅牢な性能を示し、UNI(LP) の 10 パッチ条件は、従来モデルの 500 パッチ条件と同等以上の性能を達成した。一方、パッチ数増加に伴う性能向上は基盤モデルで限定的であり、外部データセットでは性能低下も観察された。

UMAP 解析では、事前学習済み UNI は学習前から腫瘍タイプごとの明確な分離を示したが、施設内データセットでの追加学習後は過剰に分離されたクラスタリングを示し、過適合の視覚的証拠となった。GradCAM 解析では、モデルが組織学的に関連する特徴 (膠芽腫の微小血管増殖、PCNSL の血管周囲浸潤等) を適切に学習していることが確認された。

【考察】

本研究の重要な発見は、外部データセットにおいて次世代病理基盤モデル UNI で LP が FT を有意に上回った点である。この性能逆転は、大規模データで獲得された汎用的認識能力が、単一施設データでの FT により過適合を起こし、汎化

能力を失う「破滅的忘却」を示唆する。UMAP 解析でも、FT 後の過剰なクラス
タ分離が視覚的に確認された。さらに、UNI(LP)は 10 パッチ/症例という少量デ
ータでも、500 パッチの従来モデルと同等の性能を達成した。LP は計算資源の
要求も桁違いに少なく、実臨床での実装可能性が高い。

これらの知見は、病理 AI 開発における「独自データでエンコーダーを訓練する」
から「事前学習済みエンコーダーに少量データで問い合わせる」というパラダ
イムシフトを示している。すなわち、大規模な病理データ収集から、基盤モデ
ルに埋め込まれた汎用的な視覚認識フレームワークを効率的に活用する方向へ
の転換である。

【結論】

本研究は、病理画像分類における次世代病理基盤モデルの適応戦略の包括評価
を通じて、外部データセットへの汎化において LP が FT を上回ることを実証し、
データセット設計における新たなパラダイムシフトを示した。これらの知見は、
より効率的で信頼性の高い医療 AI システムの開発に向けた理論的基盤として貢
献するものである。

略語表

本文中および図中で使用した略語は以下の通りである。

A	Astrocytoma class
AA	Anaplastic Astrocytoma
AE1/AE3	Anti-Cytokeratin antibodies 1 and 3
AI	Artificial Intelligence
AO	Anaplastic Oligodendroglioma
CBTRUS	The Central Brain Tumor Registry of the United States
CD20	Cluster of Differentiation 20
CI	Confidence Interval
CNN	Convolutional Neural Network
CNS	Central Nervous System
DA	Diffuse Astrocytoma
DINO	Self-Distillation with No Labels
DINOv2	DINO version 2
EBRAINS	European Brain Research Infrastructures
EWC	Elastic Weight Consolidation
FN	False Negative
FP	False Positive
FT	Fine-Tuning
G	Glioblastoma class
GBM	Glioblastoma
GFAP	Glial Fibrillary Acidic Protein
GradCAM	Gradient-weighted Class Activation Mapping
H&E	Hematoxylin and Eosin
IDH	Isocitrate Dehydrogenase
IDH+	IDH-mutant
IDH-	IDH-wildtype
IN	ImageNet
L	Lymphoma class (PCNSL)
LLM	Large Language Model
LoRA	Low-Rank Adaptation
LP	Linear Probing

M	Metastatic tumor class
MIL	Multiple Instance Learning
NDP	NanoZoomer Digital Pathology
O	Oligodendroglioma class
PCNSL	Primary Central Nervous System Lymphoma
PEFT	Parameter-Efficient Fine-Tuning
REMEDIS	Robust and Efficient Medical Imaging with Self-supervision
RI	Random Initialization
ROI	Region of Interest
SD	Standard Deviation
SE	Standard Error
SSL	Self-Supervised Learning
STD	Standard Deviation for normalization
TN	True Negative
TP	True Positive
UMAP	Uniform Manifold Approximation and Projection
UNI	Universal Foundation Model
ViT	Vision Transformer
ViT-L	Vision Transformer Large
ViT-G	Vision Transformer Giant
WHO	World Health Organization
WSI	Whole Slide Image

緒言

悪性脳腫瘍

悪性脳腫瘍は、その発生頻度は比較的低いものの、適切な治療選択のために正確な診断が極めて重要な疾患群である(Ostrom et al., 2023)。主要な悪性脳腫瘍には、神経膠腫、原発性中枢神経系リンパ腫 Primary Central Nervous System Lymphoma: PCNSL)、および転移性脳腫瘍 (Metastatic brain tumor) が含まれる。米国中枢脳腫瘍登録局 (Central Brain Tumor Registry of the United States: CBTRUS) と日本の国立がん登録のデータを比較すると、原発性脳腫瘍における主要腫瘍の割合は、膠芽腫 (Glioblastoma) (米国 14.2%、日本 7.0%)、星細胞腫 (Astrocytoma) (米国 3.1%、日本 2.5%)、乏突起膠腫 (Oligodendroglioma) (米国 1.2%、日本 1.3%)、PCNSL (米国 1.9%、日本 3.9%) となっている。しかし、実際の病理診断現場では転移性脳腫瘍も含めた鑑別が重要となる (Villanueva-Meyer et al., 2017)。日本脳神経外科学会の報告によると、手術症例における頻度は原発性腫瘍とは異なる分布を示し、脳腫瘍手術症例における頻度は転移性脳腫瘍が 30.6%と最も高く、次いで膠芽腫 13.0%、星細胞腫 4.7%、PCNSL 4.9%、乏突起膠腫 1.9%となっている (Iihara et al., 2021)。これらの割合は実際の脳腫瘍病理診断における事前確率に近いと考えられ、診断アプローチを考える上で重要な指標となる。また、これらの腫瘍は治療方針が大きく異なるため、正確な鑑別診断が患者の予後に直接影響を与える可能性がある (van den Bent et al., 2023)。

悪性脳腫瘍の病理診断

悪性脳腫瘍の正確な診断は、適切な治療方針の決定に直結するため極めて重要である。病理診断では、H&E (Hematoxylin and Eosin) 染色による形態学的観察に加え、免疫組織化学的マーカーが補助的に用いられる。神経膠腫の診断にはアストロサイト系マーカーである GFAP (Glial Fibrillary Acidic Protein) が、PCNSL の大部分を占めるびまん性大細胞型 B 細胞リンパ腫には B 細胞系マーカーの CD20 (Cluster of Differentiation 20) が、転移性脳腫瘍の原発巣推定には上皮系マーカーの CK AE1/AE3 (Cytokeratin Anti-Cytokeratin antibodies 1 and 3) がそれぞれ用いられる。これらのマーカーの組み合わせにより、悪性脳腫瘍の主要な 3 病型である神経膠腫、PCNSL、転移性脳腫瘍の鑑別診断が可能となる。

特に術中迅速診断においては、限られた時間内での正確な判断が要求され、病理医の経験と技術に大きく依存する。術中迅速診断では凍結標本を対象とした免疫染色を行う機器は存在するものの導入施設はほとんどなく、H&E 染色等

による形態学的観察のみで神経膠腫、PCNSL、転移性脳腫瘍の鑑別を必要とされるのが現状である。一方、永久標本では免疫染色による詳細な検討が可能であるが、形態学的特徴の正確な評価には依然として病理医の専門的判断が不可欠である。

WHO 分類と分子病理学的診断

2021 年に発表された WHO 中枢神経系腫瘍分類第 5 版では、成人型びまん性神経膠腫を IDH 変異状態と 1p/19q 共欠失の有無により 3 タイプに分類している。Glioblastoma, IDH-wildtype は、IDH1/2 遺伝子に変異を持たない最も悪性度の高い神経膠腫であり、原発性に発生することが多く、予後不良である。Astrocytoma, IDH-mutant は、IDH1 または IDH2 遺伝子に変異を有し、1p/19q 共欠失を伴わない腫瘍で、野生型と比較して予後は良好である。Oligodendroglioma, IDH-mutant and 1p/19q-codeleted は、IDH 変異に加えて染色体 1p 短腕と 19q 長腕 (1p/19q) の共欠失を特徴とし、化学療法への感受性が高く、3 タイプの中で最も予後良好である。

この分類に基づき、さらに組織学的特徴と遺伝子検査結果を統合して CNS WHO grade が決定される。CNS WHO grade 2 は低悪性度腫瘍であり、細胞密度の軽度上昇と核異型を示すが、血管増殖や壊死は認められない。CNS WHO grade 3 は退形成性腫瘍として分類され、細胞密度の著明な上昇、核異型の増強、有糸分裂像の増加を特徴とする。CNS WHO grade 4 は最も悪性度が高く、微小血管増殖または壊死の存在により定義され、IDH 野生型膠芽腫では特定の分子マーカーの存在によっても grade 4 と診断される。特に、IDH 変異と 1p/19q 共欠失の検出には、サンガーシーケンシングや多重連鎖依存プローブ増幅 (Multiplex Ligation-dependent Probe Amplification: MLPA) 法などの分子診断技術が必須となっている。これらの分子診断は治療方針の決定に重要な役割を果たすが、結果を得るまでに時間を要するため、初期治療方針の決定には形態学的診断が重要な役割を果たしている。

治療方針の決定と予後

悪性脳腫瘍の治療戦略は、病理診断結果に大きく依存する。膠芽腫に対しては、最大限の安全な切除後に化学放射線療法が推奨される。低リスクのグレード 2 腫瘍 (多くの星状細胞腫と乏突起膠腫を含む) では、外科的切除後の経過観察が標準的アプローチとなる。一方、高リスクのグレード 2 およびグレード 3 腫瘍では、切除後に補助療法が行われることが多い (van den Bent et al., 2023)。PCNSL に対しては、生検後に導入療法と地固め療法からなる複合的治療が標準

的である(Sasaki & Nagane, 2021; Nakamura et al., 2022)。脳転移の治療は、原発巣の特性や患者の全身状態、症状に基づいて個別化され、手術、定位放射線治療、全脳放射線治療、全身療法を適切に組み合わせて行われる(Vogelbaum MA et al., 2023)。

悪性脳腫瘍診断における課題と機械学習技術への期待

悪性脳腫瘍は全悪性腫瘍の約2%を占める希少疾患であり(Ostrom et al., 2023)、その診断には脳神経外科医と神経病理医の緊密な連携が不可欠である。しかし、これらの専門医は限られた施設に集中しており、診断の質に施設間格差が生じやすい。永久標本による病理診断においても、形態学的特徴の評価は病理医の経験に依存し、分子診断結果の解釈と統合診断には高度な専門知識が要求される。このような背景から、診断プロセスの標準化と客観化を目指して、機械学習技術の応用が期待されている。以下では、病理診断支援における機械学習技術の発展と、特に脳腫瘍診断への応用について述べる。

機械学習の基本原則

近年の画像認識技術の飛躍的な進歩により、病理画像解析への人工知能(Artificial Intelligence: AI) 応用が注目を集めている(Redlich et al., 2024)。病理画像分類に用いられる機械学習の分類器モデルは基本的に、エンコーダー（特徴抽出層）と全結合層の2つの要素から構成される(図1)。エンコーダーは入力された病理画像から腫瘍の形態学的特徴を抽出し、1次元の固定長ベクトル（特徴量または埋め込みベクトルと呼ばれる）として数値化する。この特徴量は全結合層に渡され、最終的な診断カテゴリー（神経膠腫、PCNSL、転移性脳腫瘍など）の確率分布へと変換される。逆誤差伝播法によってエンコーダと全結合層のパラメータを最適化することで、訓練データセットに含まれるクラスのパターンを判別するモデルを獲得できる。この基本原則が、病理診断支援システムの技術的基盤となっている。

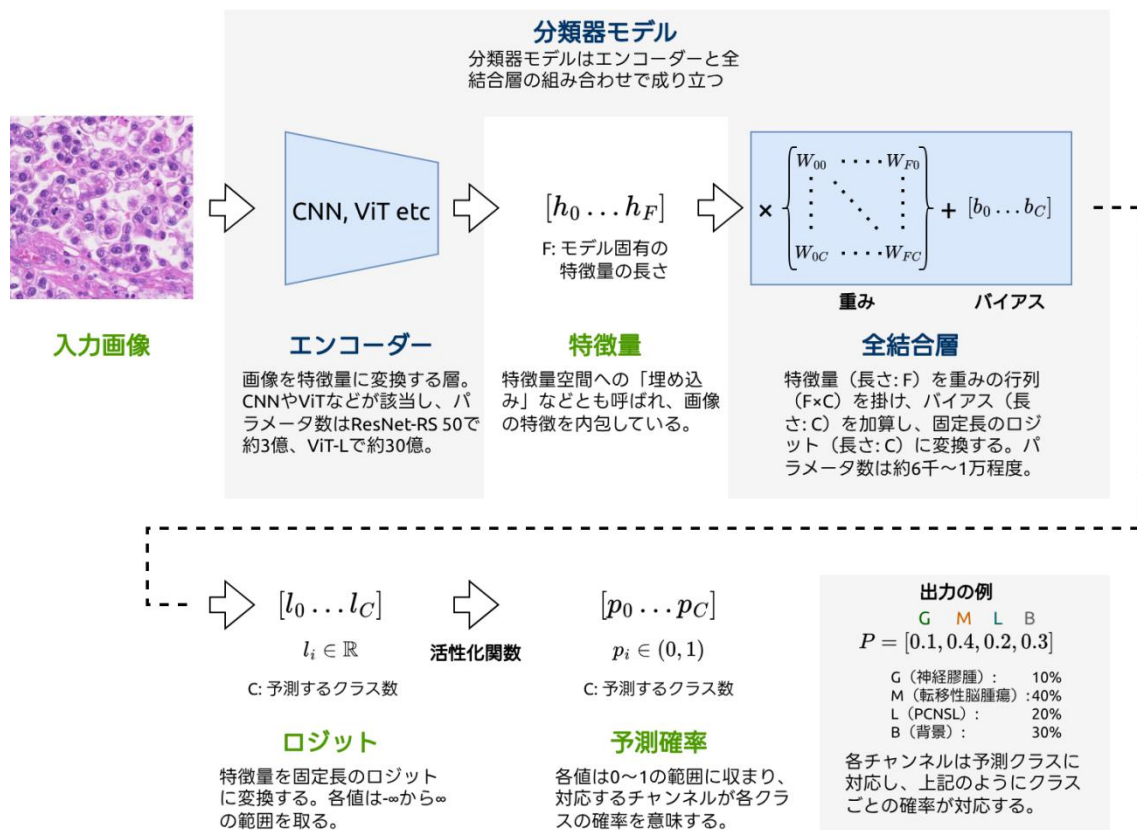


図1 機械学習分類器モデルの基本構造

病理診断への機械学習技術の応用

分子診断技術は治療選択に重要な情報を提供するものの、結果取得までの時間とリソースの制約が、特に初期治療計画において実践的な課題となっている。この文脈において、H&E 染色標本から形態学的特徴を自動的に抽出し、迅速かつ費用対効果の高い診断支援を提供する可能性のある脳腫瘍病理画像解析へのAI 応用への関心が高まっている。

病理 AI 研究の初期段階では、主に単一腫瘍タイプの分類や特定の特徴の検出に焦点が当てられていた。脳腫瘍分野では、IDH 変異に基づく神経膠腫のサブタイプ分類とグレーディングを中心とした研究が進められてきた(Hollon et al., 2023; Jin et al., 2021; W. Wang et al., 2023)。しかし、これらの先行研究の多くは神経膠腫内でのサブタイプ分類に限定されており、転移性脳腫瘍やリンパ腫を含む包括的な研究は限定的であった(Ma et al., 2023)。

病理学基盤モデルの発展

病理診断においては、症例の希少性や診断基準の複雑性から大規模なラベル付きデータセットの構築が困難であり、自己教師あり学習 (Self-Supervised Learning: SSL) による教師なし事前学習が特に有効なアプローチとなる。病理画像解析における初期の SSL モデルである CTransPath(X. Wang et al., 2022)や REMEDIS(Azizi et al., 2023)は、ImageNet で事前学習されたモデルと比較して病理画像特有の特徴を効果的に学習できることを実証し、病理 AI 分野における転移学習の有用性を確立した。しかし、これらの従来型 SSL モデルは、学習データの規模や多様性、SSL 手法自体に制約があった。

2023 年、DINOv2(Oquab et al., 2023)の登場は、SSL 手法における転換点となった。DINOv2 は、マスクドイメージモデリング、自己蒸留、コントラスト学習など、従来の SSL 手法を洗練・統合することで、意味のある特徴表現をより効率的に学習することを可能にした。この技術的革新を礎として、病理学分野で次世代の病理基盤モデルが相次いで登場した。

UNI(Chen et al., 2024)は Vision Transformer (ViT) アーキテクチャ(Dosovitskiy et al., 2020)の Large 版 (ViT-L) に基づき、10 万枚の WSI (Whole Slide Image) から抽出された 1 億枚を超える病理画像パッチを用いて DINOv2 アルゴリズムによって訓練され。Prov-GigaPath(Xu et al., 2024)は、ViT-Giant (ViT-G) に基づき、17 万枚の WSI から抽出された 13 億枚のパッチで訓練された。Virchow(Vorontsov et al., 2024)や Virchow2(Zimmermann et al., 2024)も、同様の ViT アーキテクチャと DINOv2 手法で訓練された病理基盤モデルとして相次いで登場した。

これら次世代モデルはいずれも前例のない規模で病理学的特徴を学習しており、従来の CTransPath や REMEDIS と比較して、ほぼすべての病理診断タスクにおいて有意な性能向上を示している。具体的には、UNI は腫瘍分類タスクにおいて、CTransPath と比較して AUC (Area Under the Curve) で 5-10%の改善を達成している。

病理基盤モデルの適用手法

基盤モデルなどの訓練済みモデルを実際の分類タスクに適用する際には、様々な転移学習戦略が検討されているが、Fine-tuning (FT) と linear probing (LP) が 2 つの主要なアプローチとなっている(Kim et al., 2022)。FT は大量のパラメータから構築されるエンコーダおよび全結合層の両者を訓練の過程ですべて更新する手法であるのに対し、LP は訓練済みエンコーダのパラメータは固定し全

結合層のみパラメーターを訓練の過程で更新する手法である。これらの戦略は従来の深層学習モデルで広く研究されてきたが、病理基盤モデル、特に脳腫瘍分類に特化した体系的な比較研究は限定的である。この橋渡し研究を試みることで、本研究の主要な動機の一つとなっている。

本研究の意義と目的

本研究では、まず北海道大学医学部腫瘍病理学教室で診断された症例を収集し、施設内データセットを構築した。外部データセットとして、EBRAINS(Roetzer-Pejrimovsky et al., 2022)データセットを用意した。施設内データセットを訓練データセット、EBRAINS データセットを外部の検証用データセットとして設定した。使用モデルとして、次世代の病理基盤モデルである UNI と Prov-GigaPath、従来型の病理基盤モデルである CTransPath、また ImageNet(Deng et al., 2009)で事前学習された ViT-L(Dosovitskiy et al., 2021)モデルと、CNN (Convolutional Neural Network) モデルとして ResNet-RS 50(Bello et al., 2021)モデルを用意した。これらのモデルの適用戦略として、それぞれ FT と LP アプローチで訓練し、性能指標を比較する。さらに訓練に使用する画像の枚数を変化させ、さらに各モデル・条件について比較する。

この検討を通じて、実臨床環境における病理基盤モデルの最適な訓練戦略を明らかにし、特に外部データセットに適用する際の汎化能力に焦点を当てる。さらに、実用的な応用の実証として、脳腫瘍病理画像のインタラクティブな分類のためのウェブベースシステムのプロトタイプも開発した。

方法

1. 施設内データセット

2016 年から 2023 年にかけて北海道大学医学部腫瘍病理学教室で収集した 249 症例とを後ろ向きに解析した。この悪性脳腫瘍は 254 症例で、Glioblastoma, IDH-wild type (G) 66 例、Astrocytoma, IDH-mutant (A) 26 例、Oligodendroglioma, IDH-mutant, 1p/19q co-deleted (O) 24 例、PCNSL (L) 52 例、転移性脳腫瘍 (M) 81 例の 5 種類である。さらに、背景組織として 5 症例の病理解剖から得た非腫瘍脳組織を含めた (表 1、図 2)。

Local dataset	Coarse-grained class	Fine-grained class	Cases	Patch count
Gliomas			119	73,028
Glioblastoma, IDH-wildtype	Glioma	G	66	47,943
Astrocytoma, IDH-mutant	Glioma	A	26	11,840
Oligodendroglioma, IDH-mutant, 1p/19q co-deletion	Glioma	O	24	13,245
Brain metastases	M	M	81	45,370
PCNSL	L	L	52	16,988
Background brain	B	B	5	159,873
total			254	368,287

表 1 施設内データセットの症例数とパッチ数の内訳

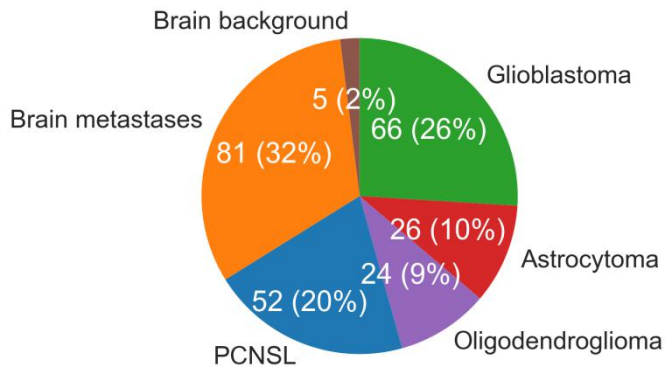


図 2 施設内データセットの症例数の内訳の円グラフ

分類タスクでは、5 種の腫瘍と 1 種の背景組織を合わせた 5+1=6 クラス分類タスクを「細粒度分類」(fine-grained classification)と定義した。また、A クラスと O クラスと G クラスは広義では Glioma にあたることから、これらを Glioma とまとめた 4=3+1 クラス分類タスクを「粗粒度分類」(coarse-grained classification)として定義した。

神経膠腫の分類は WHO 2021 基準に基づき、IDH 変異の有無については IDH1-R132H 免疫組織化学検査とサンガーシーケンス法、1p/19q 共欠失については MLPA 法を用いて評価した。すべての診断は当教室の病理専門医によって行われた。2021 年以降の症例は WHO 第 5 版基準に従っているが、それ以前の症例については IDH 変異および 1p/19q 共欠失について前述の手法によって評価を行い現行の診断基準に再分類した。品質管理のため、検体量が極度に少量の症例、IDH 変異の遺伝子検査が未施行の症例、1p/19q 共欠失が腫瘍内部で領域による不均一性が見られるような特殊な症例は除外した。標本作製では、ホルマリン固定パラフィン包埋組織から永久標本を作製し、H&E 染色を施した。

浜松ホトニクス株式会社(浜松, 日本)製の Hamamatsu NanoZoomer® XR および S210 スキャナーで標本スライドをデジタル化し、全スライド画像(Whole Slide Image: WSI)を取得した。同社が提供している WSI ビュアーである NDP.view2 を用いて 20 倍拡大(約 440 nm/pixel)で腫瘍領域を視覚的に選択し腫瘍領域の画像を獲得し、これらを 512×512 pixel 以上のパッチに分割した(図 3)。腫瘍クラスでは白色領域(グレースケール変換時に色調が 230-255 のピクセル)が 40%を超えるパッチを除去した(図 4)。一方、背景組織クラスでは非腫瘍領域を背景組織として認識させるため、意図的にこの処理を省略した。各症例のパッチ数の分布は症例により異なった(図 5)。画像の前処理として、ImageNet(Deng et al., 2009)と同じ正規化(MEAN = (0.485, 0.456, 0.406)、STD = (0.229, 0.224, 0.225))を適用した。データ拡張には Albumentations ライブラリ(Buslaev et al., 2020)を使用し、幾何学的変換、輝度・色調整、ノイズシミュレーション、構造的修正を行った。入力サイズは 224×224 pixel(約 1μm/pixel)に設定した。

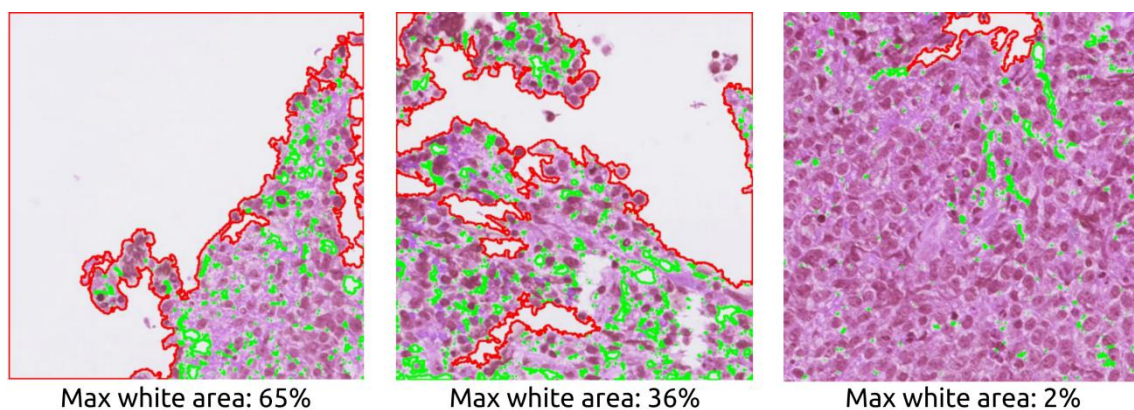


図 3 白色領域の検出

OpenCV を用いて白色領域を検出した。区切られた白色領域のうち、1000 pixel 以上の領域を赤線で囲み、それ未満の領域を緑線で囲んだ。画像下部には最大の白色領域の割合を示す。

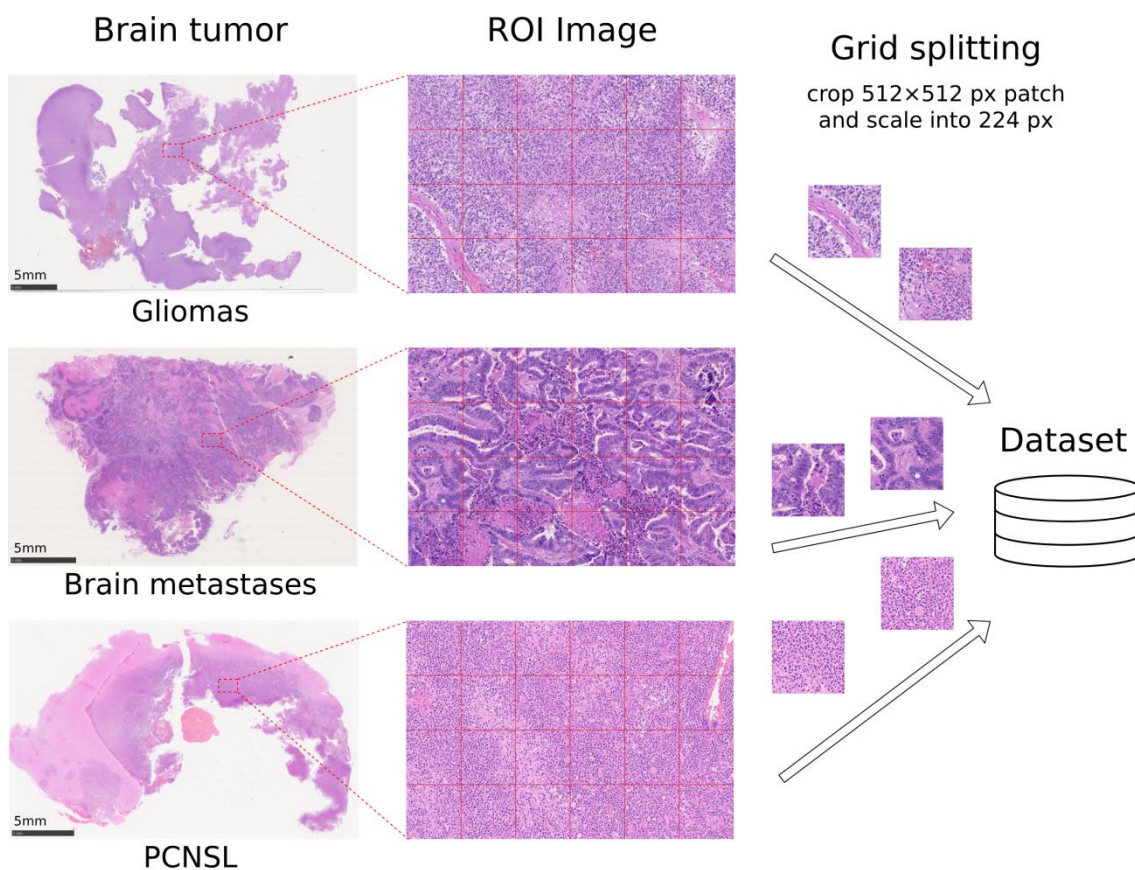


図 4 データセット作成の概略図

脳腫瘍の病理組織の WSI データをスキャン(空間解像度: 約 $0.440 \mu\text{m}/\text{pixel}$)し、腫瘍占拠領域を目視で選択し ROI 画像を取得した。その後格子状に分割しデータセットとした。学習・検証時にはオンラインで $512 \times 512 \text{ pixel}$ でクロップし、 $224 \times 224 \text{ pixel}$ にリサイズした(空間解像度:

約 $1.0 \mu\text{m}/\text{pixel}$)

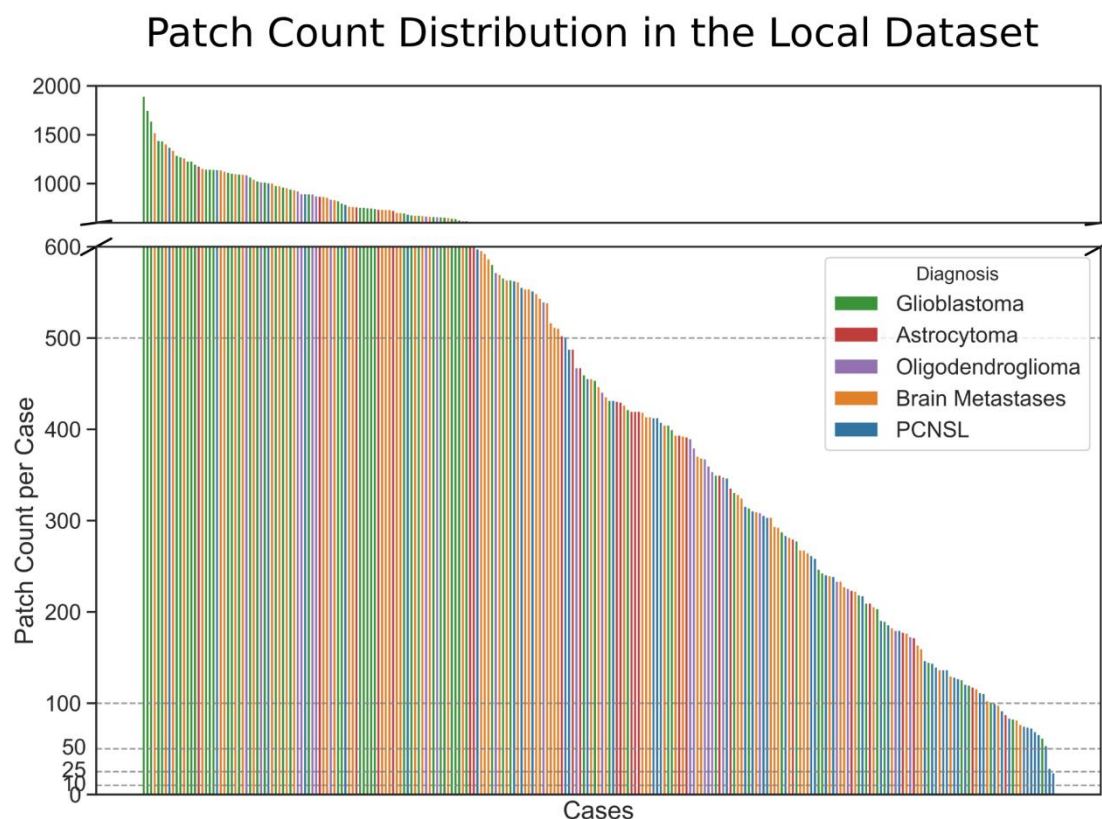


図 5 施設内データセット内での症例ごとのパッチ数の分布

1 症例ごとのパッチ数を多い順に左から並べた。訓練時に使用するパッチ枚数を 500 枚、100 枚、50 枚、25 枚、10 枚と変化させその精度変化を比較したため、その基準に補助線を描画した。

2. 機械学習モデル

画像特徴の抽出には、5 種類のエンコーダーを比較した：2024 年の次世代の病理基盤モデルである UNI(Chen et al., 2024)と Prov-GigaPath(Xu et al., 2024)、従来のベースラインモデルである CTransPath(Wang et al., 2022)、そして ViT-L(Dosovitskiy et al., 2021)と ResNet-RS 50(Bello et al., 2021)である。Prov-GigaPath は ViT-G アーキテクチャに、UNI は ViT-L アーキテクチャに基づいており、いずれも Hugging Face (Hugging Face, Inc., New York, NY, USA; <https://huggingface.co/>) で公開されている事前学習済み重みを使用した。ViT-L については UNI と同じ ViT-L モデルを使用し、ImageNet-21k で事前学習されたパラメータを timm ライブラリ(Wightman, 2019)から取得した (内部識別名: vit_large_patch16_224.augreg_in21k_ft_in1k)。ResNet-RS 50 は ImageNet-1k で事

前学習されたパラメータを同じく `timm` ライブラリから取得した（内部識別名：`resnetrs50.tf_in1`）。これらの ImageNet で事前訓練された条件には「ResNet-RS 50_{IN}」のように IN（ImageNet）の接尾語を付けて表記する。

3. 適用条件の検討

事前学習済みモデルを下流タスクに適用する際の転移学習戦略は、モデルの性能と計算効率のトレードオフを決定する重要な要素である。さらにモデルをタスクに適用する際に、3 つの訓練条件を定義した（図 6）：

- **Linear Probing (LP)**：事前学習済みエンコーダーのパラメータを固定し、全結合層のみ更新
- **Fine-tuning (FT)**：事前学習済みエンコーダーと全結合層のパラメータを同時に更新
- **Random Init (RI)**：事前学習なしでエンコーダーをランダム初期化し、全パラメータを更新

そして、これらをモデル条件に「ViT-L_{IN}(FT)」のようにさらなる接尾語として表記する。

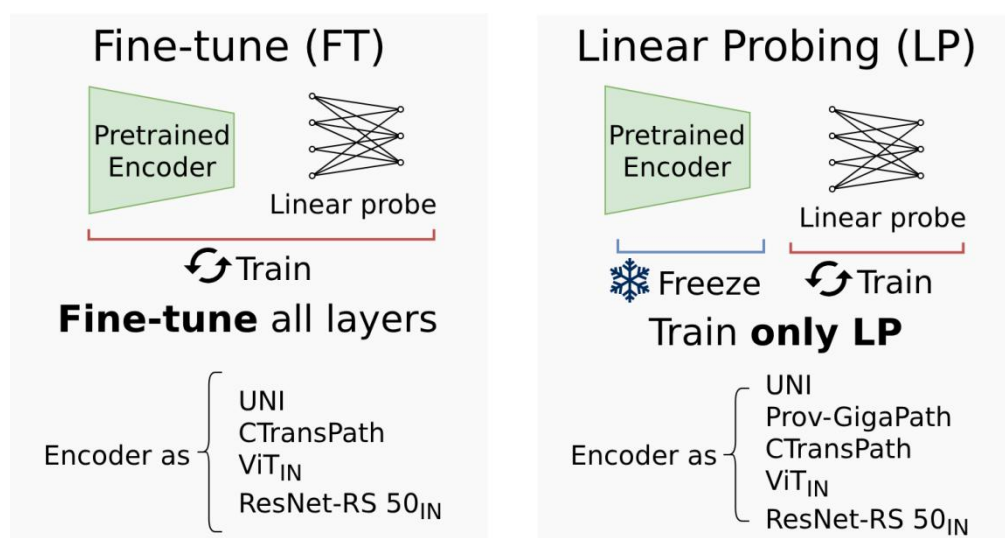


図 6 訓練時の機械学習モデルのパラメータ更新の手の比較

FT (Fine-tune) は、エンコーダーと全結合層 (Linear probe) の両方を同時に更新する。一方で、LP (Linear probe) は、エンコーダーのパラメータを固定 (凍結) し更新を行わず、全結合層のみパラメータを更新する。

これらの戦略の選択は、実用上重要な意味を持つ。FT では、エンコーダー（例：ViT-L で約 3 億パラメーター）と全結合層（6 分類モデルで約 6,000 パラメータ

一) の両方を更新するため、勾配計算とパラメーター更新に大きな計算資源を要する。一方、LP ではエンコーダーを固定し全結合層のみを更新するため、勾配計算が全結合層に限定され、計算量を大幅に削減できる。具体的には、LP ではエンコーダーを一度だけ実行して特徴量を抽出し、その後は全結合層の学習のみを行うため、訓練時間を大幅に短縮できる。LP で FT に匹敵する性能を達成できれば、限られた計算資源での効率的なモデル開発が可能となり、臨床現場での実装可能性が高まる。

さらに、RI は事前学習の効果を検証するための重要なベースラインとなる。事前学習済みモデルとランダム初期化モデルの性能差は、病理画像における転移学習の有効性を定量的に示す指標となる。

本研究では、Prov-GigaPath(LP)、UNI(FT)、UNI(LP)、CTransPath(FT)、CTransPath(LP)、ViT-L_{IN}(FT)、ViT-L_{IN}(LP)、ViT-L(RI)、ResNet-RS 50_{IN}(FT)、ResNet-RS 50_{IN}(LP)の 10 条件を比較した。各モデルのエンコーダと LP 層のパラメータ数を表に示した (表 2)。LP 層のパラメータ数はエンコーダ層の出力する特徴量の長さ: F と出力するクラス数: C を用いて以下のように求められる。

$$LP's\ parameter\ count = F \times C + C$$

なお、Prov-GigaPath は 10 億パラメータを超える ViT-G モデルに基づいており、FT を行うにはメモリ要求が過大で、同一環境ではバッチサイズ 10 で訓練することができなかったため、FT 条件は比較から除外した。

Model name	Encoder's parameter count	LP's parameter count	FT	LP	RI
ResNet-RS 50	33,642,912	12,288	✓	✓	×
ViT-L	303,301,632	6,144	✓	✓	✓
CTransPath	27,520,038	4,608	✓	✓	×
UNI (ViT-L)	303,350,784	6,144	✓	✓	×
Prov-GigaPath (ViT-G)	1,134,953,984	9,216	×	✓	×

表 2 各モデルのパラメータ数と付随する LP のパラメータ数

画像を特徴量に変換するエンコーダ層のパラメータ数と、特徴量を出力の 6 分類の各クラスの確率に変換するための LP 層のパラメータ数、FT と LP と RI の各条件での実験の実施の可否をまとめた。

3. 分類器機械学習モデルの訓練

5 分割交差検証（図 7）を用い、1 症例あたりの訓練に使用するパッチ枚数を変化させて（500、100、25、10 枚）訓練と評価を行った。パッチ数が上限を超える症例では超過分をランダムに除外し、上限に満たない症例では画像を繰り返し使用した。モデル構造として、全エンコーダーは単一の全結合層（サイズ $F \times C$ 、 F はエンコーダーの出力特徴次元、 $C=6$ ）に接続し、活性化関数には softmax 関数を使用した。粗粒度分類では、G、A、O クラスの確率を合計して Glioma クラスとして表現した。訓練設定として、出力された確率とデータセットで定義された正しい診断クラスとの損失をクロスエントロピー損失関数で計算した。訓練にはバッチサイズ 50、Adam オプティマイザー(Kingma & Ba, 2015)を使用し、学習率は訓練中固定で、LP では 0.001、ViT ベースモデル（UNI、CTransPath、ViT-L）では 0.00001、CNN モデルでは 0.0001 とした。全条件で 30 エポック訓練し、検証損失が 5 エポック連続で改善しない場合は早期停止を適用した。

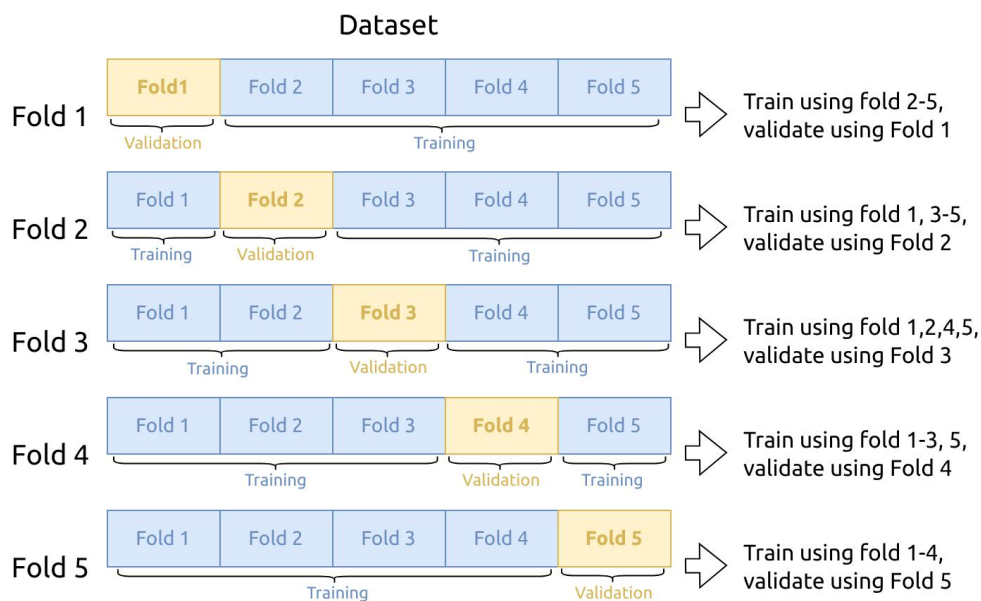


図 7 5 分割交差検証の概念図

データセット全体を 5 等分に分割し、それぞれを Fold 1-5 とする。『Fold 1』の条件では Fold 2-4 でモデルを訓練し、Fold 1 で検証を行い性能評価を行う。同様に『Fold 2』では Fold 1, 3-5 で訓練し、Fold 2 で検証を行い性能評価を行う。これを繰り返し各 Fold の評価指標を獲得し、平均値や標準偏差等を算出する。

4. 外部検証用データセット

EBRAINS(Roetzer-Pejrimovsky et al., 2022)データベースから、施設内データセットに対応する診断の Whole Slide Image (WSI) データを収集した。各診断カテゴリーでは、データ品質と計算資源を考慮しつつバランスの取れた代表性を維持することを目指した：

- Glioblastoma (G) クラス：「Glioblastoma, IDH-wildtype」199 例（474 例中、技術的制約により 200 例をダウンロードし、データ破損の 1 例を除外）、「Anaplastic astrocytoma, IDH-wildtype」47 例（全 47 例）、「Diffuse astrocytoma, IDH-wildtype」19 例（全 19 例）の計 265 例。後者 2 カテゴリーは、特定の分子的特徴を示す場合、WHO 第 5 版基準では「Molecular glioblastoma」に分類される。
- Astrocytoma (A) クラス：「Diffuse Astrocytoma, IDH-mutant」70 例（全 70 例）、「Anaplastic astrocytoma, IDH-mutant」47 例（全 47 例）、「Glioblastoma, IDH-mutant」34 例（全 34 例）の計 151 例。
- Oligodendroglioma (O) クラス：「Anaplastic oligodendroglioma, IDH-mutant and 1p/19q codeleted」91 例（全 91 例）、「Oligodendroglioma, IDH-mutant and 1p/19q codeleted」85 例（全 85 例）の計 176 例。
- Brain metastasis (M) クラス：「Metastatic tumours」47 例（全 47 例）。
- PCNSL (L) クラス：「Diffuse large B-cell lymphoma of the CNS」59 例（60 例中、ファイル破損の 1 例を除外）。

EBRAINS データセットを配信しているサーバーにネットワーク上の問題が有り、データアップロード速度が非常に低速かつ一括ダウンロードの際に不規則に接続が切断される現象が頻発したため、すべての症例をダウンロードすることができなかった。実験に使用した症例の詳細な内訳を表 3 および図 8 に示す。

EBRAINS dataset	Class	Used	Registered
Gliomas		592	867
Glioblastoma, IDH-wildtype		265	540
Glioblastoma, IDH-wildtype	G	199	474
Anaplastic astrocytoma, IDH-wildtype	G	47	47
Diffuse astrocytoma, IDH-wildtype	G	19	19
Astrocytoma, IDH-mutant		151	151
Diffuse astrocytoma, IDH-mutant	A	70	70
Anaplastic astrocytoma, IDH-mutant	A	47	47
Glioblastoma, IDH-mutant	A	34	34
Oligodendroglioma, IDH-mutant, 1p19q co-deletion		176	176
Anaplastic oligodendroglioma, IDH-mutant and 1p/19q codeleted	O	91	91
Oligodendroglioma, IDH-mutant, 1p19q co-deletion	O	85	85
Metastases brain tumor	M	47	47
PCNSL	L	59	60
total		698	974

表 3 EBRAINS データセットから今回の研究使用した症例数とリポジトリの登録された症例数
レジストリは WHO4th 以前の基準で登録されており、IDH 変異の有無から現行の診断基準に
照らし合わせて診断を読み変えてデータセットに登録した。

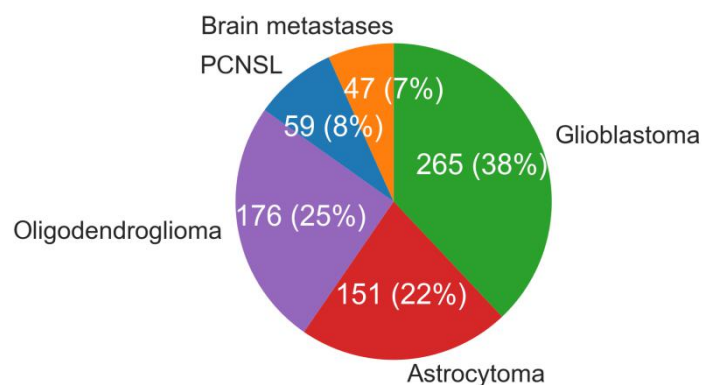


図 8 EBRAINS データセットから使用した症例の内訳

「Glioblastoma, IDH-wildtype」グループからの症例選択は、相対的に十分量の症例数が登録されていたため、ダウンロードに成功した 200 例のうち、データが破損していた 1 例を除外し 199 例を使用した。また PCNSL も 1 例の破損があったため、登録されていた 60 例中 59 例を使用した。画像前処理と ROI 抽出は

施設内データセットと同様の方法論に従い、平均 $3562.9 \pm 886.4 \times 2001.4 \pm 494.4$ Pixel（平均 $\pm$ SD）の ROI 画像から症例あたり 24.2 ± 14.6 枚（平均 $\pm$ SD）のパッチを取得した。

5. 評価指標

症例分類では、各症例のパッチに対して多数決投票を行い、最も頻繁に予測されたクラスを選択した。パッチごとおよびマクロ平均指標を以下の式で算出した：

$$\text{Patch Accuracy} = \frac{\text{Correctly classified patches}}{\text{Total patches}}$$

$$\text{Overall Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Macro Metric} = \frac{1}{K} \sum \text{Metric}_i$$

ここで、 K はクラス数、 i は各クラスを表す。 Metric_i には以下のいずれかの性能指標が代入される。つまり、マクロ指標は各クラス（G、A、O などの分類クラス）ごとの性能指標を平均したものを表す。性能指標には感度、特異度、F1 スコアが含まれ、以下のように計算される。

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

ここで、TP（真陽性）、TN（真陰性）、FP（偽陽性）、FN（偽陰性）は各クラスの正誤分類を表す。

Overall Accuracy は全症例の正解率を求めるため、施設内データセットでは各

Fold ごとに全検証用 Fold の結果を統合して算出した。EBRAINS データセットでは、各 Fold のモデルの softmax 出力を平均化したアンサンブル予測を用いてを算出した。データセットと診断間の不均衡な分布を考慮し、マクロ平均 Recall を重視しつつ、包括的評価のためパッチごとの Accuracy と Overall Accuracy も考慮した。モデル間の統計的有意性は、5 分割交差検証の検証精度に対する対応のある t 検定で評価した ($p < 0.05$)。95%信頼区間は平均 $\pm 1.96 \times \text{SE}$ として算出した (SE はフォールド間の平均の標準誤差)。

6. 推論過程の解析

2 つの相補的な可視化技術を使用した。第一に、GradCAM(Selvaraju et al., 2017)を用いて ViT ベースモデルと CNN モデルの両方の最終エンコーダー層を可視化し、分類決定に最も重要な寄与をする領域を検証した。第二に、UMAP(McInnes et al., 2018)を用いてエンコーダから得られた特徴量を散布図によって可視化することで、高次元空間における特徴の分布を解析した。これらの特徴量は、ViT モデルは最終層の出力のクラストークンを、および CNN モデルはエンコーダーの出力のプーリング後の特徴量をそれぞれ抽出することで獲得した。

7. データおよびコードの利用可能性

本研究の実装コードおよび解析スクリプトは、<https://github.com/endaaman/bt>にて公開されている。病理画像データセット (HUPatho2BrainTumor) は、「個人情報保護に関する法律」および「人を対象とする生命科学・医学系研究に関する倫理指針」に従い、すべての個人識別情報を除去した上で、Zenodo(Enda et al., 2025)にて利用可能である。

8. 倫理承認

本後ろ向き研究は、北海道大学病院倫理委員会の承認を受けて実施された (承認番号 : 024-0476)。また、書面によるインフォームドコンセントの取得については免除された。すべての研究は、「人を対象とする生命科学・医学系研究に関する倫理指針」をはじめとする関連する指針および規程に従って実施された。

結果

1. ローカルデータセットにおける結果

5 種類の腫瘍タイプ (Glioblastoma (G)、Astrocytoma (A)、Oligodendroglioma (O)、PCNSL (L)、転移性脳腫瘍 (M)) および背景組織 (B) の 6 分類タスクで性能を評価した。施設内データセットで 500 パッチ/症例で訓練した場合、細粒度分類と粗粒度分類での各 fold のマクロ平均 Recall とパッチごとの正解率を図 9 および表 4 に示す。病理基盤モデルである UNI(LP)、UNI(FT)、Prov-GigaPath(LP)は全指標で堅牢な性能を示し、互いに有意な差はなかった ($p>0.1$)。定量的には、これらのモデルは細粒度分類で同様に高いマクロ平均 Recall (Prov-GigaPath(LP): 0.88 ± 0.10 、UNI(FT): 0.87 ± 0.07 、UNI(LP): 0.88 ± 0.09) とパッチ精度

(Prov-GigaPath(LP): 0.93 ± 0.02 、UNI(FT): 0.94 ± 0.01 、UNI(LP): 0.93 ± 0.03) を達成した。

Overall Accuracy に関しては、粗粒度分類では Prov-GigaPath(LP)が最高性能を達成し (マクロ平均 Recall: 0.99 ± 0.01 、パッチ精度: 0.97 ± 0.01)、254 例中 251 例を正しく分類した。細粒度分類では Prov-GigaPath(LP)と UNI(LP)の両方が最良の性能を示し、それぞれ 254 例中 234 例を正しく分類した。

注目すべきは、LP 条件のうち UNI(LP)と Prov-GigaPath(LP)が、従来の FT モデル (ViT-L_{IN}(FT)および CTransPath(FT)) をパッチごとの精度で有意に上回った (どちらも $p<0.01$) 点である。一方、同一モデルアーキテクチャ内での訓練条件の比較では、ViT-L_{IN}と CTransPath の両方で FT 条件が LP 条件よりも高いパッチ精度を示した (どちらも $p<0.01$)。なお、ResNet-RS 50_{IN}(LP)は一貫して低い性能を示したため、グラフから除外した。

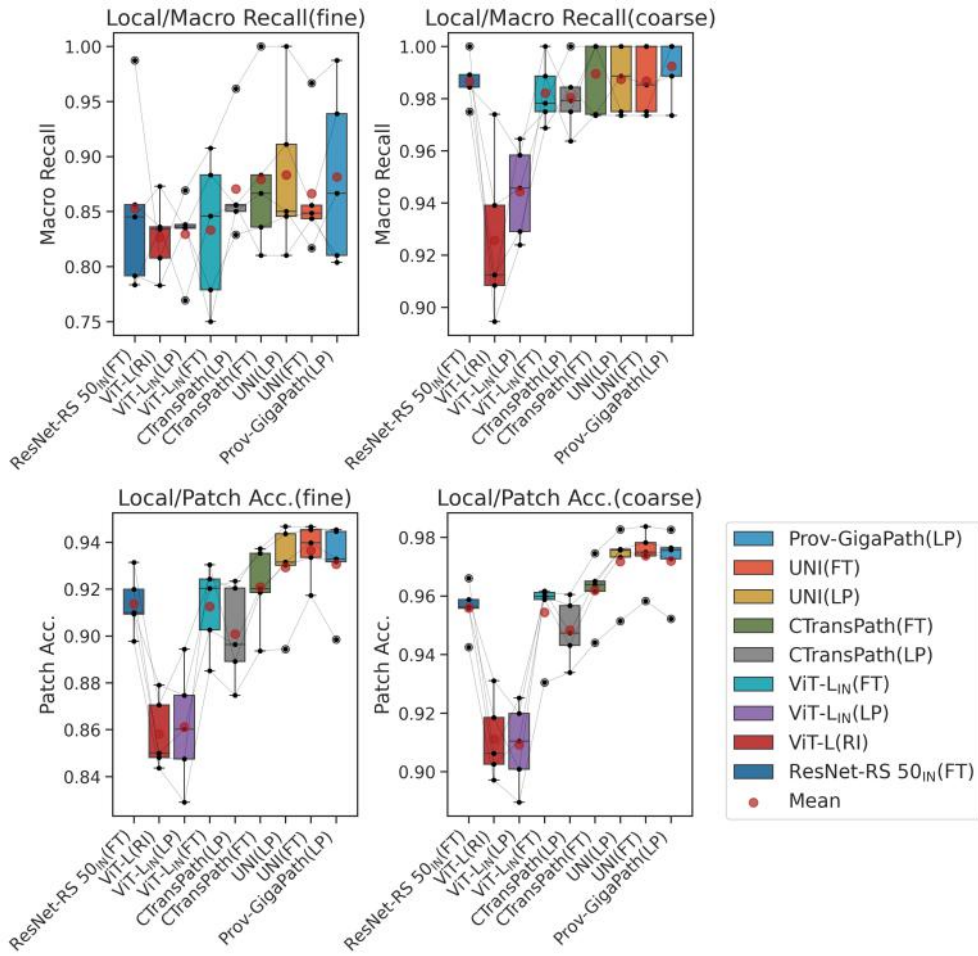


図 9 施設内データセットの性能指標

500 パッチ/症例で訓練されたモデルを使用し、施設内データセットの各 Fold でのバリデーションでの評価指標とその平均を示している。上段はマクロ平均 Recall、下段はパッチ正解率を示し、また左列は再分類、右列は粗分類の性能指標を示す。赤点が平均値、黒点は各 Fold の評価指標が対応し、同一 Fold を補助線で結んでいる。

	Fine grained			Coarse Grained		
	Macro Recall	Patch Acc.	Overall Acc.	Macro Recall	Patch Acc.	Overall Acc.
Prov-GigaPath(LP)	0.88 ± 0.10	0.93 ± 0.02	234/254	0.99 ± 0.01	0.97 ± 0.01	251/254
UNI(FT)	0.87 ± 0.07	0.94 ± 0.01	231/254	0.99 ± 0.02	0.97 ± 0.01	250/254
UNI(LP)	0.88 ± 0.09	0.93 ± 0.03	234/254	0.99 ± 0.02	0.97 ± 0.01	250/254
CTransPath(FT)	0.88 ± 0.09	0.92 ± 0.02	233/254	0.99 ± 0.02	0.96 ± 0.01	250/254
CTransPath(LP)	0.87 ± 0.06	0.90 ± 0.03	230/254	0.98 ± 0.02	0.95 ± 0.01	248/254
VIT-L _{IN} (FT)	0.83 ± 0.08	0.91 ± 0.02	225/254	0.98 ± 0.02	0.95 ± 0.02	248/254
VIT-L _{IN} (LP)	0.83 ± 0.05	0.86 ± 0.03	217/254	0.94 ± 0.02	0.91 ± 0.02	235/254
VIT-L(RI)	0.83 ± 0.04	0.86 ± 0.02	215/254	0.93 ± 0.04	0.91 ± 0.02	233/254
ResNet-RS 50 _{IN} (FT)	0.85 ± 0.10	0.91 ± 0.02	229/254	0.99 ± 0.01	0.96 ± 0.01	250/254
ResNet-RS 50 _{IN} (LP)	0.36 ± 0.07	0.38 ± 0.09	57/254	0.50 ± 0.09	0.52 ± 0.06	126/254

表 4 施設内データセットの 500 パッチ/症例で訓練時の性能指標

UNI(LP)と UNI(FT)の混同行列（図 10）では、A クラスを G クラスまたは O クラスと誤分類する症例が多く見られた。施設内データセットの混同行列は、LP と FT の両条件で高い分類精度と類似したパターンを示し、M クラスと L クラスの正確な分類が観察され、誤分類は主に Glioma クラス（G、A、O クラス）内で発生していた。

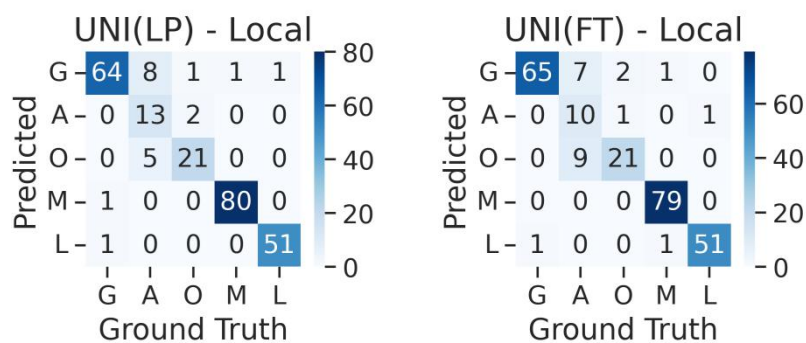


図 10 施設内データセットでの UNI(LP)および UNI(FT)の混同行列

UNI の FT と LP について、縦軸は予測クラス、横軸は正解クラスとしてそれぞれの対応する症例数を行列で示した。

2. 外部データセットでの結果

施設内データセットで 500 パッチ/症例を使用して各 Fold を訓練し、それぞれ EBRAINS データセットで評価したマクロ平均 Recall とパッチ精度を図 11 と表 5 に示す。UNI(LP)は全指標で優れた性能を示し（細粒度: マクロ平均 Recall 0.75 ± 0.01 、パッチ精度 0.65 ± 0.02 ；粗粒度: マクロ平均 Recall 0.94 ± 0.01 、パッチ精度 0.90 ± 0.02 ）、細粒度で 698 例中 479 例、粗粒度で 698 例中 648 例を正しく分類した。細粒度において、UNI(LP)は Prov-GigaPath(LP)（マクロ平均 Recall: 0.73 ± 0.01 、パッチ精度: 0.64 ± 0.01 ）をマクロ平均 Recall で有意に上回り ($p=0.020$)、パッチ精度でも高い傾向を示した ($p=0.053$)。

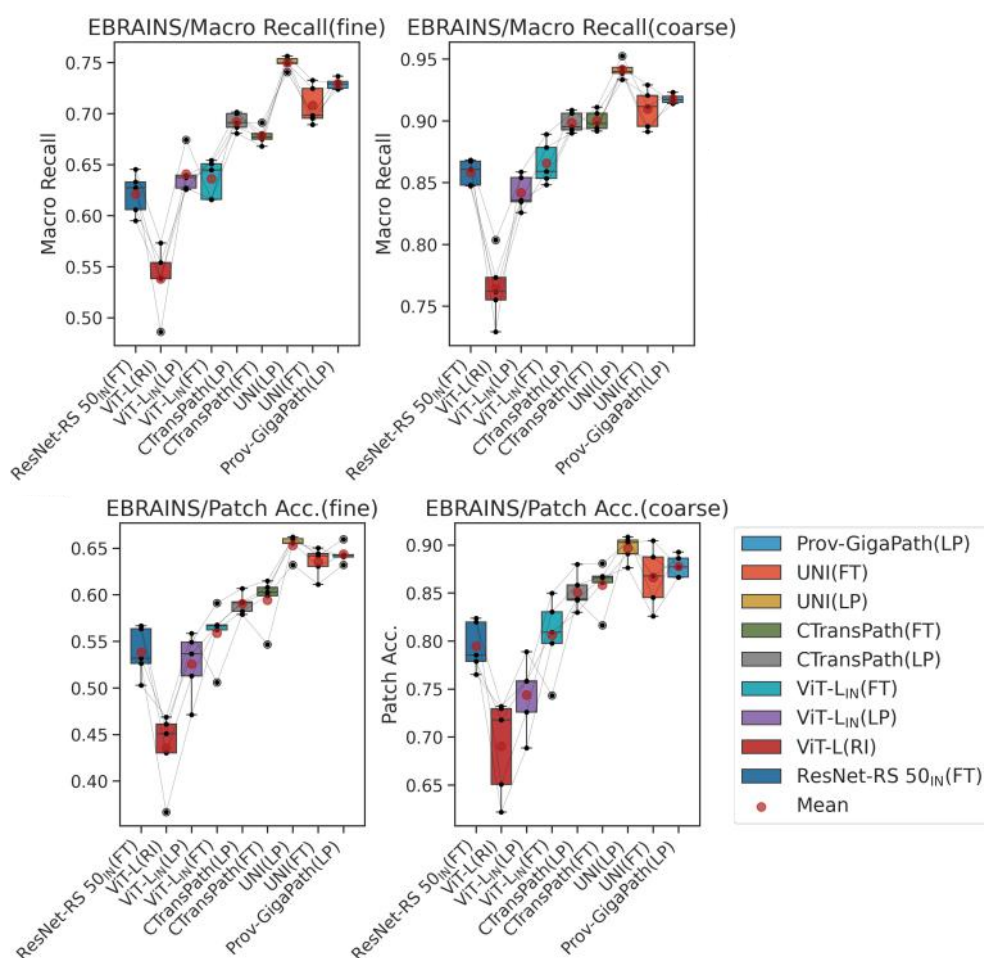


図 11 EBRAINS データセットの 500 パッチ/症例で訓練時の性能指標

施設内データセットで 500 パッチ/症例を使用し、各 Fold で EBRAINS データセットでのマクロ平均 Recall およびパッチ精度を示す。上段はマクロ平均 Recall、下段はパッチ正解率を示し、また左列は再分類、右列は粗分類の性能指標を示す。赤点が平均値、黒点は各 Fold の評価指標が対応し、同一 Fold を補助線で結んでいる。

	Fine grained			Coarse Grained		
	Macro Recall	Patch Acc.	Overall Acc.	Macro Recall	Patch Acc.	Overall Acc.
Prov-GigaPath(LP)	0.73 ± 0.01	0.64 ± 0.01	474/698	0.92 ± 0.00	0.88 ± 0.01	635/698
UNI(FT)	0.71 ± 0.02	0.64 ± 0.02	470/698	0.91 ± 0.02	0.87 ± 0.04	625/698
UNI(LP)	0.75 ± 0.01	0.65 ± 0.02	479/698	0.94 ± 0.01	0.90 ± 0.02	648/698
CTransPath(FT)	0.68 ± 0.01	0.59 ± 0.03	437/698	0.90 ± 0.01	0.86 ± 0.03	626/698
CTransPath(LP)	0.69 ± 0.01	0.59 ± 0.01	446/698	0.90 ± 0.01	0.85 ± 0.02	632/698
ViT-L _{IN} (FT)	0.64 ± 0.02	0.56 ± 0.04	422/698	0.87 ± 0.02	0.81 ± 0.05	602/698
ViT-L _{IN} (LP)	0.64 ± 0.02	0.53 ± 0.04	407/698	0.84 ± 0.02	0.74 ± 0.05	561/698
ViT-L(RI)	0.54 ± 0.04	0.44 ± 0.05	336/698	0.76 ± 0.03	0.69 ± 0.06	537/698
ResNet-RS 50 _{IN} (FT)	0.62 ± 0.03	0.54 ± 0.03	416/698	0.86 ± 0.01	0.79 ± 0.03	605/698
ResNet-RS 50 _{IN} (LP)	0.29 ± 0.04	0.13 ± 0.04	189/698	0.45 ± 0.04	0.46 ± 0.16	549/698

表 5 EBRAINS データセットの 500 パッチ/症例で訓練時の性能指標

同一モデル内での訓練条件の比較では、UNI は LP 条件が FT 条件をマクロ平均 Recall ($p < 0.01$) とパッチ精度 ($p = 0.011$) の両方で上回った。ViT-L_{IN} では、LP がマクロ平均 Recall で高く、FT がパッチ精度で高い傾向を示したが、各 Fold ごと一貫した傾向は見られず、有意差は検出されなかった（それぞれ $p = 0.78$ 、 $p = 0.24$ ）。CTransPath では、LP が FT よりも高いマクロ平均 Recall を達成し ($p = 0.037$)、FT が LP よりも高いパッチ精度を示した ($p = 0.039$)。

EBRAINS の混同行列 (図 12) は、モデル予測 (G、A、O、M、L) と分子サブタイプの真値を対比を行った。誤分類は、施設内データセット同様主に Glioma クラス (G、A、O クラス) 内で発生していた。G クラスでは、FT が LP よりもわずかに高い Recall を示した (0.76 ± 0.03 vs 0.73 ± 0.07 , $p = 0.431$)。これは特に再分類症例で顕著で、FT は Anaplastic astrocytoma, IDH-wildtype (AA, IDH-) を膠芽腫として正しく識別した症例が多く (47 例中 30 例 vs 47 例中 24 例)、Diffuse astrocytoma, IDH-wildtype (DA, IDH-) でも同様だった (19 例中 8 例 vs 19 例中 5 例)。一方、LP はこれらを星細胞腫と誤分類する傾向があった (AA: 47 例中 6 例 vs 47 例中 19 例 ; DA: 19 例中 4 例 vs 19 例中 14 例)。A クラスでは、クラス内での正解率は LP が FT を有意に上回った (0.59 ± 0.06 vs 0.34 ± 0.11 , $p < 0.001$)。この差は DA, IDH+ で最も顕著で、LP は 70 例中 58 例を星細胞腫として正しく分類したのに対し、FT は星細胞腫 (70 例中 34 例) と乏突起膠腫 (70 例中 24 例) に分けて分類した。O クラスでは、FT が LP よりも高い Recall を示した (0.59 ± 0.13 vs 0.52 ± 0.05 , $p = 0.167$)。FT は特に Oligodendroglioma (非 Anaplastic) 症例でより良い識別を示し (85 例中 65 例正解)、LP は A クラスとの誤分類が

多かった（O: 85 例中 28 例、AO: 91 例中 24 例）。L クラスでは、両条件で同様の Recall を示したが（各 57 例正解）、LP はより選択的な予測を示した（予測症例数：83 例中 57 例 vs 102 例中 57 例）。

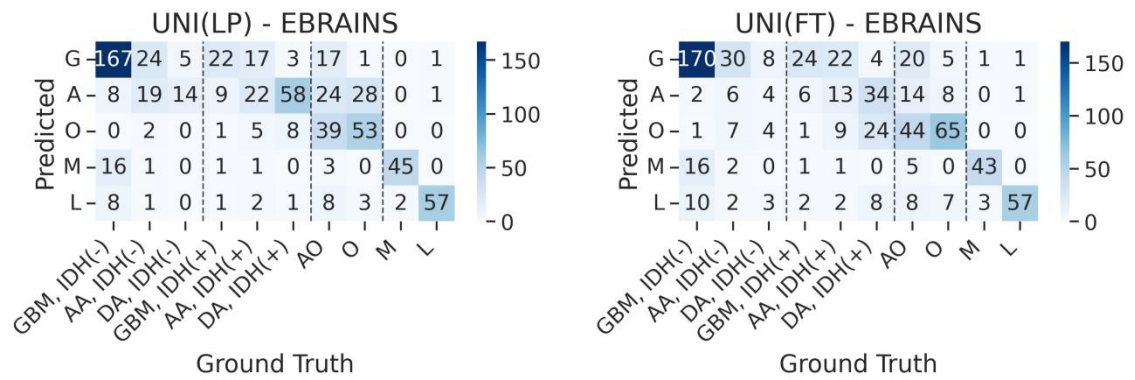


図 12 EBRAINS データセットでの UNI(LP)および UNI(FT)の混同行列

縦軸は予測クラス、横軸は正解クラスを示し、それぞれのクラスの対応する数を行列で示した。EBRAINS には WHO 4th 基準で登録されているので、元の登録名称を示している。略語はそれぞれ、GBM: Glioblastoma、AA: Anaplastic astrocytoma、DA: Diffuse astrocytoma、AO: Anaplastic oligodendroglioma、IDH(-): IDH-wild type、IDH(+): IDH-mutant を示す。

3. 訓練に使用するパッチ枚数を変化させたときの結果

症例あたりの訓練パッチ数を変化させ（500、100、25、10 枚）、施設内データセットと EBRAINS データセットの両方でモデル性能を評価した（100、25、10 パッチの詳細結果を表 5、6、7 に示す）。施設内データセットの検証 Fold では、全モデルでパッチ数の増加に伴いマクロ平均 Recall と正解パッチ数が単調に改善した。クラス別精度分析では、G、A、M、L クラスで増加傾向が見られ、A クラスは不安定ながらも全体的な改善を示した（図 13）。

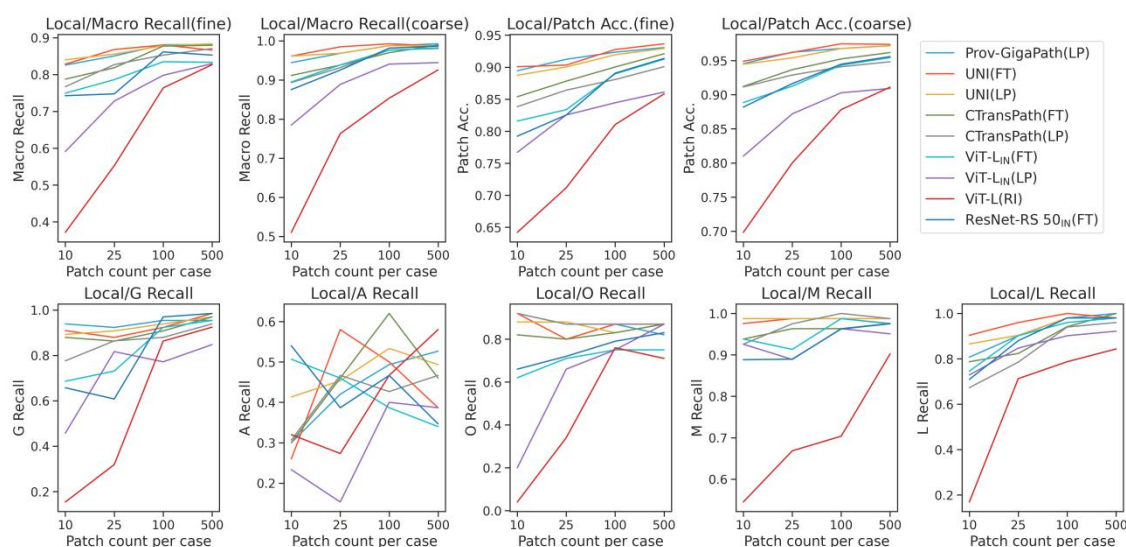


図 13 施設内データセットのパッチ数ごとの性能指標の変化

上段、左から細粒度分類のマクロ平均 Recall、粗粒度分類のマクロ平均 Recall、細粒度分類のパッチ毎精度、細粒度分類のパッチ毎精度を示す。下段は各クラスごとの Recall を示し、左から G、A、O、M、L クラスの性能を示す。

EBRAINS データセットでは、マクロ平均 Recall は概ね同様の単調な改善を示したが（図 14）、UNI(FT)と Prov-GigaPath(LP)では 100 パッチ/症例と 500 パッチ/症例で性能低下が観察された。逆に、全体精度とパッチごとの精度は、UNI モデルを含む複数の条件でパッチ数の増加に伴う減少をわずかに示した。クラス別 Recall では、G、M、L クラスは全条件でパッチ数増加に伴い一貫した改善を示した。

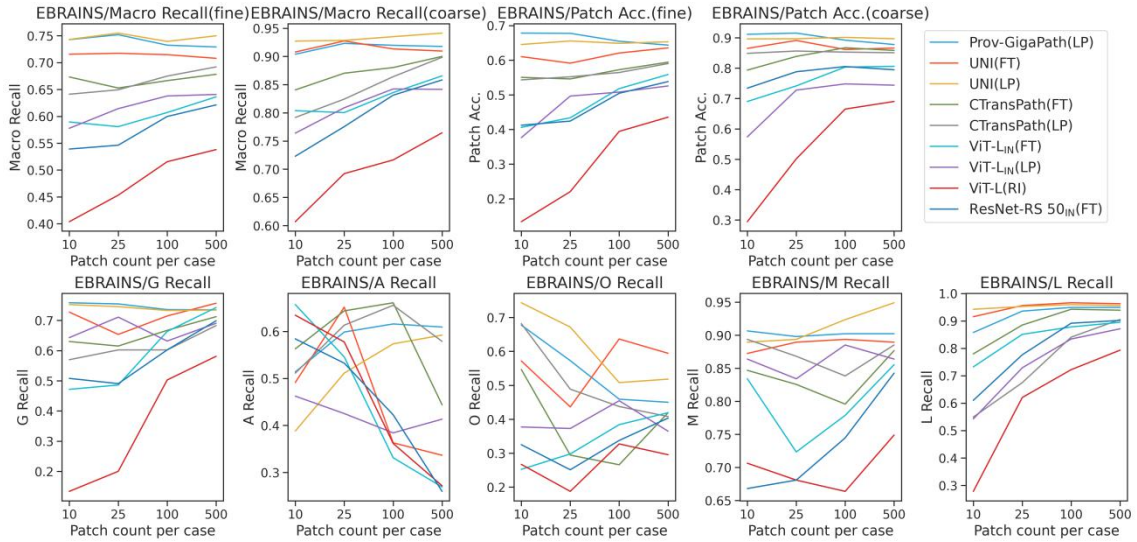


図 14 EBRAINS データセットのパッチ数ごとの性能指標の変化

上段は左から

10 パッチ/症例ベースラインからの累積性能差の定量分析 (図 15) では、施設内データセットでは、病理基盤モデルは控えめな向上を示した (10→500 パッチで Prov-GigaPath(LP): マクロ平均 Recall +0.056、UNI(LP): +0.044) のに対し、病理特化の事前学習なしのモデルは大幅な改善を示した (ViT-L(RI): +0.455、ViT-L_{IN}(LP): +0.239)。EBRAINS データセットでは、病理基盤モデルは最小限の改善または性能低下を示した (Prov-GigaPath(LP): マクロ平均 Recall -0.014、パッチ精度 -0.035 ; UNI(FT): マクロ平均 Recall -0.008) のに対し、従来モデルはパッチ数増加に伴って性能向上が見られた (ViT-L(RI): マクロ平均 Recall +0.134、パッチ精度 +0.302)。

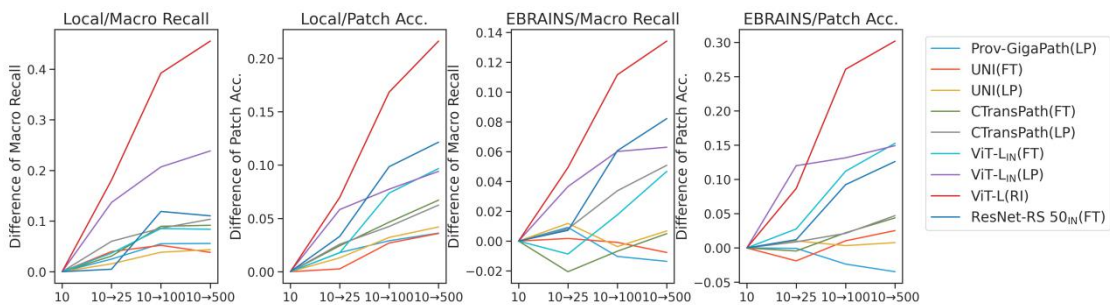


図 15 パッチ数を変化させた際の性能差分のプロット

10 パッチ/症例の性能指標を基準として、枚数を増やすごとの変化量をプロットした。

10 パッチ/症例の UNI(LP)は、EBRAINS データセットで優れた性能を示し、500

パッチ/症例の ViT-L_{IN}(FT)と CTransPath(FT)の両方を、細粒度分類 (0.743±0.008 vs 0.636±0.024 vs 0.678±0.011、p<0.001 および p<0.001) と粗粒度分類 (0.927±0.006 vs 0.866±0.022 vs 0.880±0.015、p<0.001 および p<0.001) で有意に上回った。

Local Dataset	Fine grained				Coarse Grained			
	Macro	Patch Acc.	Macro Acc.	Overall	Macro	Patch Acc.	Macro Acc.	Overall
	Recall				Recall			
Prov-GigaPath(LP)	0.88 ± 0.09	0.92 ± 0.03	0.95 ± 0.04	233/254	0.99 ± 0.02	0.97 ± 0.02	0.99 ± 0.01	250/254
UNI(FT)	0.88 ± 0.06	0.93 ± 0.02	0.94 ± 0.02	232/254	0.99 ± 0.01	0.97 ± 0.01	0.99 ± 0.01	251/254
UNI(LP)	0.88 ± 0.08	0.92 ± 0.03	0.94 ± 0.03	232/254	0.99 ± 0.02	0.97 ± 0.02	0.99 ± 0.01	250/254
CTransPath(FT)	0.88 ± 0.10	0.90 ± 0.03	0.93 ± 0.04	228/254	0.97 ± 0.02	0.95 ± 0.02	0.98 ± 0.01	245/254
CTransPath(LP)	0.85 ± 0.07	0.88 ± 0.03	0.93 ± 0.03	225/254	0.98 ± 0.03	0.94 ± 0.02	0.98 ± 0.02	247/254
ViT-L _{IN} (FT)	0.83 ± 0.04	0.89 ± 0.03	0.92 ± 0.03	224/254	0.98 ± 0.02	0.94 ± 0.01	0.98 ± 0.01	246/254
ViT-L _{IN} (LP)	0.80 ± 0.03	0.84 ± 0.04	0.89 ± 0.03	209/254	0.94 ± 0.03	0.90 ± 0.02	0.95 ± 0.02	234/254
ViT-L(RI)	0.76 ± 0.06	0.81 ± 0.03	0.84 ± 0.05	190/254	0.85 ± 0.07	0.88 ± 0.02	0.89 ± 0.05	210/254
ResNet-RS 50 _{IN} (FT)	0.86 ± 0.07	0.89 ± 0.02	0.94 ± 0.03	229/254	0.98 ± 0.02	0.94 ± 0.02	0.98 ± 0.02	248/254
ResNet-RS 50 _{IN} (LP)	0.40 ± 0.06	0.45 ± 0.06	0.48 ± 0.04	67/254	0.53 ± 0.04	0.51 ± 0.04	0.54 ± 0.05	92/254
EBRAINS Dataset	Macro	Patch Acc.	Macro Acc.	Overall	Macro	Patch Acc.	Macro Acc.	Overall
	Recall				Recall			
Prov-GigaPath(LP)	0.73 ± 0.01	0.65 ± 0.02	0.67 ± 0.02	478/698	0.92 ± 0.01	0.89 ± 0.02	0.91 ± 0.02	636/698
UNI(FT)	0.71 ± 0.01	0.62 ± 0.03	0.65 ± 0.02	475/698	0.91 ± 0.03	0.86 ± 0.04	0.89 ± 0.04	627/698
UNI(LP)	0.74 ± 0.01	0.65 ± 0.02	0.67 ± 0.02	479/698	0.94 ± 0.01	0.90 ± 0.02	0.93 ± 0.01	650/698
CTransPath(FT)	0.67 ± 0.02	0.57 ± 0.01	0.60 ± 0.02	415/698	0.88 ± 0.01	0.87 ± 0.02	0.90 ± 0.01	626/698
CTransPath(LP)	0.67 ± 0.01	0.56 ± 0.02	0.61 ± 0.02	431/698	0.86 ± 0.01	0.85 ± 0.02	0.90 ± 0.01	631/698
ViT-L _{IN} (FT)	0.61 ± 0.02	0.52 ± 0.05	0.55 ± 0.03	398/698	0.84 ± 0.02	0.80 ± 0.06	0.85 ± 0.04	594/698
ViT-L _{IN} (LP)	0.64 ± 0.03	0.51 ± 0.07	0.57 ± 0.06	415/698	0.84 ± 0.02	0.75 ± 0.06	0.82 ± 0.04	568/698
ViT-L(RI)	0.52 ± 0.02	0.39 ± 0.07	0.46 ± 0.07	341/698	0.72 ± 0.04	0.67 ± 0.06	0.75 ± 0.07	545/698
ResNet-RS 50 _{IN} (FT)	0.60 ± 0.04	0.50 ± 0.02	0.53 ± 0.04	392/698	0.83 ± 0.02	0.81 ± 0.03	0.85 ± 0.03	599/698
ResNet-RS 50 _{IN} (LP)	0.31 ± 0.03	0.06 ± 0.03	0.23 ± 0.03	173/698	0.44 ± 0.03	0.17 ± 0.09	0.45 ± 0.13	305/698

表 5 100 パッチ/症例での、施設内および EBRAINS データセットの性能指標

Local Dataset	Fine grained			Overall Acc.	Coarse Grained			Overall Acc.
	Macro	Patch Acc.	Macro Acc.		Macro	Patch Acc.	Macro Acc.	
	Recall				Recall			
Prov-GigaPath(LP)	0.85 ± 0.09	0.91 ± 0.03	0.93 ± 0.03	225/254	0.97 ± 0.03	0.96 ± 0.01	0.98 ± 0.02	246/254
UNI(FT)	0.87 ± 0.03	0.90 ± 0.04	0.93 ± 0.02	227/254	0.98 ± 0.03	0.96 ± 0.02	0.99 ± 0.02	250/254
UNI(LP)	0.86 ± 0.08	0.90 ± 0.03	0.93 ± 0.04	225/254	0.97 ± 0.04	0.95 ± 0.02	0.98 ± 0.02	246/254
CTransPath(FT)	0.82 ± 0.09	0.88 ± 0.04	0.90 ± 0.06	214/254	0.94 ± 0.05	0.94 ± 0.02	0.96 ± 0.03	238/254
CTransPath(LP)	0.83 ± 0.08	0.86 ± 0.03	0.90 ± 0.03	215/254	0.93 ± 0.04	0.93 ± 0.02	0.96 ± 0.01	237/254
VIT-L _{in} (FT)	0.79 ± 0.02	0.83 ± 0.04	0.87 ± 0.04	203/254	0.94 ± 0.02	0.91 ± 0.01	0.95 ± 0.01	235/254
VIT-L _{in} (LP)	0.73 ± 0.04	0.83 ± 0.01	0.85 ± 0.03	195/254	0.89 ± 0.01	0.87 ± 0.01	0.90 ± 0.02	216/254
VIT-L(RI)	0.55 ± 0.05	0.71 ± 0.02	0.69 ± 0.04	132/254	0.76 ± 0.07	0.80 ± 0.04	0.80 ± 0.07	174/254
ResNet-RS 50 _{in} (FT)	0.75 ± 0.07	0.83 ± 0.05	0.84 ± 0.06	190/254	0.93 ± 0.04	0.92 ± 0.01	0.94 ± 0.02	231/254
ResNet-RS 50 _{in} (LP)	0.40 ± 0.05	0.46 ± 0.06	0.48 ± 0.03	67/254	0.51 ± 0.04	0.51 ± 0.06	0.52 ± 0.04	80/254
EBRAINS Dataset	Macro	Patch Acc.	Macro Acc.	Overall	Macro	Patch Acc.	Macro Acc.	Overall
	Recall			Acc.	Recall			Acc.
Prov-GigaPath(LP)	0.75 ± 0.01	0.68 ± 0.01	0.70 ± 0.01	497/698	0.92 ± 0.01	0.92 ± 0.01	0.93 ± 0.00	654/698
UNI(FT)	0.72 ± 0.03	0.59 ± 0.05	0.64 ± 0.04	476/698	0.93 ± 0.01	0.89 ± 0.01	0.94 ± 0.02	656/698
UNI(LP)	0.75 ± 0.02	0.66 ± 0.01	0.70 ± 0.01	498/698	0.93 ± 0.01	0.90 ± 0.01	0.94 ± 0.01	657/698
CTransPath(FT)	0.65 ± 0.02	0.55 ± 0.03	0.58 ± 0.02	417/698	0.87 ± 0.03	0.84 ± 0.01	0.89 ± 0.02	629/698
CTransPath(LP)	0.65 ± 0.02	0.55 ± 0.02	0.60 ± 0.01	425/698	0.82 ± 0.03	0.86 ± 0.01	0.90 ± 0.01	631/698
VIT-L _{in} (FT)	0.58 ± 0.02	0.43 ± 0.07	0.50 ± 0.05	369/698	0.80 ± 0.03	0.74 ± 0.10	0.82 ± 0.07	589/698
VIT-L _{in} (LP)	0.61 ± 0.05	0.50 ± 0.05	0.57 ± 0.05	426/698	0.81 ± 0.05	0.73 ± 0.03	0.85 ± 0.04	602/698
VIT-L(RI)	0.45 ± 0.03	0.22 ± 0.03	0.35 ± 0.02	247/698	0.69 ± 0.05	0.50 ± 0.13	0.76 ± 0.04	533/698
ResNet-RS 50 _{in} (FT)	0.55 ± 0.02	0.42 ± 0.06	0.48 ± 0.05	356/698	0.78 ± 0.01	0.79 ± 0.03	0.85 ± 0.03	610/698
ResNet-RS 50 _{in} (LP)	0.29 ± 0.02	0.06 ± 0.02	0.20 ± 0.02	148/698	0.41 ± 0.01	0.14 ± 0.04	0.31 ± 0.03	218/698

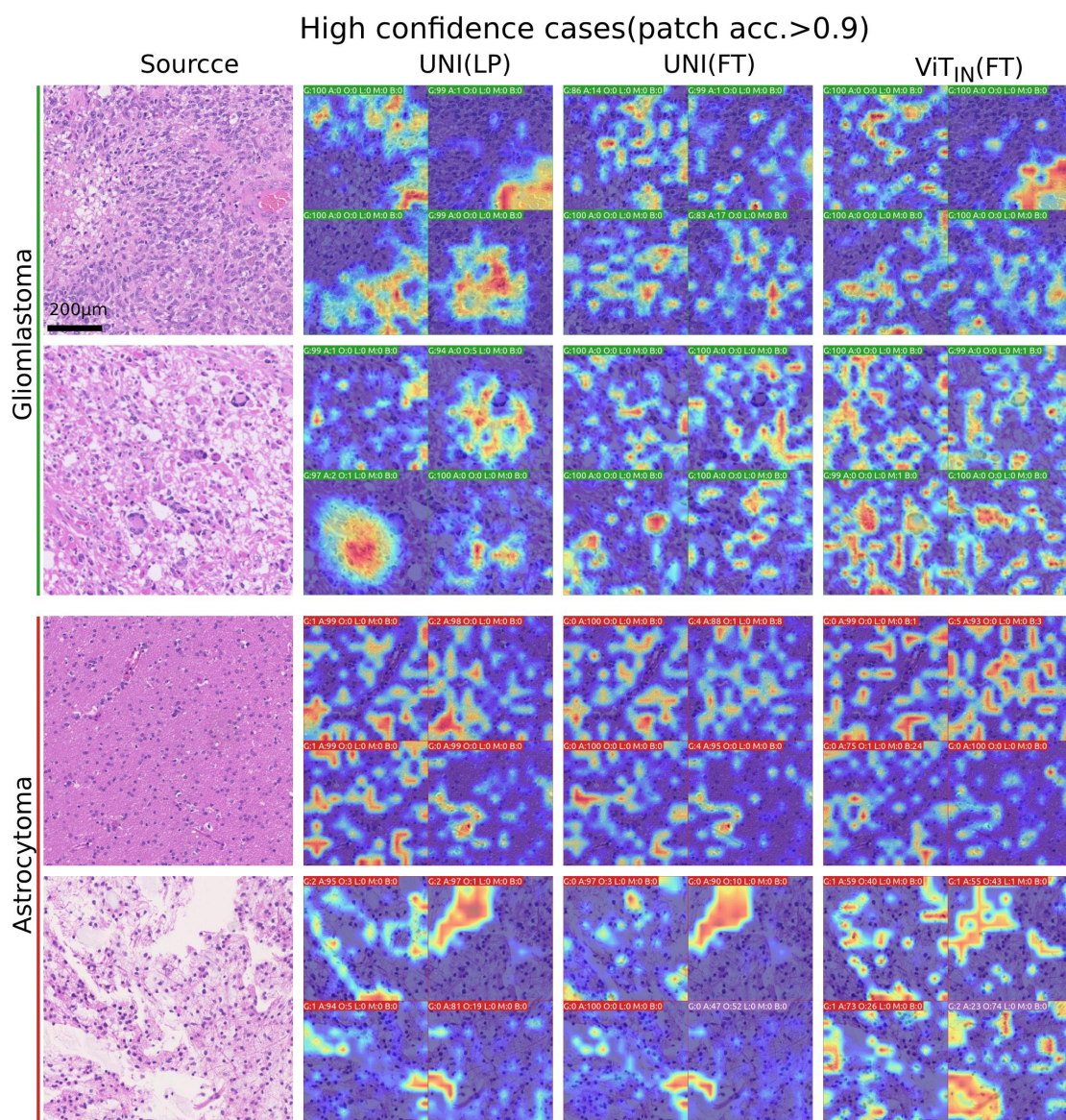
表 6 25 パッチ/症例での、施設内および EBRAINS データセットの性能指標

Local Dataset	Fine grained			Overall Acc.	Coarse Grained			Overall Acc.
	Macro	Patch Acc.	Macro Acc.		Macro	Patch Acc.	Macro Acc.	
	Recall				Recall			
Prov-GigaPath(LP)	0.83 ± 0.07	0.89 ± 0.03	0.91 ± 0.03	225/254	0.94 ± 0.05	0.95 ± 0.02	0.97 ± 0.03	241/254
UNI(FT)	0.83 ± 0.07	0.90 ± 0.03	0.91 ± 0.03	227/254	0.96 ± 0.03	0.95 ± 0.02	0.97 ± 0.02	243/254
UNI(LP)	0.84 ± 0.08	0.89 ± 0.03	0.92 ± 0.04	225/254	0.96 ± 0.04	0.94 ± 0.02	0.98 ± 0.02	245/254
CTransPath(FT)	0.79 ± 0.08	0.85 ± 0.03	0.88 ± 0.04	214/254	0.91 ± 0.05	0.91 ± 0.02	0.94 ± 0.04	229/254
CTransPath(LP)	0.77 ± 0.08	0.84 ± 0.03	0.85 ± 0.06	215/254	0.89 ± 0.05	0.91 ± 0.02	0.93 ± 0.03	228/254
VIT-L _{in} (FT)	0.75 ± 0.09	0.82 ± 0.04	0.85 ± 0.06	203/254	0.89 ± 0.05	0.89 ± 0.02	0.92 ± 0.03	224/254
VIT-L _{in} (LP)	0.59 ± 0.09	0.77 ± 0.04	0.76 ± 0.07	195/254	0.78 ± 0.08	0.81 ± 0.04	0.80 ± 0.08	175/254
VIT-L(RI)	0.37 ± 0.07	0.64 ± 0.04	0.55 ± 0.08	132/254	0.51 ± 0.10	0.70 ± 0.04	0.60 ± 0.07	96/254
ResNet-RS 50 _{in} (FT)	0.74 ± 0.07	0.79 ± 0.04	0.83 ± 0.06	190/254	0.88 ± 0.06	0.88 ± 0.02	0.91 ± 0.03	219/254
ResNet-RS 50 _{in} (LP)	0.35 ± 0.07	0.48 ± 0.03	0.48 ± 0.05	67/254	0.49 ± 0.04	0.51 ± 0.03	0.49 ± 0.04	65/254
EBRAINS Dataset	Macro	Patch Acc.	Macro Acc.	Overall	Macro	Patch Acc.	Macro Acc.	Overall
	Recall			Acc.	Recall			Acc.
Prov-GigaPath(LP)	0.74 ± 0.01	0.68 ± 0.02	0.70 ± 0.01	498/698	0.90 ± 0.01	0.91 ± 0.01	0.94 ± 0.01	656/698
UNI(FT)	0.72 ± 0.01	0.61 ± 0.05	0.66 ± 0.03	492/698	0.91 ± 0.02	0.87 ± 0.05	0.93 ± 0.03	656/698
UNI(LP)	0.74 ± 0.01	0.65 ± 0.02	0.70 ± 0.01	493/698	0.93 ± 0.01	0.90 ± 0.01	0.95 ± 0.00	660/698
CTransPath(FT)	0.67 ± 0.03	0.55 ± 0.02	0.62 ± 0.02	447/698	0.84 ± 0.02	0.79 ± 0.04	0.88 ± 0.02	619/698
CTransPath(LP)	0.64 ± 0.02	0.54 ± 0.04	0.61 ± 0.02	427/698	0.79 ± 0.03	0.85 ± 0.01	0.90 ± 0.01	626/698
VIT-L _{in} (FT)	0.59 ± 0.04	0.41 ± 0.04	0.50 ± 0.05	373/698	0.80 ± 0.02	0.69 ± 0.10	0.83 ± 0.07	595/698
VIT-L _{in} (LP)	0.58 ± 0.02	0.38 ± 0.07	0.54 ± 0.05	403/698	0.76 ± 0.02	0.57 ± 0.13	0.86 ± 0.04	602/698
VIT-L(RI)	0.40 ± 0.03	0.13 ± 0.01	0.33 ± 0.03	229/698	0.61 ± 0.07	0.29 ± 0.04	0.78 ± 0.08	558/698
ResNet-RS 50 _{in} (FT)	0.54 ± 0.04	0.41 ± 0.10	0.50 ± 0.07	375/698	0.72 ± 0.03	0.73 ± 0.10	0.85 ± 0.03	617/698
ResNet-RS 50 _{in} (LP)	0.28 ± 0.03	0.06 ± 0.01	0.18 ± 0.03	133/698	0.40 ± 0.03	0.11 ± 0.05	0.29 ± 0.10	191/698

表 7 10 パッチ/症例での、施設内および EBRAINS データセットの性能指標

4. GradCAM パターンの解析

症例単位でパッチ精度が高い (>0.9) および低い (<0.5) 症例を選択して GradCAM パターンを分析した (図 16、17、18、19)。G クラスでは、強い GradCAM 信号が異型核と微小血管増殖を強調し、壊死領域は限定的な応答を示した。L クラスでは、重要な診断的特徴である血管周囲浸潤パターンで明確な信号が観察された。A クラスおよび O クラスでは、腫瘍領域全体により広範な GradCAM 信号分布を示し、M クラスのような充実性腫瘍でのより局在化した信号と対照的だった。これらのパターンは、モデルが分類のために組織学的に関連する特徴を学習したことを示唆している。



High confidence cases(patch acc.>0.9)

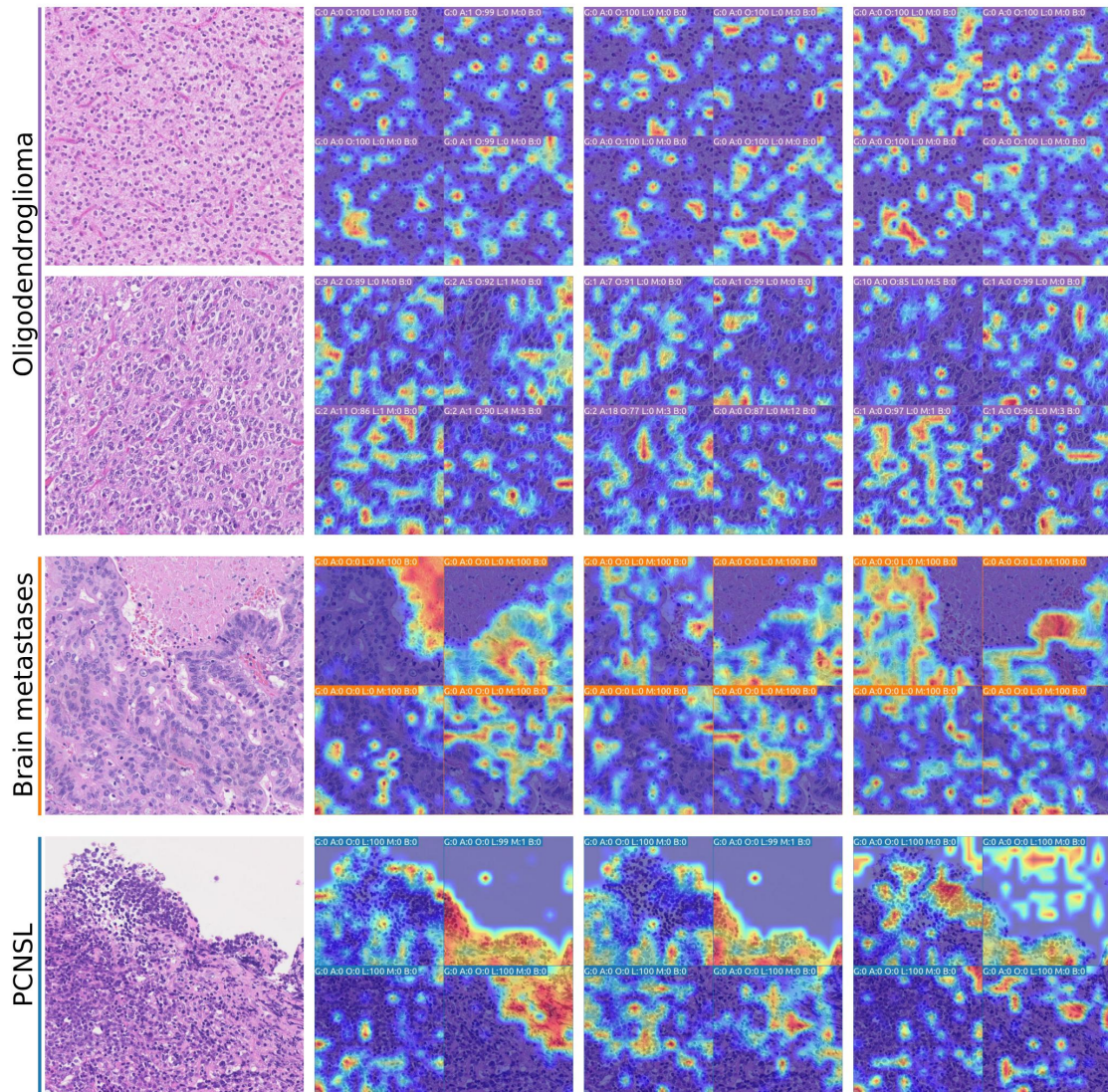
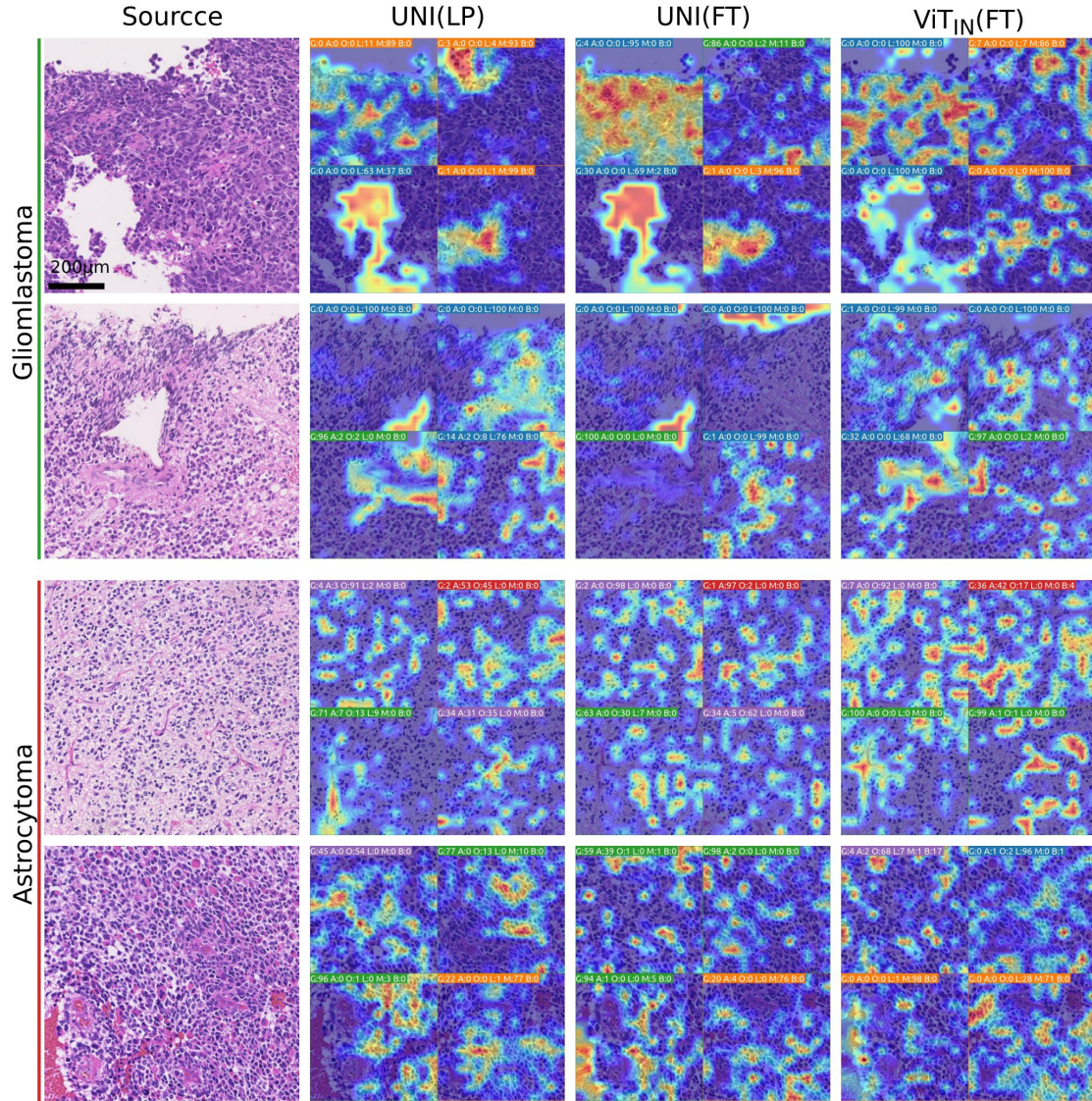


図 16 施設内データセットの高精度症例の GradCAM 分析

施設内データセットで、1 症例に含まれるパッチごとの正解率が 90%を超える症例での、パッチごとの予測クラスと GradCAM 出力結果を、モデルごとに出力した。

Low confidence cases(patch acc.<0.5)



Low confidence cases(patch acc.<0.5)

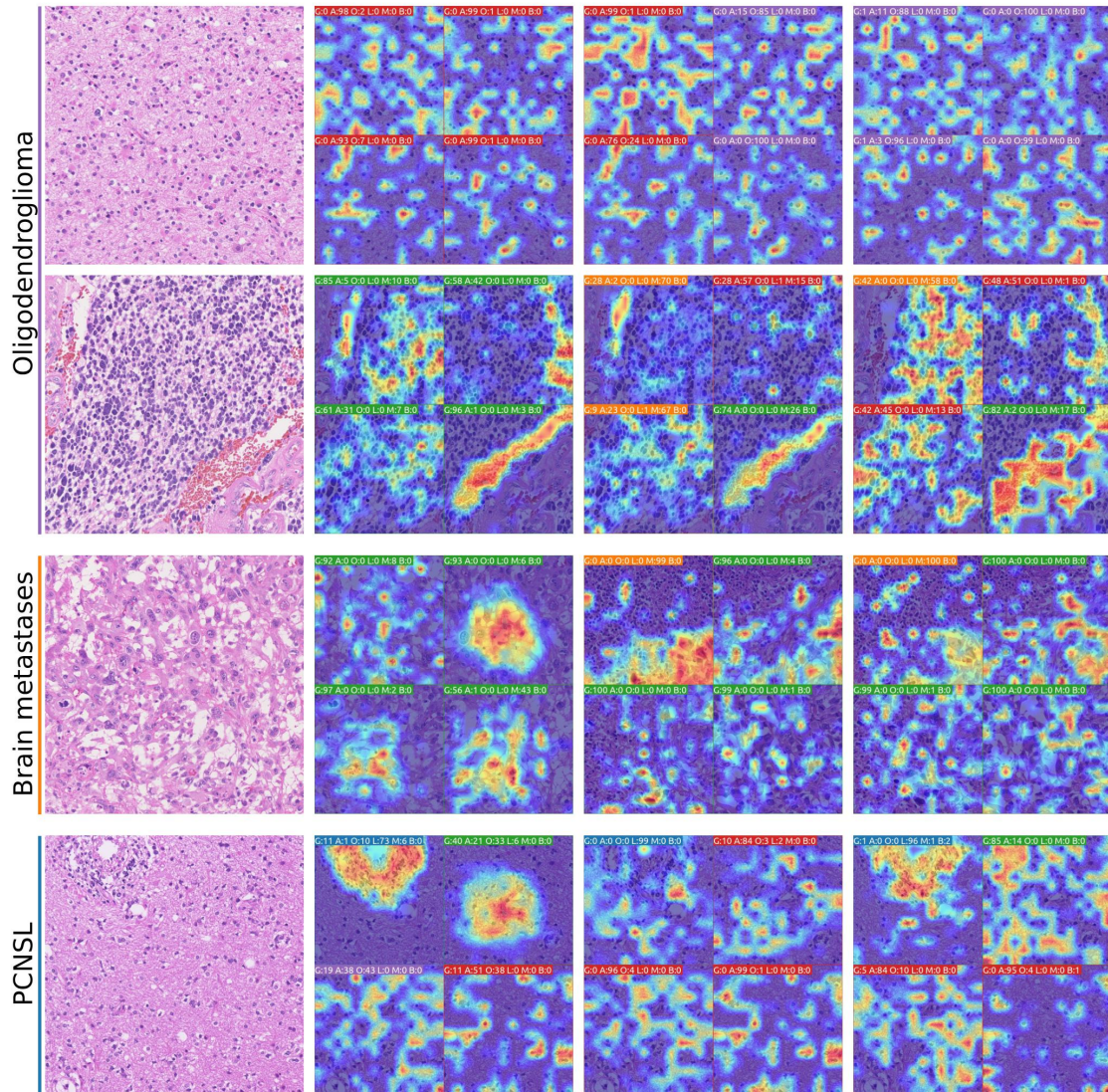
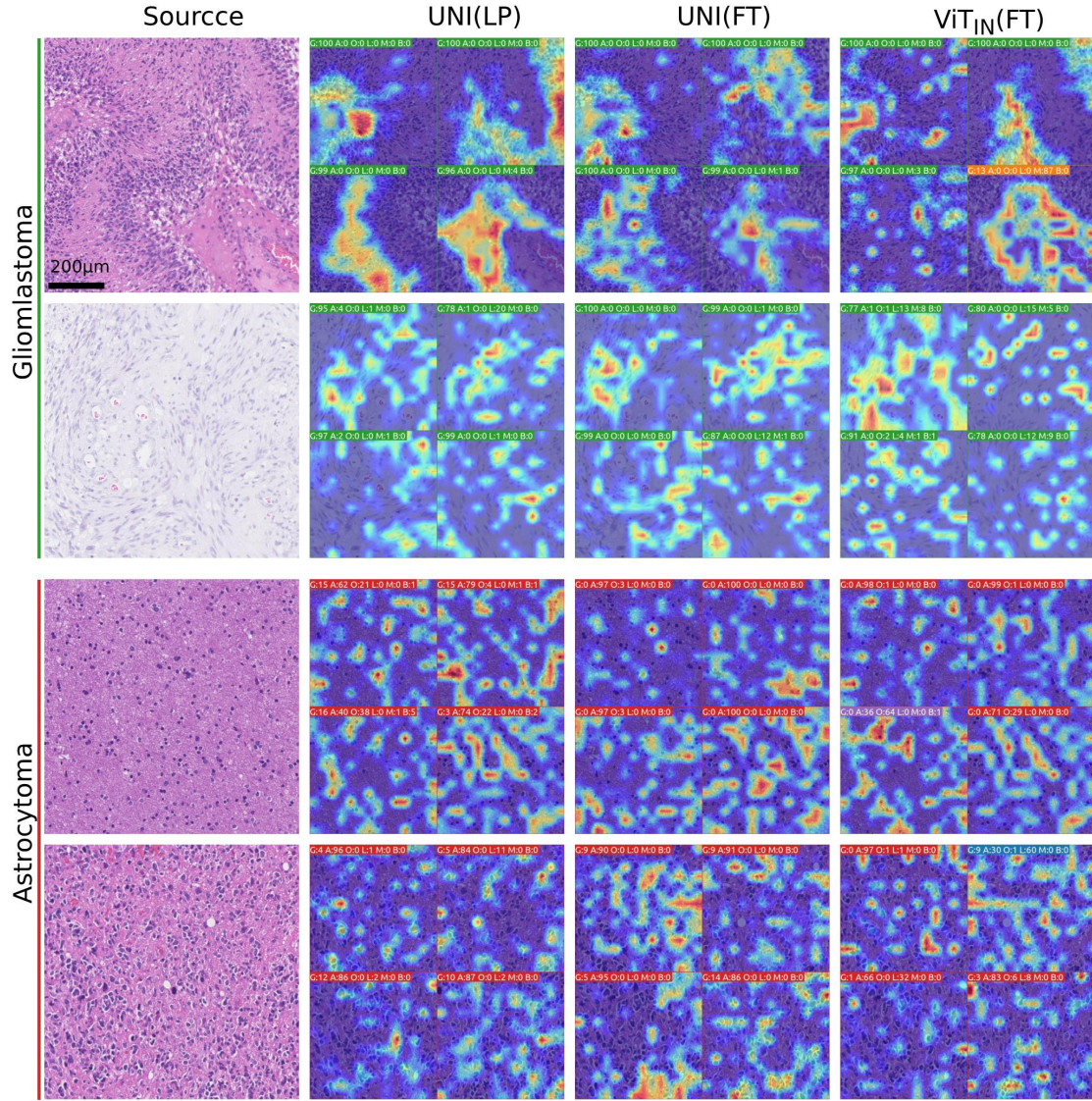


図 17 施設内データセットの低精度症例の GradCAM 分析

施設内データセットで、1 症例に含まれるパッチごとの正解率が 50%を下回った症例での、パッチごとの予測クラスと GradCAM 出力結果を、モデルごとに出力した。

High confidence cases(patch acc.>0.9)



High confidence cases(patch acc.>0.9)

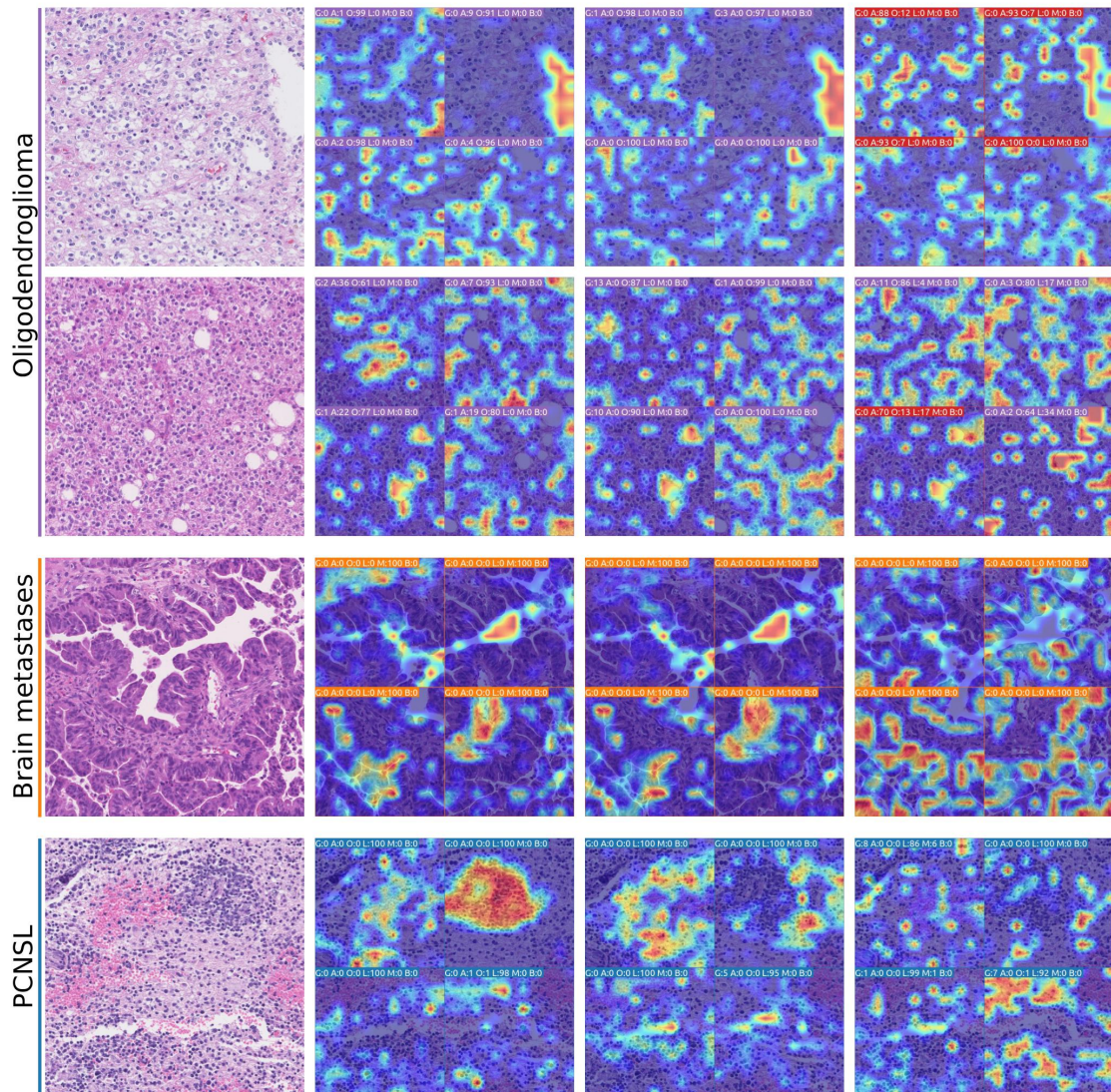
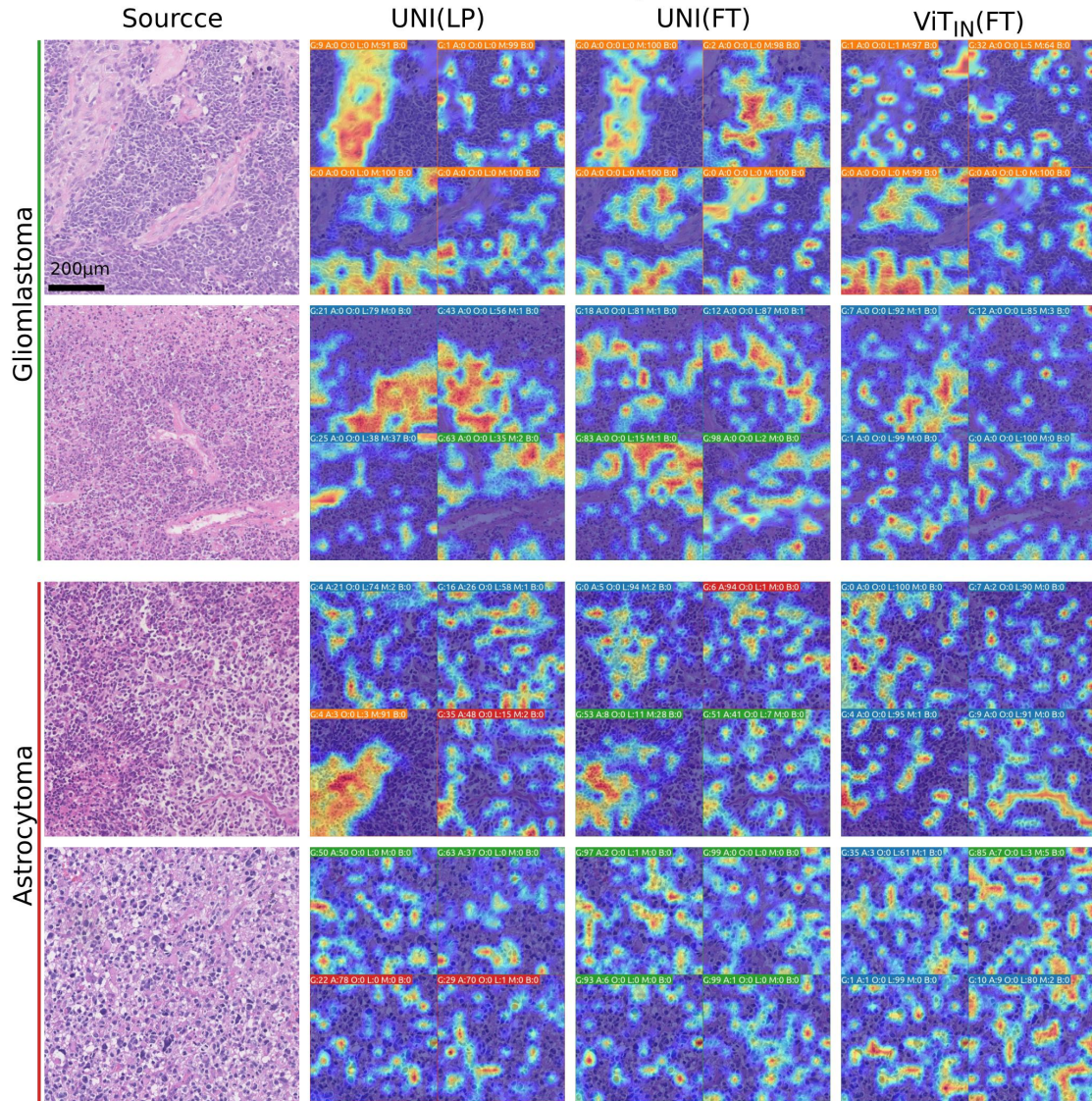


図 18 EBRAINS データセットの高精度症例の GradCAM 分析

EBRAINS データセットで、1 症例に含まれるパッチごとの正解率が 90%を超えた症例での、パッチごとの予測クラスと GradCAM 出力結果を、モデルごとに出力した。

Low confidence cases(patch acc.<0.5)



Low confidence cases(patch acc.<0.5)

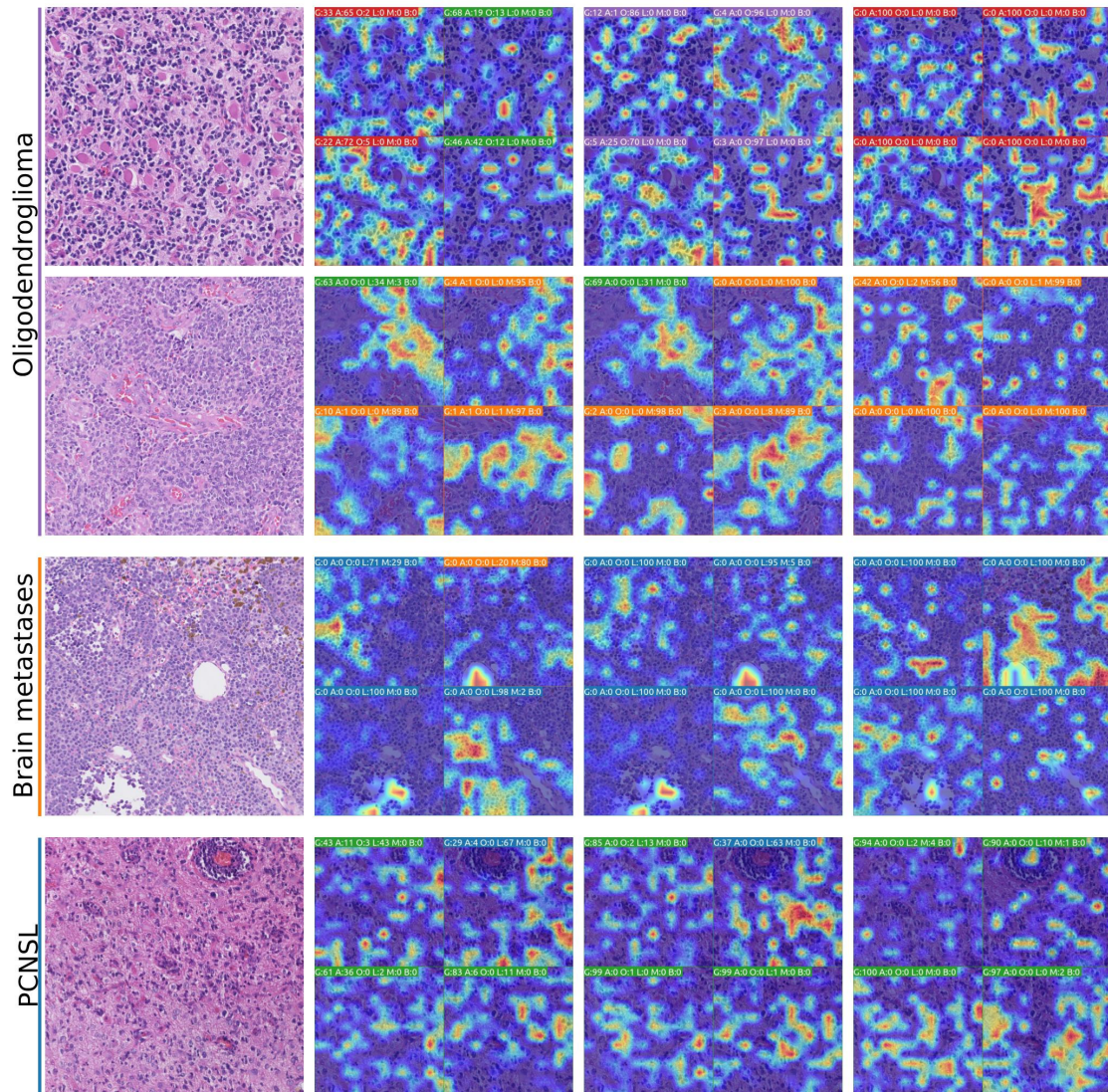


図 19 EBRAINS データセットの低精度症例の GradCAM 分析

EBRAINS データセットで、1 症例に含まれるパッチごとの正解率が 50%を下回った症例での、パッチごとの予測クラスと GradCAM 出力結果を、モデルごとに出力した。

5. 特徴量空間解析

UMAP 可視化により各モデルの特徴量空間を分析した。分析には 1 つの fold の施設内データ (train+val、10 パッチ/症例) と EBRAINS データ (5 パッチ/症例) を使用し、4 つのモデル条件 (元の ViT-L_{IN}、施設内データセットで訓練された ViT-L_{IN}、元の UNI、施設内データセットで訓練された UNI) ごとに、それぞれのパッチに対応した特徴量をデータセットやサブタイプごとの分布を比較した (図 20、21、22、23)。

ImageNet の自然画像のみで訓練された元々の ViT-L_{IN} は、施設内と EBRAINS データセット間でわずかに分布の違いを示したが、腫瘍タイプ間の分離は不明瞭だった (図 20)。施設内データセットでの学習後、ViT-L_{IN} は腫瘍タイプ間の分離が明瞭となった (図 21)。対照的に、病理画像で事前学習された元の UNI は、ローカルな学習なしでも両データセットで腫瘍タイプごとの分離を示した (図 22)。特に、施設内データセットのサンプルは特徴量空間の周辺部に広がるように配置し、EBRAINS サンプルは中心付近に集中する傾向を示した。施設内データセットでの学習後は、ViT-L_{IN} と同様に顕著なクラスター形成を示した (図 23)。EBRAINS データセットのグリオーマ症例の詳細分析では、UNI モデルが IDH 変異の有無よりも形態学的グレードを反映する分布パターンを示した。また、原発部位別の転移性脳腫瘍の分析では、施設内データセットの学習後の全モデルで原発部位によるクラスタリング傾向が認められた。

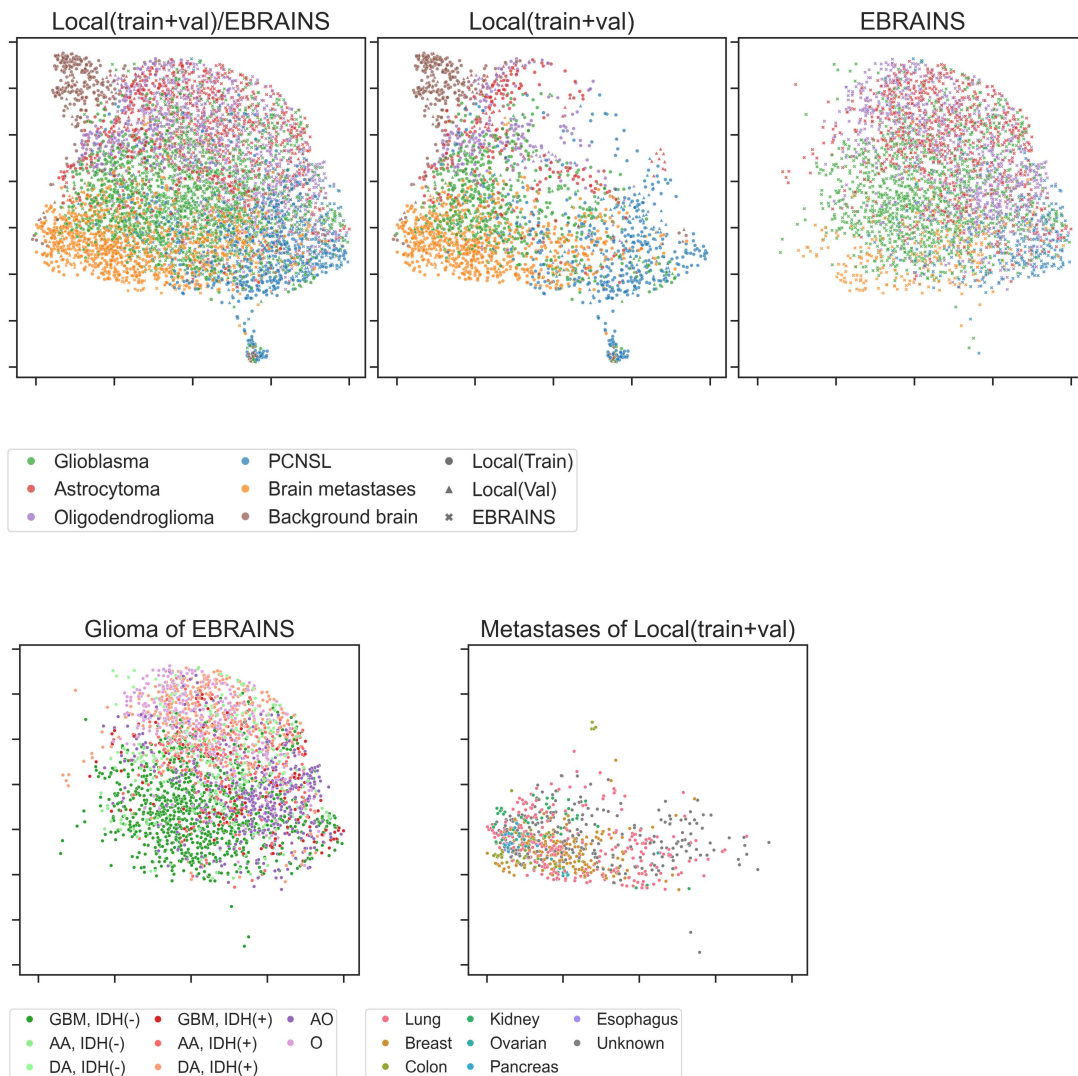


図 20 Original ViT- L_{IN} の UMAP 投射

各パッチを ImageNet で事前訓練された ViT-L に入力し、得られた特徴量をデータセットおよびクラスごとにマーカーをわけてプロットした。上段左から、施設内（訓練と評価）と EBRAINS データセットを統合したもの、施設内データセットのみ、EBRAINS データセットのみの UMAP 散布図を示す。下段は EBRAINS データセットの Glioma クラスのみの UMAP 散布図であり、レジストリに登録された診断でマーカーをわけている。下段右は施設内データセットの M クラス（転移性脳腫瘍）のみの UMAP 散布図であり、既往歴や免疫染色等で原発の腫瘍タイプを特定できたものについて原発の腫瘍種別をマーカーとして表示している。

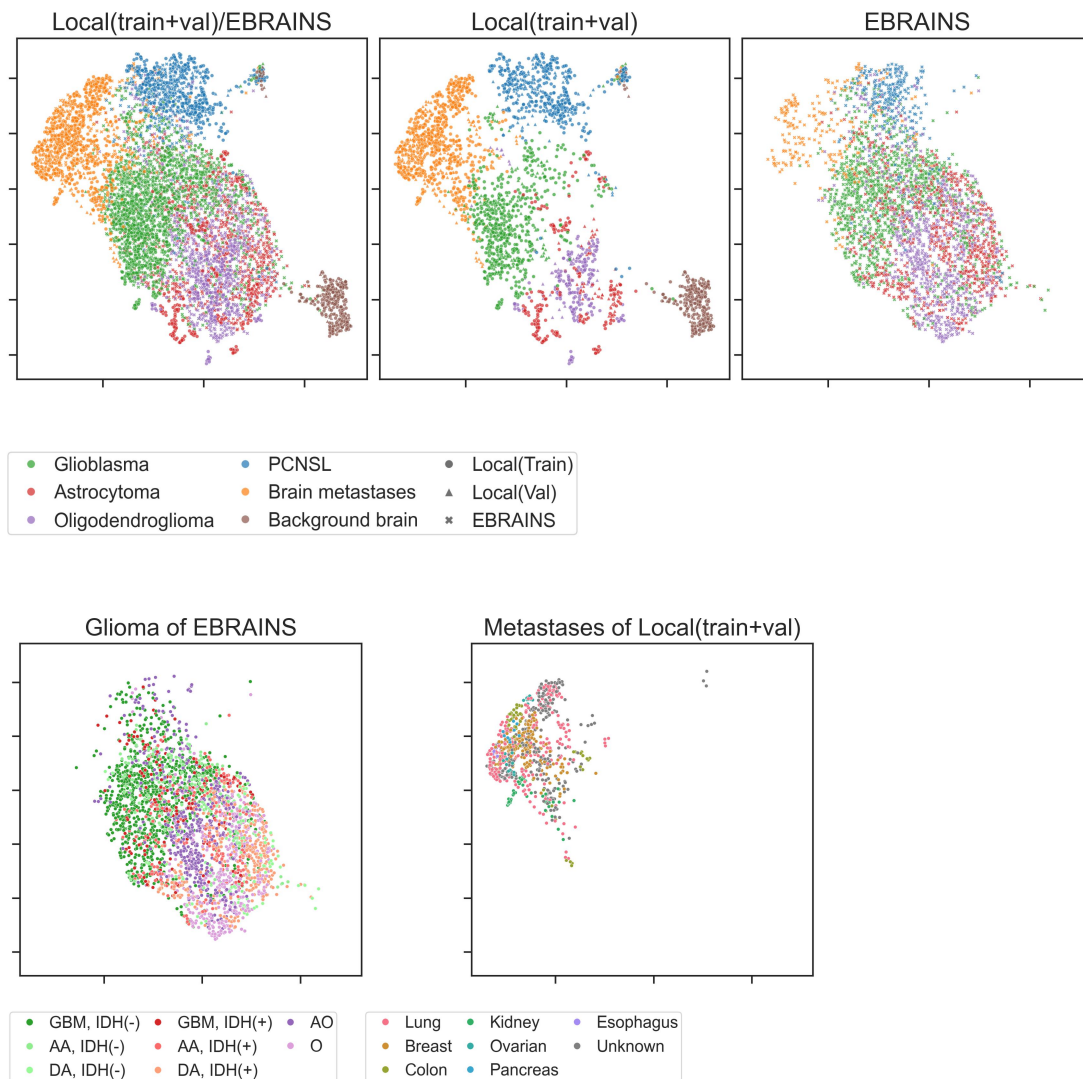


図 21 施設内データセットで訓練された ViT-L_{IN} の UMAP 投射

各パッチを施設内データセットで Fine-tuning した ViT-L_{IN} に入力し、得られた特徴量をデータセットおよびクラスごとにマーカーをわけてプロットした。図の並びは図 20 と同様。

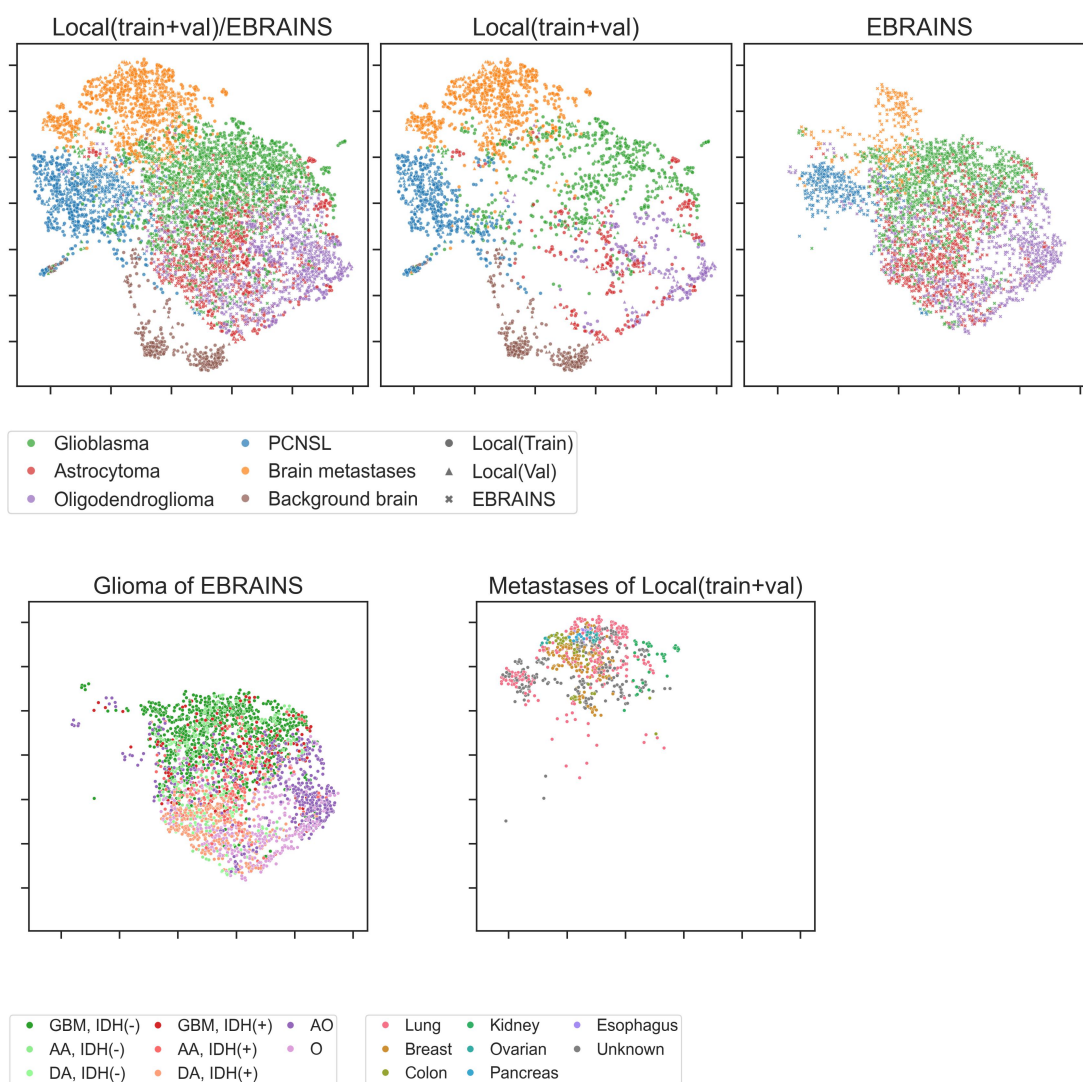


図 22 Original UNI の UMAP 投射

各パッチを事前訓練済みの UNI に入力し、得られた特徴量をデータセットおよびクラスごとにマーカーをわけてプロットした。図の並びは図 20 と同様。

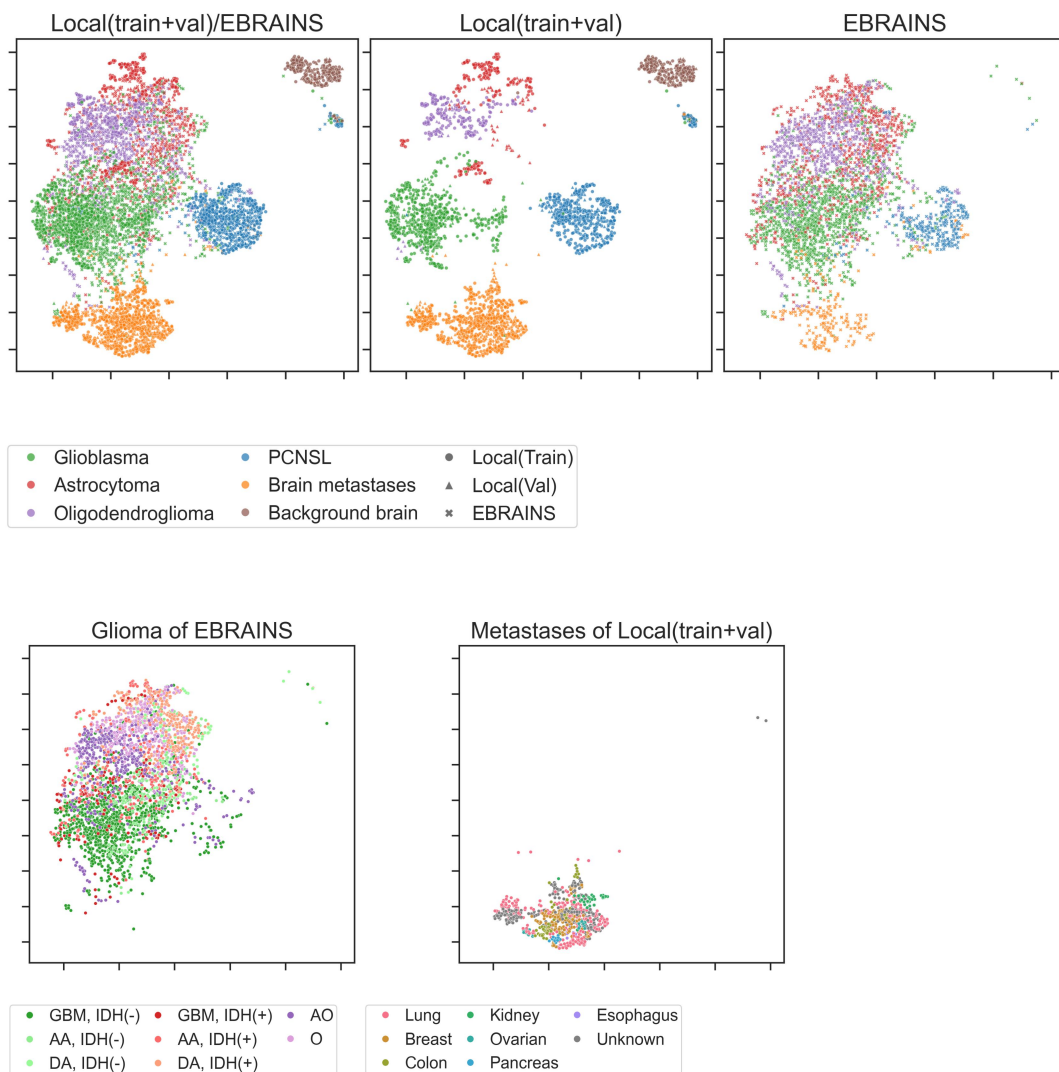


図 23 施設内データセットで訓練された UNI の UMAP 投射

各パッチを施設内データセットで Fine-tuning した UNI に入力し、得られた特徴量をデータセットおよびクラスごとにマーカーをわけてプロットした。図の並びは図 20 と同様。

6. ウェブベース診断支援システム

実用的応用の実証として、病理画像解析のためのインタラクティブなウェブベースシステムを開発した。このシステムは、顕微鏡カメラ統合によるリアルタイム分類、透明度調整可能な GradCAM による決定過程の可視化、分析画像をデータセットパッチの UMAP 可視化に投影することによる特徴空間の探索を可能にした。

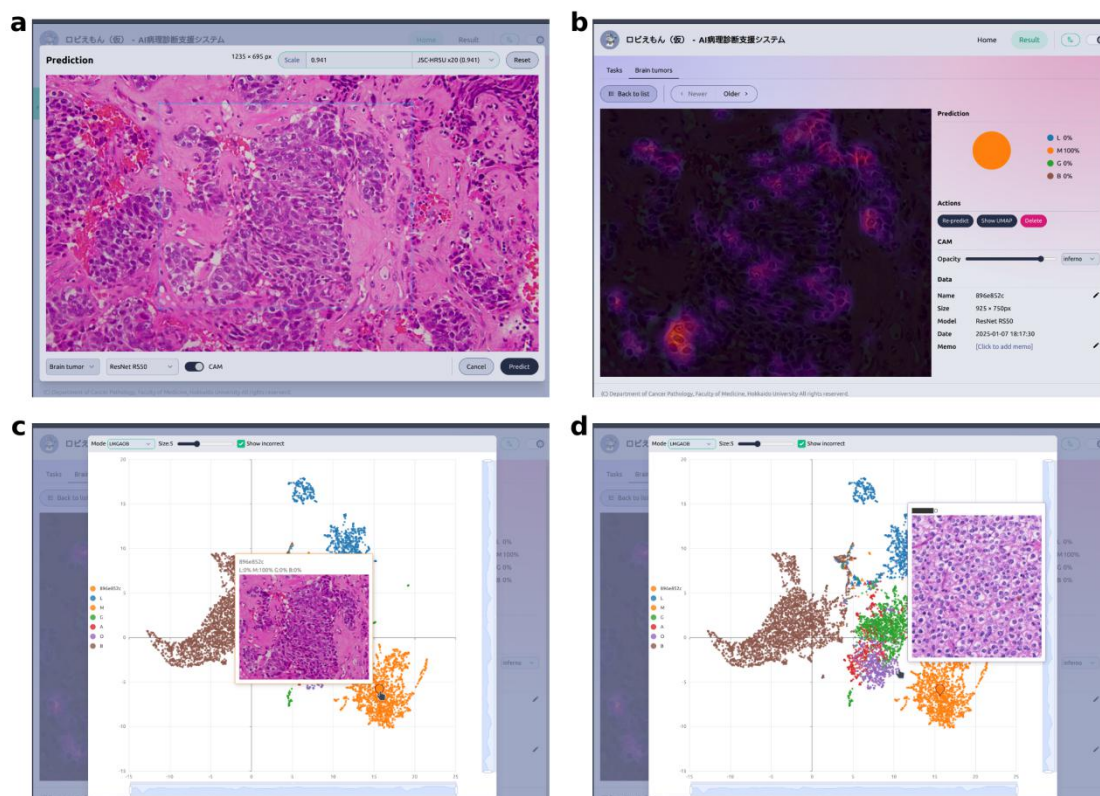


図 24 ウェブベース脳腫瘍病理画像分類アプリケーションの使用例

考察

本研究のもっとも着目すべき結果は、施設内データセットの交差検証では、すべてのモデルで $FT > LP$ の一貫した関係が見られ、外部データセットでも従来型モデルは、CTransPath のパッチ精度を除いてすべて $FT > LP$ の関係は示した一方で、次世代型の病理基盤モデルである UNI は外部データセットにおいてすべての評価指標で $LP > FT$ という逆の関係を示したことである（表 8）。

Model name	Local Dataset	EBRAINS Dataset
ViT-L _{IN}	$FT > LP$	$FT \geq LP$
CTransPath	$FT > LP$	$FT < LP$ and $FT < LP$
UNI	$FT > LP$	$FT < LP$

表 8 モデルごとの FT と LP の性能指標の関係

また、パッチ数増加に伴う性能軌跡の相違においても、施設内データセットでは ViT-L_{IN} 一貫して精度向上が見られたものの、CTransPath は外部データセットでは横ばいないし僅かに増加であり、次世代の病理基盤モデルは向上が見られないもしくは劣化する傾向が見られた（表 9）。

Model name	Local Dataset	EBRAINS Dataset
ViT-L _{IN}	↗	↗
CTransPath	↗	↗ or →
UNI	↗	↘ or →

表 9 モデルと訓練パッチ数増加に伴う性能指標の変化の比較

さらにパッチ数と適用手法を横断した比較では、UNI と Prov-GigaPath は、わずか 10 パッチ/症例という少量のデータでも堅牢な分類性能を示し、LP 手法で 50 倍のパッチ数を必要とする ViT_{IN} や CTransPath といった従来モデルの FT 手法と同等以上の性能を達成した。

これらを考察するにあたり、まず次世代の病理基盤モデルは成り立ちとして、人体病理組織について臓器および疾患について横断的かつ網羅的な超大規模なデータセットを、DINOv2 という洗練された事前学習アルゴリズムによって訓練され、その形態学的な認識能力は高度な汎用性を持つように構築されている点

について認識する必要がある。

このように潜在的に汎用性に焦点を置いて作成モデルを単一施設の限られたデータセットで FT することは、汎化能力を犠牲にして施設特有のパターンに急速に適応することを示している。その結果 FT は LP に対して、施設内のベンチマークではわずかながらも適応能力の向上が見られるものの、外部データセットでは性能劣化に繋がったと考察される。さらに、外部データセットにおいて FT と LP の条件に関わらずパッチの枚数の増加が大きな性能向上に寄与しない（あるいは劣化を起こす）という結果は、次世代の病理基盤モデルのエンコーダは十分な組織認識能力をすでに持ち合わせているためと説明され、これらの考察を支持する。

その一端は UNI の UMAP 投射にも現れており、腫瘍の粗粒度分類については大域的に分離し、グリオーマ内部では G、A、O クラスのサブタイプおよび Grade が局所的な配列として表現される（図 22）一方で、施設内データセットで追加訓練した UNI では、施設内データセットはクラスごとに明瞭なクラスタとして分離されている（図 23）。この過剰に分離されたクラスタリングの様は可適合を視覚的に表現しており、施設内データセットでのわずかながらの性能向上に寄与すると同時に、外部データセットでの汎用性低下の原因になったと解釈できる。

これらの結果の統合によって、「独自の病理データで新しいエンコーダーを訓練する」から「事前学習済みエンコーダーに、データセットを通じて問い合わせる」というパラダイムシフトが示唆される。これは大規模病理データの収集から、事前学習された特徴表現の効率的活用への移行を意味する。つまり病理基盤モデルに埋め込まれた汎用性の高い視覚的認識フレームにアクセスする手段として、データセット設計を位置づけるのである。具体的には、画像数やデータ量は少量でも、広範な疾患をカバーするような薄いデータセットを用いて、病理基盤モデルの生成する特徴量に対して、linear probing のような非破壊的学習等を通じて問い合わせ、新たな形態学的認識の知見を抽出を試みるようなアプローチが実験として成立しうると考える。

また、本研究では LP が FT よりも良い性能を示した意味は大きい。それは、LP は FT に比してその性質上計算資源の要求が文字通り桁違いに少なく、環境負荷や実験の施行回数の増加や効率化につながる点にある。LP で FT に対して十分に競合的な性能を示せるのであればその選択の妥当性は高い。

本実験で見られた病理基盤モデルでの FT<LP の性能の逆転現象は、特有の過学習の一形態を表していると解釈できる。外部データセットでの FT が LP に対して示した性能劣化は、ニューラルネットワークで知られる「破滅的忘却 (catastrophic forgetting)」(French, 1999)に類似している。これは、モデルが新しいタスクに適応する過程で過適合を起こし事前学習された汎化能力を失う現象である。この問題に対処するため、EWC (Elastic Weight Consolidation) (Kirkpatrick et al., 2017)、LoRA (Low-Rank Adaptation) (Hu et al., 2021)、Bottleneck Adapter(Zhang et al., 2023)などのパラメータ効率的手法が開発されてきた。実際、公開データセットを用いた病理病理基盤モデルの適応戦略の検討(Lee et al., 2024)では、FT と LP が同一データセット内で競合的な性能を示した一方、LoRA を用いた PEFT (Parameter-Efficient Fine Tuning) がより優れた汎化能力を示した。すでに有効性が証明された LoRA に加え、EWC や Bottleneck Adapter などのパラメータ効率的手法の導入が有望な解決策となる可能性がある。

これらの方法論の検討に加えて、病理基盤モデル自体の選択も重要な考慮点となる。Prov-GigaPath は同一手法で事前訓練された UNI よりも、モデルパラメーター数と事前訓練データセットの規模の両面で上回るが、本研究の性能指標では UNI が Prov-GigaPath を上回るものが多かった。ただ複数のベンチマークで Prov-GigaPath が優れた性能を示す報告が見られるなど、この優劣関係は一様ではなく、同一文献内でも腫瘍種別などタスク依存的な変動が報告されている (Zimmermann et al., 2024; Chen et al., 2024)。病理基盤モデルの事前学習データセットの臓器レベルの分布は公開されているが、組織型レベルの詳細情報は利用できず、これらの分布の違いが各モデルの特性を決定している可能性が高い。したがって、モデル選択では、モデルパラメーターや事前学習データセットの規模だけでなく、公開されたベンチマーク結果の慎重な検討と実際のテストデータでの検証性能を考慮すべきである。

本研究を通してデータセット設計上の実践的な課題も明らかになった。それは、UMAP 散布図で見られた施設内と EBRAINS データ間の明確な分布の乖離から見出すことができ、これは主に施設間の染色変動を反映している可能性が高い。実際に図組織画像を比較すると、施設内データセットはエオジンの鮮やかな好酸性染色が見られ染色性は症例ごとに比較的均一なのに対し、EBRAINS データセットは青みが強くその色調の程度は症例ごとに異なっている。この問題に対処するためには、染色の多様性を考慮した訓練方法を検討すべきである(Ochi et al., 2024)。簡単なアプローチとして、色正規自体に機械学習モデルを用いるよ

うな画像前処理技術(Tellez et al., 2019)がこの課題を克服できる可能性がある。

さらに本研究で得られたデータセットやモデルを活用したシステムの実用化において特に懸念される点は、A クラスと O クラスの細粒度分類における低い精度である。混同行列が示すように、これらの腫瘍型間で頻繁な誤分類が発生した。この問題の背景には、症例数の限界という課題がある。実際、施設内データセットにおけるこれらの症例の少なさは、日本脳神経外科学会の全国調査(Iihara et al., 2021)で報告された実際の分布を反映しており、有限期間での単一施設データ収集の固有の制約を浮き彫りにしている。この問題に対処するには、より大規模な多施設データ収集とドメイン適応技術による汎化能力の向上が必要である一方、このような希少症例に対する効率的な学習方法の開発も同様に重要である。

これらの技術的課題にもかかわらず、粗粒度分類で達成された高い精度は、臨床 AI システムの応用、特に術中迅速診断への強い可能性を示している。神経膠腫、PCNSL、転移性脳腫瘍の鑑別は術中の意思決定に直接影響する(Sasaki & Nagane, 2021; Nakamura et al., 2022)。注目すべきことに、これら 3 つの腫瘍に特化した迅速診断モデルはこれまで報告されていない(Zhang et al., 2024; Shi et al., 2023; Nasrallah et al., 2023)。本研究で開発したウェブ診断支援システムは、リアルタイム分類表示、GradCAM 可視化、形態学的特徴の相対化を実装しており、診断支援と教育の両面での潜在的な利益を提供する。現在の腫瘍領域を事前を選択する必要性の制約は、顕微鏡カメラに接続された診断支援システムには大きな影響を与えない可能性があるが、より包括的な分析のために、Multiple Instance Learning(Ilse et al., 2018)や関心機構を考慮した実装(Ding et al., 2023)による全スライド画像解析は検討に値する。

結論

本研究は、病理画像分類における次世代の病理基盤モデルや機械学習モデルの適応戦略の包括評価を通じて、データセット設計における新たなパラダイムシフトを実証した。これらの知見は、より効率的で信頼性の高い医療 AI システムの開発に向けた理論的基盤として貢献するものである。

謝辞

適切な助言と直接の御指導を賜りました北海道大学大学院医学研究院腫瘍病理学教室、田中伸哉教授、小田義崇助教に感謝の意を表します。最後に、本論文を作成するに当たってご指導・ご協力を頂いたすべての皆様に、重ねて御礼申し上げます。

利益相反

開示すべき利益相反状態はない。

引用文献

- Azizi, S. et al., 2023. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature biomedical engineering*, 7(6), pp.756–779.
- Bello, I. et al., 2021. Revisiting ResNets: Improved training and scaling strategies. *arXiv [cs.CV]*. Available at: <http://arxiv.org/abs/2103.07579>.
- van den Bent, M.J. et al., 2023. Primary brain tumours in adults. *Lancet*, 402(10412), pp.1564–1579.
- Buslaev, A. et al., 2020. Albumentations: Fast and flexible image augmentations. *Information (Basel)*, 11(2), p.125.
- Chen, R.J. et al., 2024. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3), pp.850–862.
- Deng, J. et al., 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). IEEE, pp. 248–255.
- Ding, J. et al., 2023. LongNet: Scaling transformers to 1,000,000,000 tokens. *arXiv [cs.CL]*. Available at: <http://arxiv.org/abs/2307.02486> [Accessed June 28, 2024].
- Dosovitskiy, A. et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv [cs.CV]*. Available at: <http://arxiv.org/abs/2010.11929> [Accessed December 2, 2024].
- Enda, K. (2025) ‘HUPatho2BrainTumor’. Hokkaido University. Available at: <https://doi.org/10.5281/ZENODO.16693636>.
- French, R.M., 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4), pp.128–135.
- Hollon, T. et al., 2023. Artificial-intelligence-based molecular classification of diffuse gliomas using rapid, label-free optical imaging. *Nature medicine*, 29(4), pp.828–832.
- Hu, E.J. et al., 2021. LoRA: Low-Rank Adaptation of large language models. *arXiv [cs.CL]*. Available at: <http://arxiv.org/abs/2106.09685> [Accessed January 10, 2025].
- Iihara, K. et al., 2021. The Japan Neurosurgical Database: Statistics update 2018 and 2019. *Neurologia medico-chirurgica*, 61(12), pp.675–710.

- Ilse, M., Tomczak, J.M. & Welling, M., 2018. Attention-based deep multiple instance learning. *arXiv [cs.LG]*. Available at: <http://arxiv.org/abs/1802.04712>.
- Jin, L. et al., 2021. Artificial intelligence neuropathologist for glioma classification using deep learning on hematoxylin and eosin stained slide images and molecular markers. *Neuro-oncology*, 23(1), pp.44–52.
- Kim, H.E. et al., 2022. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1), p.69.
- Kingma, D.P. & Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv [cs.LG]*. Available at: <http://arxiv.org/abs/1412.6980> [Accessed December 2, 2024].
- Kirkpatrick, J. et al., 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13), pp.3521–3526.
- Lee, J. et al., 2024. Benchmarking pathology foundation models: Adaptation strategies and scenarios. *arXiv [cs.CV]*. Available at: <http://arxiv.org/abs/2410.16038> [Accessed January 24, 2025].
- Louis, D.N. et al., 2021. The 2021 WHO classification of tumors of the Central Nervous System: A summary. *Neuro-oncology*, 23(8), pp.1231–1251.
- Ma, Y. et al., 2023. Histopathological auxiliary system for brain tumour (HAS-Bt) based on weakly supervised learning using a WHO CNS5-style pipeline. *Journal of neuro-oncology*, 163(1), pp.71–82.
- MahmoodLab/UNI · Hugging Face. Available at: <https://huggingface.co/MahmoodLab/UNI> [Accessed October 31, 2025a].
- McInnes, L. et al., 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of open source software*, 3(29), p.861.
- Nakamura, H. et al., 2022. The Japan Society for Neuro-Oncology guideline on the diagnosis and treatment of central nervous system germ cell tumors. *Neuro-oncology*, 24(4), pp.503–515.
- Nasrallah, M.P. et al., 2023. Machine learning for cryosection pathology predicts the 2021 WHO classification of glioma. *Med (New York, N.Y.)*, 4(8), pp.526–540.e4.
- Oquab, M. et al., 2023. DINOv2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193. Available at: <http://dx.doi.org/10.48550/arXiv.2304.07193>.

- Ostrom, Q.T. et al., 2023. CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2016-2020. *Neuro-oncology*, 25(12 Suppl 2), pp.iv1–iv99.
- Anon, prov-gigapath/prov-gigapath · Hugging Face. Available at: <https://huggingface.co/prev-gigapath/prev-gigapath> [Accessed October 31, 2025b].
- Redlich, J.-P. et al., 2024. Applications of artificial intelligence in the analysis of histopathology images of gliomas: a review. *Npj imaging*, 2(1), p.16.
- Roetzer-Pejrimovsky, T. et al., 2022. The Digital Brain Tumour Atlas, an open histopathology resource. *Scientific data*, 9(1), p.55.
- Sasaki N. & Nagane M., 2021. Treatment of primary central nervous system lymphoma: Standard treatments according to the Japanese 2019 guideline and novel treatments. *Shinkei kenkyu no shinpo [Brain and nerve]*, 73(10), pp.1107–1114.
- Selvaraju, R.R. et al., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 618–626.
- Shi, L. et al., 2023. Contribution of whole slide imaging-based deep learning in the assessment of intraoperative and postoperative sections in neuropathology. *Brain pathology (Zurich, Switzerland)*, 33(4), p.e13160.
- Tellez, D. et al., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58(101544), p.101544.
- The Japan Neurosurgical Society and The Japanese Society of Pathology, 2023. *General rules for clinical and pathological studies on brain tumors. 5th edition*, Kanahara Shuppan.
- Villanueva-Meyer, J.E., Mabray, M.C. & Cha, S., 2017. Current clinical brain tumor imaging. *Neurosurgery*, 81(3), pp.397–415.
- Vorontsov, E. et al., 2024. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 30(10), pp.2924–2935.
- Wang, W. et al., 2023. Neuropathologist-level integrated classification of adult-type diffuse gliomas using deep learning from whole-slide pathological images. *Nature communications*, 14(1), p.6359.

Wang, X. et al., 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81(102559), p.102559.

Wightman, R. et al., 2023. *rwightman/pytorch-image-models: v0.8.10dev0 Release*, Zenodo. Available at: <https://github.com/rwightman/pytorch-image-models>.

Xu, H. et al., 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015), pp.181–188.

Zhang, R. et al., 2023. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv [cs.CV]*. Available at: <http://arxiv.org/abs/2303.16199> [Accessed January 24, 2025].

Zhang, X. et al., 2024. A multicenter proof-of-concept study on deep learning-based intraoperative discrimination of primary central nervous system lymphoma. *Nature communications*, 15(1), p.3768.

Zimmermann, E. et al., 2024. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv [cs.CV]*. Available at: <http://arxiv.org/abs/2408.00738> [Accessed November 28, 2024].