

ARTICLE

PathRes_FM: A regularized ResNet-50 framework
for multi-class colorectal histopathology tissue
classificationPrakashini Ambati^{1*} and Harikrishna Bommala²¹Department of Computer Science and Engineering, Bharatiya Engineering Science & Technology Innovation University, Anantapur, Andhra Pradesh, India²Department of Computer Science and Engineering, KG Reddy College of Engineering and Technology, Hyderabad, Telangana, India

Abstract

Accurate tissue-level classification in colorectal histopathology is important for developing reliable computer-aided pathology systems, particularly when multiple tissue categories show high morphological similarity. This study proposes PathRes_FM, a regularized ResNet-50-based deep learning framework for multi-class colorectal histopathology tissue classification using the PathMNIST benchmark dataset. The framework classifies nine colorectal tissue categories: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium. The proposed pipeline uses an ImageNet-pretrained ResNet-50 backbone with a task-specific dropout-regularized classification head. To improve generalization, the training strategy incorporates image augmentation, MixUp, CutMix, label smoothing, AdamW optimization, and OneCycleLR scheduling. Experimental evaluation demonstrated strong predictive performance on the PathMNIST benchmark. The proposed PathRes_FM model achieved an overall accuracy of 94.04%, an area under the receiver operating characteristic curve of 0.9890, a weighted F1-score of 0.9399, a weighted precision of 0.9410, and a weighted recall of 0.9404. Class-wise analysis showed strong performance for adipose, lymphocytes, mucus, normal colon mucosa, and colorectal adenocarcinoma epithelium, while cancer-associated stroma and smooth muscle remained comparatively challenging due to morphological overlap. Comparative evaluation further showed that PathRes_FM outperformed Auto-Sklearn, ResNet-18, and ResNet-50 baseline models on the same benchmark setting. Additional analyses using confusion matrix, receiver operating characteristic curves, precision-recall curves, t-distributed stochastic neighbor embedding visualization, calibration assessment, robustness plots, and gradient-weighted class activation mapping-style interpretation supported the discriminative and interpretable behavior of the learned representations. The results indicate that PathRes_FM is an effective supervised image-level framework for colorectal histopathology tissue classification and provides a strong baseline for future extensions toward whole-slide and pathology-specific representation-learning systems. This work explicitly positions PathRes_FM as a reproducible benchmark-level supervised transfer-learning framework rather than as a new whole-slide foundation model or clinical deployment system.

***Corresponding author:**Prakashini Ambati
(2022PCSE019@bestiu.edu.in)

Citation: Ambati, P., & Bommala, H. (2026). PathRes_FM: A regularized ResNet-50 framework for multi-class colorectal histopathology tissue classification. *Int J Systematic Innovation*, 10(5): 026210102.
[https://doi.org/10.6977/IJoSI.202610_10\(5\).0010](https://doi.org/10.6977/IJoSI.202610_10(5).0010)

Received: May 18, 2026**Revised:** June 11, 2026**Accepted:** June 29, 2026**Published online:** October 15, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: Colorectal histopathology; PathMNIST; Tissue classification; ResNet-50; Deep learning; Computational pathology; MixUp and CutMix; Gradient-weighted class activation mapping

1. Introduction

Digital pathology has reshaped histopathological practice by converting conventional glass slides into high-resolution whole-slide images (WSIs), thereby enabling scalable computational analysis of tissue morphology for diagnosis, prognosis, biomarker prediction, and workflow support. Early progress in this area was driven by supervised and weakly supervised deep learning frameworks, which demonstrated that histopathology images contain rich diagnostic information that can be learned directly from image data (Campanella *et al.*, 2019; Coudray *et al.*, 2018; Lu *et al.*, 2021; Shao *et al.*, 2021; Yang *et al.*, 2023). These studies established that artificial intelligence (AI) can assist pathology not only at the patch level but also at the slide level, where gigapixel WSIs must be interpreted under limited annotation availability. As a result, computational pathology has evolved from task-specific image classifiers toward more general representation-learning systems capable of supporting multiple downstream clinical applications (Campanella *et al.*, 2019; Chen *et al.*, 2024; Coudray *et al.*, 2018; H. Xu *et al.*, 2024).

However, a major limitation is that typical WSI pathology AI is tightly tied to specific labels and training images. When datasets are limited, especially for rare diseases and underrepresented clinical settings, annotated WSIs may be sparse, and dedicated task-specific models may be difficult to train. This has recently driven research in digital pathology toward weakly supervised learning, transferable pretraining, and self-supervised learning (Chen *et al.*, 2024; Lu *et al.*, 2021; Shao *et al.*, 2021; Wang *et al.*, 2022). In particular, transformer-based and contrastive pretraining methods such as TransMIL and CTransPath demonstrated that histopathology-specific feature learning can improve slide-level understanding beyond standard natural-image initialization (Shao *et al.*, 2021; Wang *et al.*, 2022). These advances laid the technical foundation for a new generation of pathology foundation models, where the objective is not merely to solve a single classification task but to learn universal tissue representations from large-scale unlabeled or weakly paired pathology corpora (Chen *et al.*, 2024; Lu, Chen, Williamson, Chen, Liang, *et al.*, 2024; Vorontsov *et al.*, 2024; H. Xu *et al.*, 2024).

Recent 2024–2025 studies have accelerated this transition through the introduction of large-scale pathology foundation models. UNI showed that self-

supervised pretraining on large diagnostic WSI collections can yield a general-purpose pathology encoder with broad downstream transferability (Chen *et al.*, 2024). CONCH further demonstrated the value of vision-language alignment in computational pathology, indicating that biomedical text can provide useful semantic supervision beyond image-only training (Lu, Chen, Williamson, Chen, Liang, *et al.*, 2024). Prov-GigaPath extended this direction by combining large-scale representation learning with explicit whole-slide modeling, while Virchow established that clinical-grade foundation models can support pan-cancer analysis and rare-cancer detection at scale (Vorontsov *et al.*, 2024; H. Xu *et al.*, 2024). More recent systems such as TITAN, mSTAR, PathOrchestra, and related multimodal foundation models have expanded the foundation-model paradigm toward multimodal slide-text learning, large-scale generalization, and clinically diverse task coverage (Ding *et al.*, 2025; Y. Xu *et al.*, 2025; Yan *et al.*, 2025). Collectively, these works show that pathology foundation models are emerging as a central paradigm for transferable and data-efficient histopathology intelligence (Chen *et al.*, 2024; Ding *et al.*, 2025; Lu, Chen, Williamson, Chen, Liang, *et al.*, 2024; Vorontsov *et al.*, 2024; Xu *et al.*, 2024).

At the same time, clinically meaningful pathology AI is moving beyond purely recognition-based tasks. PathChat illustrated that a pathology-adapted vision encoder can be connected with generative language modeling for interactive pathology assistance (Lu, Chen, Williamson, Chen, Zhao, *et al.*, 2024). HistoGPT and subsequent report-generation frameworks further suggested that pathology systems may evolve from classification-only tools to assistants capable of producing draft structured reports from gigapixel image evidence (Hu *et al.*, 2025; Tran *et al.*, 2025). This transition is important because pathology reporting remains labor-intensive and highly dependent on domain expertise, particularly when specialist availability is limited. Nevertheless, automatic report generation in pathology remains challenging due to the need for faithful image-text grounding, clinically constrained outputs, and robust generalization under limited-data conditions (Hu *et al.*, 2025; Lu, Chen, Williamson, Chen, Zhao, *et al.*, 2024; Tran *et al.*, 2025).

Despite these advances, several research challenges remain. Many recent pathology foundation models

derive their strength from extremely large pretraining resources, which may not be readily available in many institutions. In addition, real-world deployment still faces difficulties related to stain variation, scanner variation, out-of-distribution samples, and data-efficient downstream adaptation (Campanella *et al.*, 2025; Nicke *et al.*, 2025; Pocevičiūtė *et al.*, 2025; Vorontsov *et al.*, 2024). Benchmarking studies have shown that adaptation strategy is often as important as encoder scale, while robustness studies indicate that transferability alone does not eliminate clinical deployment risks (Campanella *et al.*, 2025; Lee *et al.*, 2025; Neidlinger *et al.*, 2025; Pocevičiūtė *et al.*, 2025). These issues are especially relevant for data-scarce pathology settings, where the need for universal slide representations is strongest but the opportunity for large-scale supervised fine-tuning is weakest (Lee *et al.*, 2025; Nicke *et al.*, 2025; Pocevičiūtė *et al.*, 2025).

Although recent pathology foundation models have shown the potential of large-scale self-supervised and multimodal learning, the current study focuses on a practical and reproducible supervised image-level classification problem. Specifically, this work investigates whether a regularized ResNet-50-based transfer-learning pipeline can provide accurate and balanced classification of nine colorectal tissue types using the PathMNIST benchmark. This focus is appropriate because benchmark-level image classification remains an important stage in computational pathology research, especially for evaluating model discrimination, class-wise robustness, and interpretability before extending the framework toward whole-slide or multimodal systems.

Deep learning has an increasingly important role in colorectal cancer (CRC) diagnosis because it can learn discriminative morphological patterns from histopathology, endoscopy, radiology, and related diagnostic data. Recent reviews have emphasized that deep learning-based systems may support CRC detection, tissue classification, lesion characterization, biomarker prediction, and clinical decision support, although careful validation is still required before routine clinical implementation (Bousis *et al.*, 2023). In the present work, this motivation is narrowed to the histopathology image-classification setting, where accurate recognition of colorectal tissue types can support the development of reliable computer-aided pathology workflows.

Healthcare staff also play an essential preventive role in CRC control by identifying eligible individuals, educating patients about screening options, encouraging adherence to fecal testing or colonoscopy pathways, and ensuring follow-up of abnormal findings. Screening is recommended for average-risk adults beginning at 45

years and continuing through 75 years, with individualized decisions for selected adults aged 76 to 85 years (Davidson *et al.*, 2021). Therefore, AI-based histopathology systems should be viewed as complementary tools within a broader prevention and screening pathway led by trained healthcare professionals, rather than replacing clinical screening programs.

The main contribution of this study is the development and evaluation of PathRes_FM, a regularized deep learning framework that combines ImageNet-pretrained ResNet-50 feature extraction, a dropout-based classification head, standard image augmentation, MixUp, CutMix, label smoothing, AdamW optimization, and OneCycleLR scheduling. In addition to global classification metrics, the study evaluates class-wise performance, confusion behavior, ROC and precision-recall characteristics, feature-space separability, calibration, robustness, and gradient-weighted class activation mapping (Grad-CAM)-style visual explanation. Therefore, the proposed work is positioned as a supervised, interpretable, and reproducible colorectal histopathology classification framework rather than a fully developed whole-slide foundation model.

To avoid overstatement of novelty, this study clarifies that the methodological contribution is not a new neural network architecture or a new pathology foundation model. The contribution lies in a carefully regularized and reproducible ResNet-50 transfer-learning pipeline, supported by consistent benchmark evaluation, class-wise analysis, calibration assessment, feature-space visualization, and Grad-CAM-style interpretation for nine-class colorectal histopathology tissue classification.

2. Literature review

Computational pathology has progressed from conventional patch-based histopathology classifiers toward increasingly scalable slide-level and representation-learning paradigms. Early deep-learning studies established that diagnostic and molecular information can be inferred directly from histopathology images, thereby demonstrating the feasibility of end-to-end pathology AI for clinically relevant tasks (Campanella *et al.*, 2019; Coudray *et al.*, 2018). Subsequent work addressed the weak-label nature of pathology data by introducing multiple-instance and attention-based whole-slide learning strategies, which enabled slide-level prediction without exhaustive region annotation (Lu *et al.*, 2021; Shao *et al.*, 2021). In parallel, standardized benchmarks such as MedMNIST and its PathMNIST subset made colorectal histopathology image classification more accessible by providing lightweight, reproducible patch-level datasets derived from CRC histology, thus supporting rapid

methodological comparison across models (Yang *et al.*, 2023). These developments are directly relevant to the current study implementation, which is based on colorectal histopathology tissue classification using image-level data rather than full WSI pretraining.

A second major direction has focused on improving pathology-specific feature extraction through self-supervised and transformer-based representation learning. Generic ImageNet pretraining is often suboptimal for pathology because histomorphological patterns differ substantially from natural-image statistics. To address this, CTransPath introduced transformer-based unsupervised contrastive learning tailored to histopathological images and showed that pathology-specific pretraining can yield stronger and more transferable features than conventional initialization (Wang *et al.*, 2022). Related WSI learning frameworks such as TransMIL further demonstrated that long-range contextual dependencies are important in histopathology, especially when tissue patterns must be interpreted in relation to the surrounding microenvironmental structures rather than as isolated patches (Shao *et al.*, 2021). These studies are methodologically important because they support the broader motivation behind transferable pathology representations learned from pathology data, even though the current implementation remains an image-level supervised classification pipeline.

The most influential recent advance is the emergence of pathology foundation models. UNI demonstrated that a self-supervised pathology encoder pretrained on more than 100 million images from over 100,000 diagnostic hematoxylin and eosin WSIs across 20 tissue types can transfer strongly across 34 downstream computational pathology tasks (Chen *et al.*, 2024). CONCH extended this paradigm by showing that visual-language pretraining using more than 1.17 million image-caption pairs improves clinically relevant pathology understanding beyond vision-only baselines (Lu, Chen, Williamson, Chen, Liang, *et al.*, 2024). Prov-GigaPath then showed that explicit whole-slide modeling combined with real-world large-scale data can achieve state-of-the-art results across a broad range of digital pathology tasks (H. Xu *et al.*, 2024). Virchow further strengthened the case for pathology foundation models by reporting clinical-grade pan-cancer detection and strong performance in rare cancers, highlighting the promise of large pretrained encoders for data-scarce diagnostic settings (Vorontsov *et al.*, 2024).

Recently, the field has started branching out beyond pure images to address vision-language learning tasks in pathology. PathChat combines a vision model with a language model and uses instruction-tuning across

pathology image-text prompts for question answering and interactive pathology assistance (Lu, Chen, Williamson, Chen, Zhao, *et al.*, 2024). TITAN extends this direction to WSIs using reports and captions, showing that slide-level vision-language pretraining can support diagnosis, prognosis, biomarker prediction, retrieval, and report-related tasks (Ding *et al.*, 2025). mSTAR similarly incorporates pathology reports and molecular context into whole-slide foundation modeling, demonstrating how textual and molecular information may enrich pathology representation learning (Y. Xu *et al.*, 2025). These directions are relevant for future multimodal extensions, but they are beyond the scope of the current image-level PathMNIST implementation.

The generation of pathology reports and support for clinical workflows are another important research direction. HistoGPT uses a vision-language model to generate dermatopathology reports from gigapixel WSIs (Tran *et al.*, 2025). Knowledge-retrieval and multi-level region-selection strategies have also been proposed to improve report faithfulness and focus on clinically salient regions (Hu *et al.*, 2025). These studies are relevant to future extensions of the present work, but they also show that report generation remains less mature and more tissue- and setting-specific than image classification or representation-learning tasks (Hu *et al.*, 2025; Tran *et al.*, 2025). Accordingly, the current image-only colorectal tissue classification study provides a focused benchmark-level contribution within this context.

Recent benchmarking and deployment-focused research in pathology shows that strong pretraining performance does not automatically guarantee generalization and stability. Benchmarking across multiple datasets has shown that the choice of adaptation method, task type, and test setting can substantially affect pathology model performance (Lee *et al.*, 2025; Neidlinger *et al.*, 2025). Efficient parameter-based tuning methods can be practical in low-resource settings, whereas cross-cohort evaluations reveal variation in generalization when models are tested outside their training scope (Campanella *et al.*, 2025; Lee *et al.*, 2025). Studies on out-of-distribution detection and real-world deployment further indicate that scanner variation, stain variability, institutional shift, and low-resource settings remain major obstacles to clinical translation (Campanella *et al.*, 2025; Pocevičiūtė *et al.*, 2025). These findings highlight why class-wise performance, confusion behavior, calibration, and robustness analysis are important in colorectal tissue classification.

From the perspective of the present study, the literature supports three conclusions. First, colorectal histopathology

image classification remains an active and meaningful problem, particularly in benchmark settings derived from PathMNIST and related tissue datasets (Yang *et al.*, 2023). Second, pathology-specific representation learning consistently outperforms purely generic initialization strategies, motivating the broader objective of universal pathology representation learning (Chen *et al.*, 2024; Vorontsov *et al.*, 2024; Wang *et al.*, 2022; H. Xu *et al.*, 2024). Third, the field is clearly moving toward foundation-model and multimodal pathology systems, but the current evidence implemented in this study corresponds to an image-only classification stage, rather than a completed WSI-scale self-supervised or report-generation system. Therefore, the literature review positions the present study as an implemented colorectal histopathology classification framework developed in the context of and motivated by the broader transition toward pathology foundation models.

Based on the above literature, the present study does not claim to introduce a new whole-slide pathology foundation model or a multimodal report-generation system. Instead, it contributes a regularized supervised image-level framework for colorectal tissue classification using the PathMNIST benchmark. Foundation-model and multimodal pathology studies are discussed only to motivate the broader research direction and future extension of the work. The implemented contribution remains focused on robust classification, class-wise analysis, benchmark comparison, and interpretability for nine-class colorectal histopathology tissue recognition.

This distinction is important because the foundation-model literature is included only to explain the broader research context and the future direction of computational pathology. No foundation model, multimodal model, or whole-slide report-generation model is implemented in the current experiments. Accordingly, all claims and conclusions in the present study are restricted to supervised image-level classification on PathMNIST.

The representative studies reviewed in the literature are summarized in Table 1 to provide a concise comparison of computational pathology, pathology foundation models, and pathology report-generation approaches relevant to the present work.

3. Problem statement

Automated colorectal histopathology tissue classification is challenging because different tissue categories may exhibit high intra-class variation and inter-class morphological similarity. In PathMNIST, tissue types such as cancer-associated stroma, smooth muscle, debris, normal colon mucosa, and colorectal adenocarcinoma epithelium can

contain visually overlapping structural patterns, making accurate multi-class classification difficult. A reliable model should therefore not only achieve high overall accuracy but also maintain balanced precision, recall, and F1-score across all tissue categories.

The problem addressed in this study is the development and evaluation of a supervised image-level deep learning framework that can classify nine colorectal histopathology tissue types accurately and robustly using the PathMNIST benchmark. The study specifically investigates whether a regularized ResNet-50-based transfer-learning pipeline, enhanced with augmentation, MixUp, CutMix, label smoothing, AdamW optimization, and interpretability analysis, can improve classification performance compared with selected baseline models. The scope of the present work is limited to benchmark-level image classification and does not include whole-slide self-supervised pretraining or multimodal pathology report generation.

The problem statement also recognizes that reliance on a single benchmark dataset limits claims about clinical generalizability. External validation using multi-center whole-slide cohorts, stain-normalized datasets, and scanner-diverse images is required before the model can be considered for clinical translation.

4. Proposed methodology

The proposed PathRes_FM framework is designed as a supervised image-level classification pipeline for colorectal histopathology tissue recognition. The framework uses the PathMNIST benchmark dataset and classifies images into nine tissue categories. The implemented model does not perform whole-slide self-supervised pretraining or multimodal report generation; therefore, the methodology is presented strictly as a regularized ResNet-50-based classification framework supported by the provided implementation and experimental results.

The dataset is loaded through the MedMNIST API under the parameter `data_flag = "pathmnist."` The number of target classes is directly determined from the dictionary used for labels in the dataset's metadata. The histopathological regions include nine categories of colorectal tissue: adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium. First, image transforms for the central processing unit-side preprocessing to convert images into tensor and float formats are implemented. Afterward, graphics processing unit-side transforms are separated into training and evaluation pipelines. For training, each image is resized to 224×224 pixels, flipped randomly along the horizontal and vertical directions, randomly rotated by up to 15° ,

Table 1. Comparative review of representative studies in computational pathology, pathology foundation models, and pathology report generation

Study	Task/scope	Core methodology	Main strength	Main limitation	Relevance to present work
Coudray <i>et al.</i> (2018)	Histopathology-based cancer classification and mutation prediction	CNN-based supervised learning on pathology images	Demonstrated clinical information can be learned directly from histology	Task-specific and not designed for transferable representation learning	Establishes the importance of image-based pathology prediction
Campanella <i>et al.</i> (2019)	Whole-slide clinical-grade pathology	Weakly supervised deep learning on WSIs	Showed the feasibility of clinical-grade WSI analysis with weak labels	Focused on specific diagnostic tasks rather than universal encoders	Important background for slide-level pathology AI
Lu <i>et al.</i> (2021)	Whole-slide pathology	Data-efficient and weakly supervised MIL	Reduced annotation burden and improved WSI learning practicality	Still downstream-task-specific	Relevant to low-annotation pathology settings
Shao <i>et al.</i> (2021); TransMIL	WSI classification	Transformer-based correlated MIL	Captures long-range contextual dependency in WSIs	Requires slide-level pipeline design beyond simple patch classifiers	Important for contextual pathology learning
Yang <i>et al.</i> (2023); MedMNIST v2/ PathMNIST	Standardized biomedical benchmark, including colorectal tissue classification	Lightweight benchmark dataset derived from colorectal histology patches	Reproducible and accessible evaluation benchmark	Patch-level simplification cannot represent full WSI complexity	Directly relevant to the present dataset setting
Wang <i>et al.</i> (2022); CTransPath	Histopathology representation learning	Transformer-based unsupervised contrastive pretraining	Strong pathology-specific pretrained features	Not itself a general-purpose WSI foundation model	Supports the value of pathology-specific feature learning
Chen <i>et al.</i> (2024); UNI	General-purpose pathology representation learning	Self-supervised pretraining on >100 million images from >100,000 WSIs	Strong transfer across many pathology tasks	Requires very large-scale data resources	Provides direct motivation for pathology-specific representation learning
Lu, Chen, Williamson, Chen, Liang, <i>et al.</i> (2024); CONCH	Computational pathology vision-language learning	Contrastive image-text pretraining	Brings semantic biomedical text into pathology representation learning	Moves beyond the image-only scope of the current implementation	Important broader foundation-model context
H. Xu <i>et al.</i> (2024); Prov-GigaPath	Whole-slide pathology foundation modeling	Large-scale pretraining with explicit whole-slide modeling	Strong performance across many digital pathology tasks	Needs real-world WSI-scale infrastructure	Key motivation for WSI-level representation learning
Vorontsov <i>et al.</i> (2024); Virchow	Clinical-grade pathology foundation model	Large pretrained pathology encoder	Strong pan-cancer and rare-cancer performance	Heavy pretraining requirement	Relevant to data-scarce pathology motivation
Lu, Chen, Williamson, Chen, Zhao, <i>et al.</i> (2024); PathChat	Multimodal pathology assistance	Pathology vision encoder + LLM	Interactive pathology reasoning and question answering	Not part of current image-only implementation	Relevant as a longer-term extension
Ding <i>et al.</i> (2025); TITAN	Multimodal whole-slide foundation model	WSI-report-caption multimodal pretraining	Supports diagnosis, retrieval, prognosis, and report-related tasks	Requires multimodal slide-text resources	Motivates multimodal future scope

(Cont'd...)

Table 1. (Continued)

Study	Task/scope	Core methodology	Main strength	Main limitation	Relevance to present work
Y. Xu <i>et al.</i> (2025); mSTAR	Knowledge-enhanced pathology FM	WSI + reports + molecular context	Incorporates richer pathology supervision	Beyond the scope of current code/results	Relevant to future multimodal extensions
Tran <i>et al.</i> (2025); HistoGPT	Pathology report generation	Vision-language report generation from gigapixel slides	Directly addresses pathology reporting workload	Specialized report-generation setting	Relevant to future work, not present implementation
Hu <i>et al.</i> (2025)	WSI report generation	Knowledge retrieval + multi-level region selection	Improves report faithfulness through salient-region focus	Report generation still less mature than classification	Useful for future report-generation framing
Lee <i>et al.</i> (2025)	Foundation-model benchmarking	Comparison of adaptation strategies across datasets	Shows parameter-efficient adaptation is practical	Benchmarking does not itself solve the deployment shift	Important for discussing data-efficient adaptation
Neidlinger <i>et al.</i> (2025)	Large benchmark of pathology FMs	Multi-model, multi-task external benchmarking	Strong evidence on cross-model generalization and leakage-aware evaluation	Benchmark study rather than a new deployment model	Supports the rationale for the present study
Nicke <i>et al.</i> (2025)	Supervised pathology foundation modeling	Tissue concepts supervised FM	Reduces pretraining burden while preserving generalization	Still dependent on curated development setting	Relevant alternative to large self-supervised pipelines
Pocevičiūtė <i>et al.</i> (2025)	OOD detection in digital pathology	Robustness analysis for pathology FMs	Highlights deployment risks under distribution shift	Does not directly improve classification accuracy	Important for clinical robustness discussion
Campanella <i>et al.</i> (2025)	Real-world FM deployment	Prospective biomarker deployment study	Demonstrates practical translation of pathology FM ideas	Deployment-specific and not colorectal tissue classification	Important evidence for eventual clinical relevance

Abbreviations: AI: Artificial intelligence; CNN: Convolutional neural network; FM: Foundation model; LLM: Large language model; MIL: Multiple instance learning; OOD: Out-of-distribution; WSI: Whole slide image.

and subjected to photometric augmentation (brightness, saturation, contrast, and hue adjustments) using a predefined parameter space, and finally normalized using mean [0.7021, 0.5462, 0.6964] and standard deviation [0.2350, 0.2820, 0.2165]. For evaluation, the process is restricted to resizing and normalizing each image, thereby establishing a deterministic inference path.

The PathMNIST benchmark officially contains 89,996 training images, 10,004 validation images, and 7,180 test images. In the implemented pipeline used for the reported results, the official training split was used for model learning, and the official test split was used only for final performance evaluation. The official validation split is reported for dataset completeness but was not used for checkpoint selection in the present implementation. Therefore, this study avoids the phrase “best validation accuracy” and clarifies that no validation-based model

selection was performed on the unused validation split.

For robust and stable training, the code further implements sample-level regularization with MixUp and CutMix schemes. More specifically, MixUp(alpha = 0.2) and CutMix(alpha = 1.0) are initialized in the implementation. Next, the regularization effect of these two sample augmentation schemes is applied using RandomChoice, whereby each training batch is regularized by one of these two selected data mixing strategies. Notably, this feature is important because many colorectal tissues possess structures visually overlapping with one another. Thus, mixing training samples can allow decision boundaries to become smoother and alleviate issues associated with overfitting to certain class characteristics. Hence, the PathRes_FM pipeline proposed here is not simply a transfer-learning classifier; it is a pathology classification pipeline that involves data augmentation, label smoothing,

and sample mixing to enhance robustness within one training mechanism.

At the architectural level, the model uses ResNet-50 initialized with ImageNet1K_V2 pretrained weights as the visual backbone. The original final fully connected layer is replaced with a task-specific head composed of a dropout layer with probability 0.3 followed by a linear layer that maps the extracted feature vector to the nine output classes. This means that the backbone performs hierarchical visual feature extraction from colorectal histopathology patches, while the modified terminal classifier adapts the representation to the target tissue-recognition task. Since this exact structure is used in the implemented model, this study presents it as the proposed PathRes_FM architecture.

Training was performed with a batch size of 128 for 35 epochs. The loss function was cross-entropy with label smoothing = 0.1, the optimizer was AdamW with a learning rate of 1×10^{-3} and weight decay of 1×10^{-2} , and the learning-rate schedule was controlled using OneCycleLR across the full training duration. Mixed-precision training with autocast and GradScaler was used to improve computational efficiency. Because the official validation split was not loaded in the implemented pipeline, the present study does not claim validation-based checkpoint selection. The final reported metrics correspond to evaluation on the official PathMNIST test split after training under the stated configuration.

After training, the framework performs a comprehensive quantitative and qualitative evaluation. The implemented evaluation pipeline computes overall accuracy, multi-class area under the receiver operating characteristic curve (AUC-ROC) under a one-vs-rest setting, weighted F1-score, weighted precision, weighted recall, and class-wise classification statistics. It also supports confusion-matrix analysis, ROC visualization, precision-recall curves, average precision analysis, calibration assessment, t-distributed stochastic neighbor embedding (t-SNE) feature projection, and gradient-based localization through a Grad-CAM-style interpretability module. In the explainability stage, the model extracts activations and gradients from the final convolutional layer and combines them to form class-discriminative heatmaps over the pathology patch, thereby allowing visual inspection of the regions that contributed most strongly to the prediction. This makes PathRes_FM both a classification model and an interpretable tissue-recognition framework.

Accordingly, the methodology of the present study should be described as a regularized ResNet-50-based colorectal histopathology classification framework on PathMNIST, implemented under the name PathRes_FM, with official PathMNIST train/test usage corresponding

to 89,996 training samples and 7,180 test samples in the implemented pipeline. The official validation split of 10,004 images belongs to the benchmark dataset but is not loaded in the current implementation and therefore should not be presented as part of the implemented training pipeline unless the code is explicitly revised.

4.1. Architecture of the proposed PathRes_FM framework

The implemented architecture can be summarized in the following sequence:

- (i) Input layer: PathMNIST colorectal histopathology image patch.
- (ii) Preprocessing layer: Tensor conversion, resizing to 224×224 , and normalization.
- (iii) Augmentation layer: Horizontal flip, vertical flip, rotation, and color jitter.
- (iv) Regularization layer: Random MixUp or CutMix.
- (v) Feature extraction layer: Pretrained ResNet-50 backbone.
- (vi) Classification layer: Dropout (0.3) + linear projection.
- (vii) Prediction layer: Nine-class logits and softmax probabilities.
- (viii) Optimization layer: Label-smoothed cross-entropy + AdamW + OneCycleLR.
- (ix) Interpretation layer: Confusion matrix, ROC, precision-recall, calibration, t-SNE, and Grad-CAM.

The Grad-CAM-based visual explanation workflow used to interpret the proposed PathRes_FM predictions is illustrated in Figure 1.

4.2. Mathematical model/equations

Let the input colorectal histopathology image be denoted as $x \in \mathbb{R}^{H \times W \times 3}$. After preprocessing, the resized and normalized image is $\hat{x} = \mathcal{N}(\mathcal{R}(x))$, where $\mathcal{R}(\cdot)$ denotes resizing to 224×224 and $\mathcal{N}(\cdot)$ denotes channel-wise normalization using the code-defined mean and standard deviation.

For training, stochastic augmentation is applied:

$$x^{aug} = \mathcal{A}(\hat{x}), \quad (1)$$

where $\mathcal{A}(\cdot)$ includes random horizontal flip, vertical flip, rotation, and color jitter.

For MixUp, given two samples (x_i, y_i) and (x_j, y_j) , the transformed sample is $\tilde{x} = \lambda x_i + (1 - \lambda)x_j$ and $\tilde{y} = \lambda y_i + (1 - \lambda)y_j$, where $\lambda \sim \text{Beta}(\alpha, \alpha)$ and the code uses $\alpha = 0.2$.

For CutMix, a binary mask M replaces a region of one image with another:

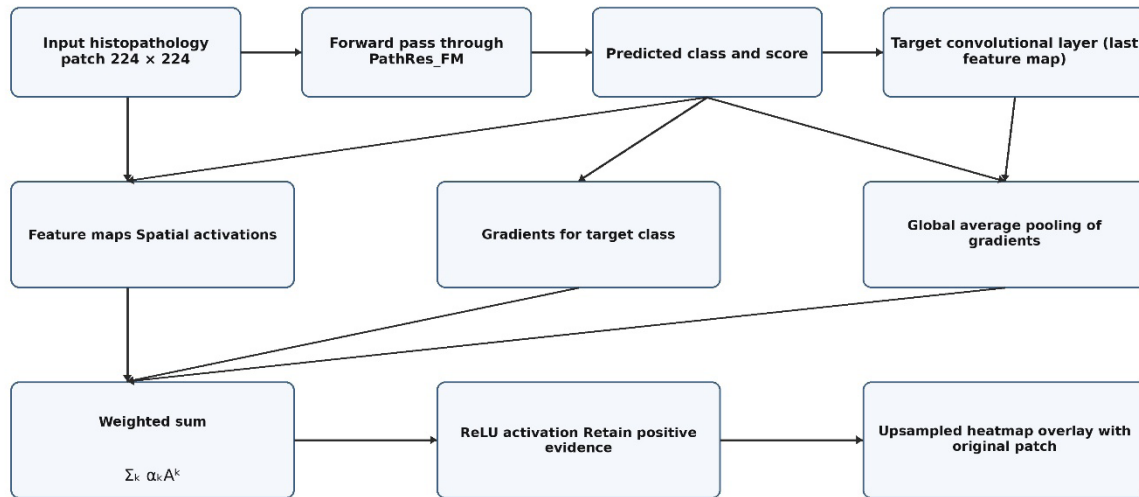


Figure 1. Gradient-weighted class activation mapping-based visual explanation mechanism in the proposed PathRes_FM model. The input colorectal histopathology patch is first passed through the trained PathRes_FM classifier to obtain the predicted class score. Gradients of the target class with respect to the final convolutional feature maps are then computed and globally pooled to obtain channel-wise importance weights. These weights are used to form a weighted combination of feature maps, followed by rectified linear unit (ReLU) activation and upsampling, producing a heatmap that highlights the most discriminative image regions contributing to the final tissue prediction.

$$\tilde{x} = M \odot x_i + (1 - M) \odot x_j, \quad \tilde{y} = \lambda y_i + (1 - \lambda) y_j,$$

where λ is determined by the retained image area and the code uses $\alpha = 1.0$.

Let $f_\theta(\cdot)$ denote the pretrained ResNet-50 backbone. The deep representation is $h = f_\theta(\tilde{x})$

The classification head applies dropout and a linear mapping:

$$z = W \cdot \text{Dropout}(h) + b, \quad (2)$$

where $z \in \mathbb{R}^K$ and $K=9$ is the number of tissue classes. The posterior probability for class k is obtained through softmax:

$$p_k = \frac{e^{z_k}}{\sum_{c=1}^K e^{z_c}}, k=1, 2, \dots, K. \quad (3)$$

The training procedure uses label smoothing with $\varepsilon = 0.1$, so the smoothed target is:

$$y_k^{LS} = (1 - \varepsilon) y_k + \frac{\varepsilon}{K}. \quad (4)$$

The corresponding multi-class cross-entropy loss is:

$$\mathcal{L}_{CE} = -\sum_{k=1}^K y_k^{LS} \log(p_k). \quad (5)$$

Since MixUp and CutMix generate soft targets, this

same formulation remains valid for mixed labels.

The model parameters are updated using AdamW:

$$\theta_{t+1} = \theta_t - \eta_t \cdot \text{AdamW}(\nabla_{\theta} \mathcal{L}_{CE}), \quad (6)$$

where η_t is the time-varying learning rate defined by the OneCycleLR scheduler over 35 epochs.

The final predicted label is:

$$\hat{y} = \arg \max_k p_k. \quad (7)$$

Overall accuracy is computed as:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(\hat{y}_i = y_i). \quad (8)$$

For each class k , precision, recall, and F1-score are computed as:

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \quad (9)$$

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k}, \quad (10)$$

$$F1_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \quad (11)$$

For visual explainability, the Grad-CAM-style heatmap for class c is formed from the final convolutional activations A^m and their pooled gradients:

$$\alpha_c^m = \frac{1}{Z} \sum_i \sum_j \frac{\partial z_c}{\partial A_{ij}^m}, \quad (12)$$

$$L_c^{\text{Grad-CAM}} = \text{ReLU} \left(\sum_m \alpha_c^m A^m \right). \quad (13)$$

This heatmap highlights the discriminative image regions that contributed most to the model's pathology decision.

The main hyperparameter settings used in the implemented PathRes_FM pipeline are summarized in Table 2, and the mathematical formulations supporting classification and explainability are summarized in Table 3.

4.3. Experimental setup

The experimental evaluation of the proposed PathRes_FM framework was conducted as a supervised nine-class colorectal histopathology tissue classification task using the PathMNIST benchmark. According to the uploaded implementation, the dataset was loaded through the MedMNIST interface using the PathMNIST configuration, and the model was trained and evaluated on the train and test splits used in the implemented pipeline. The target classes in the current study are adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium, confirming that the present study addresses colorectal tissue-level image classification rather than full whole-slide pathology modeling.

All pathology images were first converted to floating-point tensor representation and resized to 224×224 pixels before being passed to the network. The preprocessing stage further applied channel-wise normalization using the dataset statistics explicitly defined in the code, with mean values of [0.7021, 0.5462, 0.6964] and standard deviation values of [0.2350, 0.2820, 0.2165]. For training, the model used a richer augmentation pipeline composed of random horizontal flip, random vertical flip, random rotation (15°), and color jitter. To further improve generalization, the training pipeline incorporated MixUp with $\alpha = 0.2$ and CutMix with $\alpha = 1.0$ through a random-choice mechanism, whereas test-time preprocessing remained deterministic with only resizing and normalization.

The classification architecture was implemented using a ResNet-50 backbone initialized with ImageNet1K_V2 pretrained weights. The original final classifier was replaced with a task-adapted head consisting of a dropout layer ($p = 0.3$) followed by a fully connected layer for nine-class tissue prediction. The implementation identifies this final configuration as the “PathRes_FM PIPELINE: FINAL PERFORMANCE VERSION,” and this name should therefore be used consistently throughout the study.

Table 2. Hyperparameter configuration of the implemented PathRes_FM pipeline

Component	Setting/value
Model name	PathRes_FM
Dataset	PathMNIST (MedMNIST)
Implemented splits in code	Train and test
Official PathMNIST split sizes	Train: 89,996; validation: 10,004; test: 7,180
Input image size	224×224
Number of classes	9
Backbone network	ResNet-50
Pretrained weights	ImageNet1K_V2
Classification head	Dropout ($p = 0.3$) + fully connected layer
Training batch size	128
Number of epochs	35
Loss function	Cross-entropy loss
Label smoothing	0.1
Optimizer	AdamW
Initial learning rate	1×10^{-3}
Weight decay	1×10^{-2}
Learning-rate scheduler	OneCycleLR
Mixed precision	Enabled (autocast + GradScaler)
Data augmentation	Horizontal flip, vertical flip, rotation (15°), color jitter
Mixed-sample regularization	MixUp ($\alpha = 0.2$), CutMix ($\alpha = 1.0$), RandomChoice
Normalization mean	[0.7021, 0.5462, 0.6964]
Normalization standard deviation	[0.2350, 0.2820, 0.2165]
Model checkpoint selection	No validation-based checkpoint selection; official test split used only for final evaluation
Explainability module	Grad-CAM-style visualization
Evaluation outputs	Accuracy, AUC-ROC, weighted F1, weighted precision, weighted recall, confusion matrix, ROC, precision-recall, calibration, t-SNE

Abbreviations: AUC: Area under the curve; Grad-CAM: Gradient-weighted class activation mapping; ROC: Receiver operating characteristic; t-SNE: t-distributed stochastic neighbor embedding.

Model training was carried out with a batch size of 128

Table 3. Mathematical formulation of the PathRes_FM classification and explainability model

Module	Mathematical formulation	Description
Preprocessing	$\hat{x} = \mathcal{N}(\mathcal{R}(x))$	Input image x is resized to 224×224 and normalized
Data augmentation	$x^{aug} = \mathcal{A}(\hat{x})$	Training-time stochastic augmentation using flips, rotation, and color jitter
MixUp	$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$	Two samples are linearly combined using $\lambda \sim \text{Beta}(\alpha, \alpha)$, with $\alpha = 0.2$
MixUp labels	$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$	Target labels are mixed in the same ratio as the images
CutMix	$\tilde{x} = M \odot x_i + (1 - M) \odot x_j$	A spatial region from one image is inserted into another using a binary mask M
Feature extraction	$h = f_{\theta}(\tilde{x})$	ResNet-50 backbone extracts deep visual features
Classification head	$z = W \cdot \text{Dropout}(h) + b$	Dropout and linear mapping produce class logits
Softmax output	$p_k = \frac{e^{z_k}}{\sum_{c=1}^K e^{z_c}}$	Probability of class k , where $K = 9$
Label smoothing	$y_k^{LS} = (1 - \varepsilon) y_k + \frac{\varepsilon}{K}$	Smoothed target distribution with $\varepsilon = 0.1$
Loss function	$\mathcal{L}_{CE} = -\sum_{k=1}^K y_k^{LS} \log(p_k)$	Multi-class cross-entropy loss used in training
Optimization	$\theta_{t+1} = \theta_t - \eta_t \cdot \text{AdamW}(\nabla_{\theta} \mathcal{L}_{CE})$	Parameter update using AdamW with OneCycleLR-controlled learning rate
Prediction rule	$\hat{y} = \text{argmax}_k p_k$	Final predicted class is the one with maximum probability
Accuracy	$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i = y_i)$	Fraction of correctly classified samples
Precision	$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}$	Class-wise positive predictive value
Recall	$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k}$	Class-wise sensitivity
F1-score	$F1_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}$	Harmonic mean of precision and recall
Grad-CAM weights	$\alpha_c^m = \frac{1}{Z} \sum_i \sum_j \frac{\partial z_c}{\partial A_{ij}^m}$	Channel importance weights from globally pooled gradients
Grad-CAM heatmap	$L_c^{\text{Grad-CAM}} = \text{ReLU}(\sum_m \alpha_c^m A^m)$	Final localization map for the target class

Abbreviations: Grad-CAM: Gradient-weighted class activation mapping; ReLU: Rectified linear unit.

for 35 epochs. Optimization used AdamW with an initial learning rate of 1×10^{-3} and weight decay of 1×10^{-2} . The loss function was cross-entropy loss with label smoothing of 0.1, and the learning-rate schedule was controlled using OneCycleLR. The experiments were executed in Python 3 using PyTorch, torchvision, and the MedMNIST API on a compute unified device architecture (CUDA)-enabled environment with an NVIDIA T4 GPU. Mixed precision

was enabled using autocast and GradScaler.

For reproducibility, the exact package versions, CUDA version, operating system, and random seed should be reported when the experiments are reproduced. If the experiments are rerun, fixed seeds should be applied to Python, NumPy, and PyTorch to reduce random variation. Table 4 summarizes the experimental configuration, while Table 5 details the preprocessing, augmentation, and

optimization settings used in the pipeline.

Table 4 provides an overview of the implemented experimental configuration of the PathRes_FM framework.

Table 4. Experimental configuration of the implemented PathRes_FM framework

Parameter	Configuration
Model name	PathRes_FM
Task	Multi-class colorectal histopathology tissue classification
Dataset	PathMNIST (MedMNIST)
Implemented dataset splits in code	Train and test
Official benchmark split sizes	Train: 89,996; validation: 10,004; test: 7,180
Number of classes	9
Tissue classes	Adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium
Input image size	224 × 224
Backbone architecture	ResNet-50
Pretrained initialization	ImageNet1K_V2
Classification head	Dropout ($p = 0.3$) + fully connected layer
Batch size	128
Number of training epochs	35
Loss function	Cross-entropy loss
Label smoothing	0.1
Optimizer	AdamW
Initial learning rate	1×10^{-3}
Weight decay	1×10^{-2}
Learning-rate scheduler	OneCycleLR
Mixed precision	Enabled (autocast + GradScaler)
Device selection	CUDA if available, otherwise CPU
cuDNN setting	torch.backends.cudnn.benchmark = True
PyTorch CUDA allocation setting	expandable_segments: True
Checkpoint selection	No validation-based checkpoint selection; final test evaluation on official PathMNIST test split
Evaluation metrics	Accuracy, AUC–ROC, weighted F1-score, weighted precision, and weighted recall
Additional analyses	Confusion matrix, ROC curves, precision–recall curves, average precision, calibration, t-SNE, Grad-CAM
Hardware platform	NVIDIA T4 GPU/CUDA-enabled environment
Programming language	Python 3
Deep learning framework	PyTorch with torchvision
Dataset API	MedMNIST
Backbone library	torchvision

Abbreviations: AUC: Area under the curve; API: Application programming interface; CPU: Central processing unit; CUDA: Compute unified device architecture; Grad-CAM: Gradient-weighted class activation mapping; GPU: Graphics processing unit; ROC: Receiver operating characteristic; t-SNE: t-distributed stochastic neighbor embedding.

Table 5. Data preprocessing, augmentation, and optimization settings used in PathRes_FM

Stage	Operation/setting	Value/description
Preprocessing	Tensor conversion	ToImage() + ToDtype(torch.float32, scale=True)
Preprocessing	Image resizing	224 × 224
Preprocessing	Normalization mean value	[0.7021, 0.5462, 0.6964]
Preprocessing	Normalization standard deviation value	[0.2350, 0.2820, 0.2165]
Training augmentation	Random horizontal flip	Enabled
Training augmentation	Random vertical flip	Enabled
Training augmentation	Random rotation	15°
Training augmentation	Color jitter	Enabled
Mixed-sample regularization	MixUp	$\alpha = 0.2$
Mixed-sample regularization	CutMix	$\alpha = 1.0$
Mixed-sample strategy	Application mode	RandomChoice
Training loss	Cross-Entropy Loss	With label smoothing
Label smoothing factor	ϵ	0.1
Optimization	Optimizer	AdamW
Optimization	Learning rate	1×10^{-3}
Optimization	Weight decay	1×10^{-2}
Learning policy	Scheduler	OneCycleLR
Training precision	Mixed precision	autocast + GradScaler
Model selection	Checkpoint rule	Final trained checkpoint; test split used only for final evaluation
Explainability	Heatmap generation	Grad-CAM-style method
Feature-space analysis	Projection	t-SNE
Calibration analysis	Reliability assessment	Enabled in evaluation

Abbreviations: Grad-CAM: Gradient-weighted class activation mapping; t-SNE: t-distributed stochastic neighbor embedding.

The framework was evaluated as a supervised image-level colorectal histopathology classifier on the PathMNIST dataset. It uses a ResNet-50 backbone initialized with ImageNet pretrained weights and a dropout-regularized fully connected head for nine-class tissue discrimination. Training was conducted for 35 epochs with a batch size of 128. The final trained checkpoint was used for reporting, and the official test split was reserved exclusively for final evaluation. The evaluation includes global classification metrics and additional analyses of discrimination, calibration, feature-space organization, and visual interpretability.

Table 5 details the data preprocessing, augmentation, and optimization settings used in the PathRes_FM pipeline. Images were converted into PyTorch tensors, resized to 224 × 224 pixels, and normalized using dataset-

specific channel-wise mean and standard deviation values. The training augmentation pipeline included random horizontal flipping, random vertical flipping, random rotation, color jitter, MixUp, and CutMix regularization. Model optimization used AdamW with label-smoothed cross-entropy loss, OneCycleLR scheduling, mixed-precision computation, and CUDA acceleration when available.

5. Results and discussion

The experimental evaluation on the PathMNIST benchmark demonstrates that the proposed PathRes_FM framework achieves strong multi-class classification performance across nine colorectal tissue categories. The overall performance metrics are summarized in Table 6. These metrics describe not only overall classification accuracy but also the balance between precision and recall.

High values for these measures indicate that the classifier distinguishes tissue categories effectively while limiting both false-positive and false-negative predictions. The AUC–ROC value of 0.9890 further confirms the strong discriminative capability of the learned features across the nine colorectal tissue classes.

Table 6. Overall performance metrics of the proposed PathRes_FM framework

Metric	Value
Overall accuracy	94.04%
Overall AUC–ROC	0.9890
Weighted F1-score	0.9399
Weighted precision	0.9410
Weighted recall	0.9404

Abbreviations: AUC: Area under the curve; ROC: Receiver operating characteristic.

Based on the weighted performance metrics, the proposed PathRes_FM model demonstrates balanced classification performance on the PathMNIST test set. The model achieved an overall accuracy of 94.04%, an AUC–ROC of 0.9890, a weighted precision of 0.9410, a weighted recall of 0.9404, and a weighted F1-score of 0.9399. The closeness of weighted precision, recall, and F1-score indicates that the model maintains a stable balance between false-positive and false-negative predictions across the nine colorectal tissue categories.

Table 7 presents the class-wise classification results. Among all categories, lymphocytes achieved the highest precision of 0.9967 and the highest F1-score of 0.9808,

indicating highly reliable discrimination of this class. The adipose class also achieved excellent performance, with a precision of 0.9700, a recall of 0.9910, and an F1-score of 0.9804, indicating low false-positive predictions while maintaining high sensitivity. Colorectal adenocarcinoma epithelium, one of the clinically important categories in the dataset, achieved a precision of 0.9439, a recall of 0.9830, and an F1-score of 0.9631. This suggests that the framework can discriminate cancer-related epithelial regions from non-epithelial or normal tissue patterns. Mucus and normal colon mucosa also showed strong classification quality, with F1-scores of 0.9579 and 0.9521, respectively.

As expected, cancer-associated stroma, smooth muscle, and debris were the most challenging categories. Cancer-associated stroma produced the lowest F1-score of 0.7854 and recall of 0.7435, indicating that this class remained the most difficult for the model to classify consistently. Smooth muscle obtained a precision of 0.8155, a recall of 0.8514, and an F1-score of 0.8331. Debris achieved a precision of 0.9000 and a recall of 0.8496, suggesting that some debris samples were still misclassified. These results indicate that the remaining difficulties are concentrated in tissue types that share morphologically similar patterns with surrounding categories.

The support values in Table 7 further help explain class-wise behavior. Adipose had the largest support of 1,338 samples and achieved an F1-score above 0.98, while colorectal adenocarcinoma epithelium had 1,233 samples and achieved an F1-score above 0.96. However, lymphocytes achieved an F1-score of 0.9808 with only 634 samples, showing that a higher sample count alone does not guarantee better performance. Conversely, cancer-associated stroma had 421 samples but remained the

Table 7. Class-wise performance analysis of PathRes_FM on colorectal histopathology images

Tissue class	Precision	Recall	F1-score	No. of samples
Adipose	0.9700	0.9910	0.9804	1,338
Background	0.9177	1.0000	0.9571	847
Debris	0.9000	0.8496	0.8741	339
Lymphocytes	0.9967	0.9653	0.9808	634
Mucus	0.9958	0.9227	0.9579	1,035
Smooth muscle	0.8155	0.8514	0.8331	592
Normal colon mucosa	0.9666	0.9379	0.9521	741
Cancer-associated stroma	0.8324	0.7435	0.7854	421
Colorectal adenocarcinoma epithelium	0.9439	0.9830	0.9631	1,233

weakest class. This indicates that tissue morphology and inter-class similarity may influence performance more strongly than support size alone.

As shown in Table 8, PathRes_FM achieved higher accuracy than all listed baselines and higher AUC-ROC than Auto-Sklearn and ResNet-18, while matching the ResNet-50 AUC-ROC. The proposed model achieved an accuracy of 94.04% and an AUC-ROC of 0.9890, compared with 71.6% and 0.934 for Auto-Sklearn, 90.7% and 0.983 for ResNet-18, and 89.2% and 0.989 for ResNet-50. These results indicate that the complete regularized PathRes_FM pipeline improves classification accuracy over the selected baseline configurations for PathMNIST classification.

For a fair interpretation of the comparative results, all reported baseline performances are discussed under the PathMNIST benchmark setting and evaluated using the same official test split. The proposed PathRes_FM model uses the same nine-class colorectal tissue classification task but improves the ResNet-50 transfer-learning pipeline through dropout regularization, data augmentation, MixUp, CutMix, label smoothing, AdamW optimization, and OneCycleLR scheduling. The comparison therefore highlights the performance gain obtained by the complete regularized training strategy rather than by the backbone architecture alone.

To make the comparison clearer, the baseline models are interpreted under the same PathMNIST task definition, input-size setting, and official test-split evaluation protocol. However, because the reported Auto-Sklearn, ResNet-18, and ResNet-50 baselines may originate from different published or benchmark configurations, the comparison is presented as indicative rather than as a definitive state-of-the-art comparison. Future work should include direct retraining of all baselines under identical preprocessing, augmentation, optimization, epoch, and seed settings, and should also compare with pathology-specific convolutional neural network, transformer, and foundation-model

feature extractors.

PathRes_FM improved accuracy by 22.44 percentage points over Auto-Sklearn, 3.34 percentage points over ResNet-18, and 4.84 percentage points over ResNet-50. Its AUC-ROC was higher by 0.0550 than Auto-Sklearn and by 0.0060 than ResNet-18, while matching ResNet-50 (difference = 0.0000).

Figure 2 presents the global performance of the proposed model and is consistent with the quantitative results in Table 6. It shows strong performance in accuracy, AUC-ROC, and F1-score. Accuracy alone may be insufficient in imbalanced medical-image classification settings; therefore, the use of multiple metrics provides a more reliable view of the model's predictive behavior.

Figure 3 presents the multi-class ROC curves for the proposed classifier. Consistent with the high AUC-ROC reported in Table 6, the class-specific ROC curves show strong separation ability, indicating that the model can obtain high true-positive rates while maintaining relatively low false-positive rates across tissue categories.

Figure 4 presents the precision-recall curves for colorectal tissue classification. These curves are particularly useful when class frequencies differ across tissue categories because they directly show the trade-off between precision and recall. The curves complement the ROC analysis and provide additional evidence that the model maintains strong retrieval and classification behavior across most tissue classes.

Figure 5 presents the confusion-matrix heatmap and provides a class-level view of prediction errors. The lower class-wise results for cancer-associated stroma, smooth muscle, and debris are consistent with the confusion behavior shown in this visualization. The heatmap helps identify which tissue categories are most frequently confused and supports the interpretation of the class-wise results reported in Table 7.

Table 8. Comparative benchmarking of PathRes_FM against baseline models

Model architecture	Accuracy (%)	AUC-ROC	Evaluation protocol
Auto-Sklearn (baseline)	71.6%	0.934	PathMNIST test split
ResNet-18 (28)	90.7%	0.983	PathMNIST test split
ResNet-50 (224)	89.2	0.989	PathMNIST test split
PathRes_FM	94.04	0.9890	PathMNIST test split

Source for baseline benchmark values: MedMNIST v2 official PathMNIST benchmark, <https://medmnist.com/v2>. PathRes_FM values are from the present study.

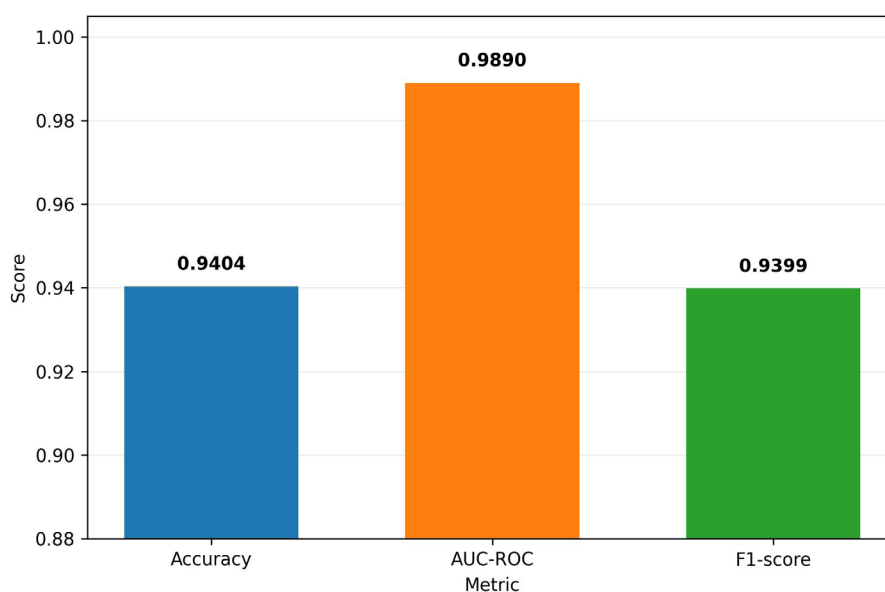


Figure 2. Global performance summary of the proposed PathRes_FM model
Abbreviations: AUC: Area under the curve; ROC: Receiver operating characteristic.

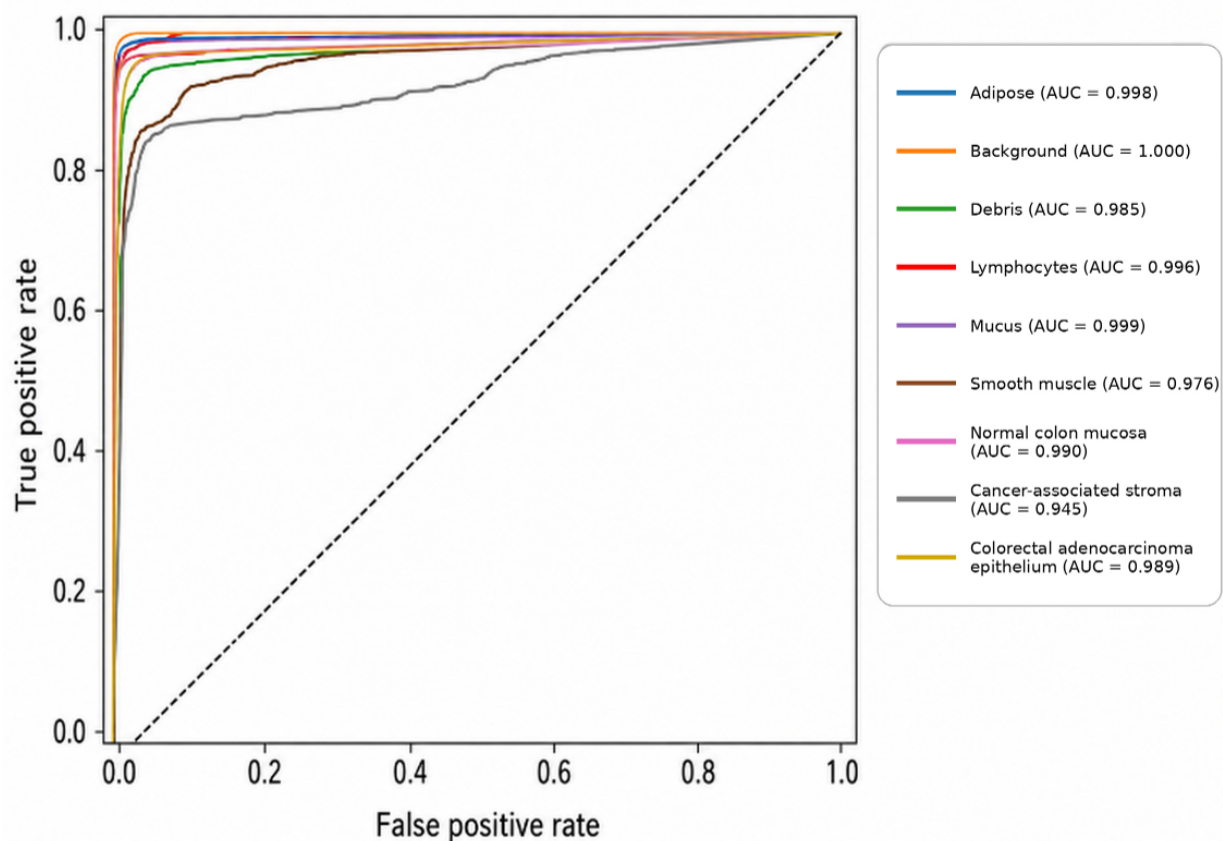


Figure 3. Multi-class receiver operating characteristic curves of the proposed PathRes_FM framework
Abbreviation: AUC: Area under the curve.

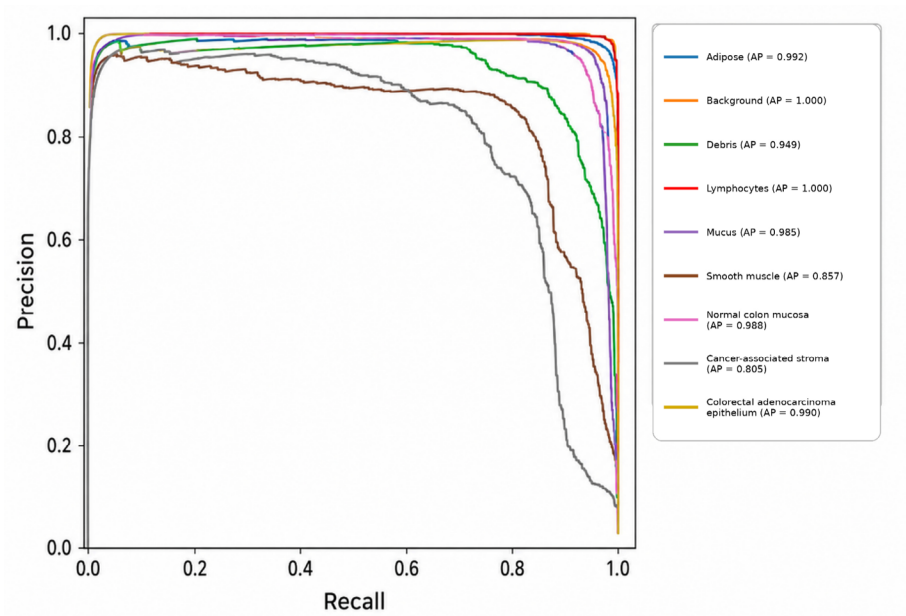


Figure 4. Precision–recall curves for colorectal tissue classification
Abbreviation: AP: Average precision.

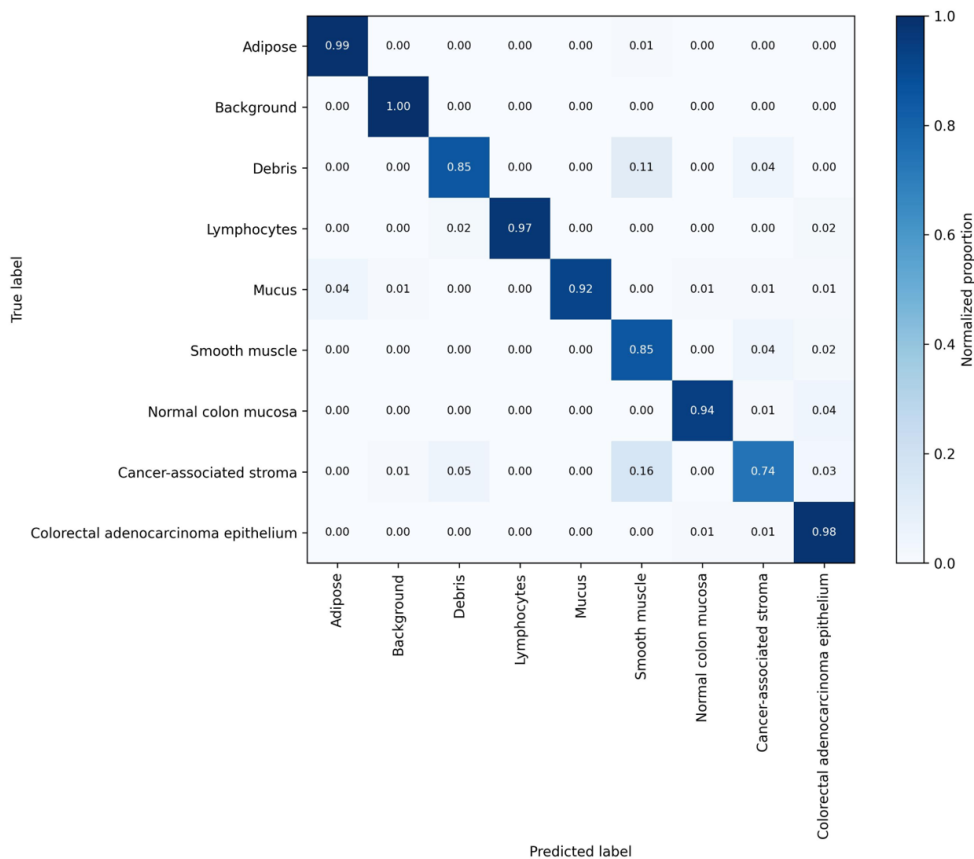


Figure 5. Confusion-matrix heatmap of PathRes_FM predictions

Figure 6 illustrates the comparative performance of PathRes_FM against the listed baseline models. Consistent with Table 8, the proposed model achieved the highest accuracy and tied for the highest AUC–ROC with ResNet-50 among the evaluated approaches, supporting the performance benefit of the final PathRes_FM configuration for PathMNIST classification.

Figure 7 shows the class-wise F1-score distribution across the nine tissue categories. The strongest categories include lymphocytes, adipose, and colorectal adenocarcinoma epithelium, whereas cancer-associated stroma and smooth muscle remain comparatively weaker. This plot provides a direct visual summary of the model's strengths and the tissue categories that require further improvement.

Figure 8 presents the t-SNE feature-space visualization and provides insight into the separability of the learned internal representations. The visualization indicates that well-separated clusters correspond to effective feature learning, whereas overlapping clusters suggest morphologically similar categories. This is consistent with the reported high overall AUC–ROC and strong class-wise results for visually distinctive tissues, while also explaining why some structurally related categories remain harder to classify.

In Figure 9, the three-dimensional feature manifold visualization shows that the learned deep representations form several compact and clearly separated clusters in the embedded feature space. This indicates that the PathRes_

FM backbone learns discriminative internal features for multiple colorectal tissue types. At the same time, a few clusters appear closer to one another or partially elongated, which is consistent with the class-level difficulty seen for morphologically similar tissues. Thus, the three-dimensional manifold visualization offers qualitative evidence that the model has learned a structured feature organization rather than arbitrary class boundaries.

The reliability diagram (Figure 10) analyzes how accurate the PathRes_FM is at expressing its confidence, or whether its high predictions truly translate into consistent accuracy. In medical imaging classification, this type of analysis is a necessity, as it shows when confident predictions can turn out to be incorrect—a major concern that may undermine physician confidence in automated systems. A perfectly calibrated classifier would exhibit reliability along the dashed diagonal reference line, so higher calibration quality indicates that predicted confidence is appropriately aligned with performance (e.g., predicting with confidence = 0.90 corresponds to a 90% actual accuracy).

Figure 11 plots class support against F1-score and provides another view of class-level robustness. Adipose and colorectal adenocarcinoma epithelium show both high support and high F1-score, while lymphocytes achieve one of the best F1-scores despite having fewer samples than several other classes. In contrast, cancer-associated stroma has the lowest F1-score even though it is not the least represented class, and smooth muscle also shows moderate support with a lower F1-score than many other categories. These observations support the interpretation

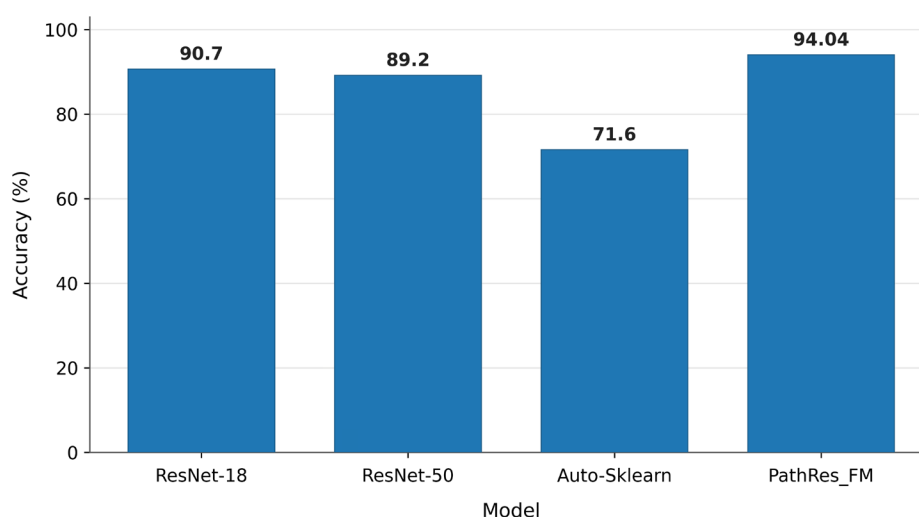


Figure 6. Comparative performance of PathRes_FM against baseline models
Abbreviations: Acc: Accuracy; SOTA: State of the art.

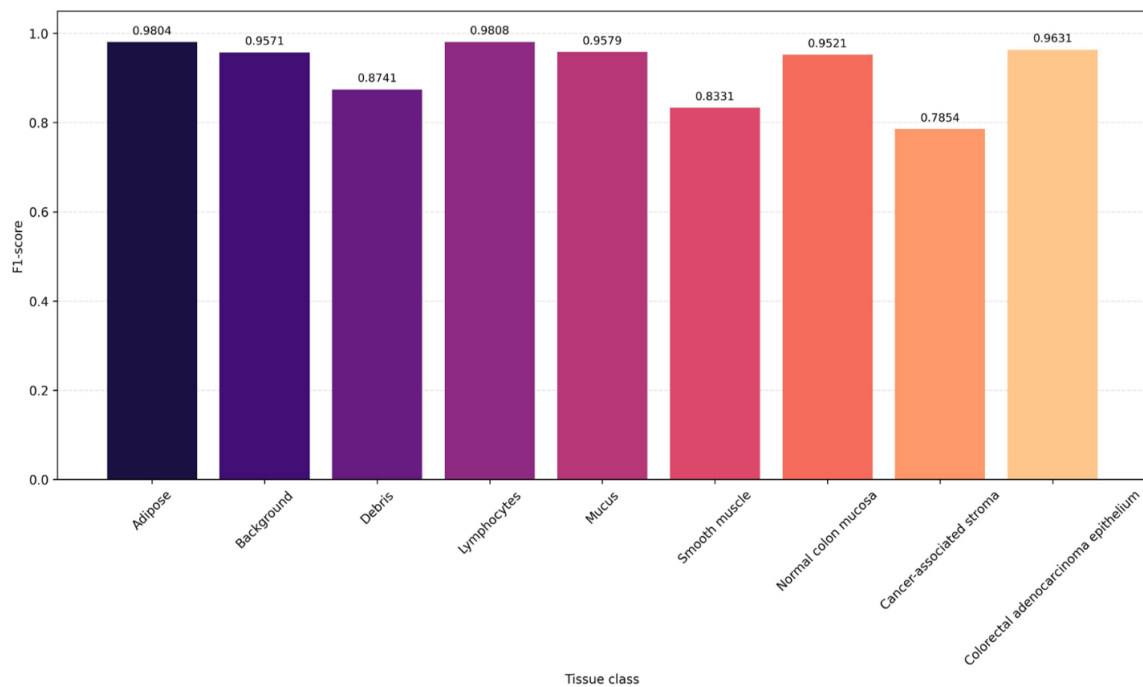


Figure 7. Class-wise F1-score distribution across tissue categories

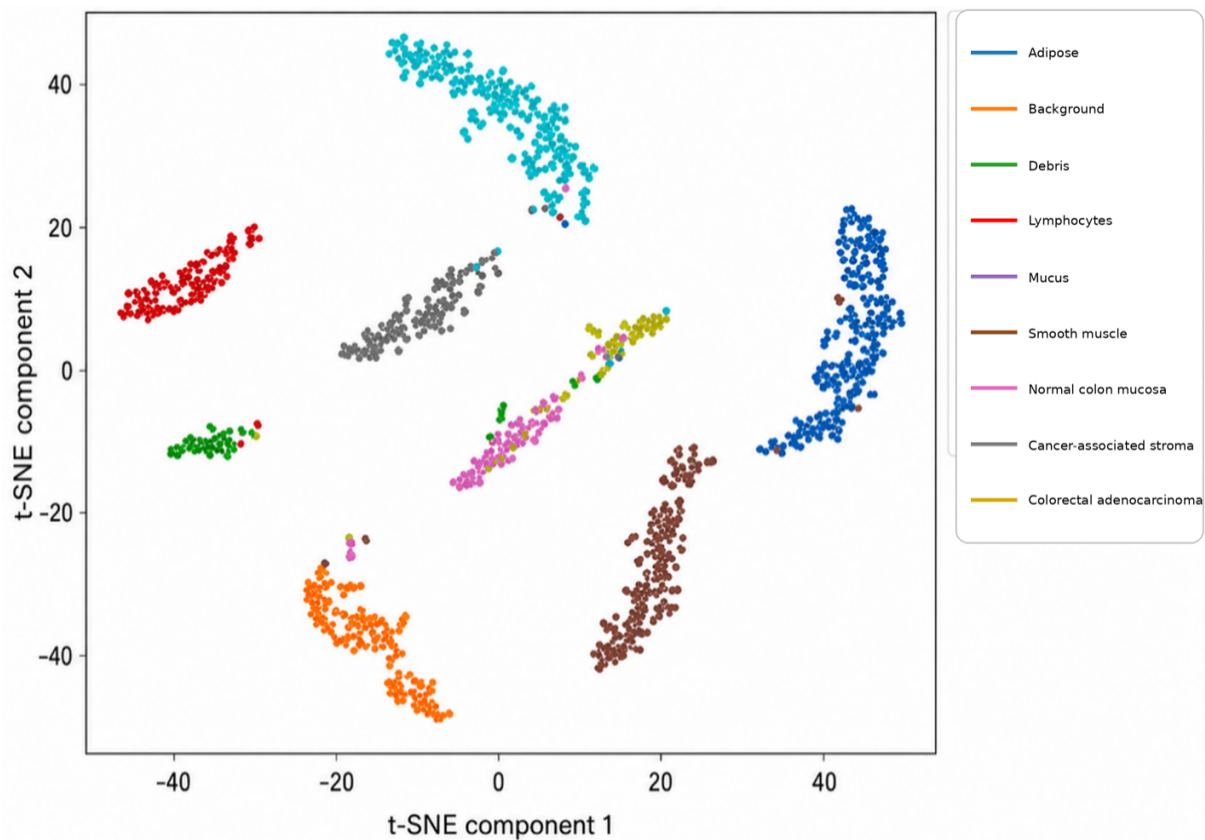


Figure 8. t-distributed stochastic neighbor embedding (t-SNE) visualization of learned feature representations

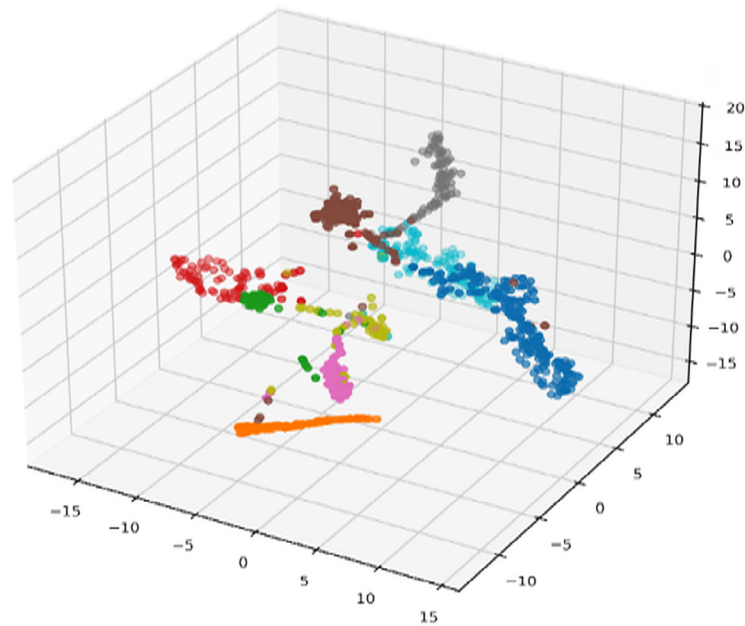


Figure 9. Three-dimensional feature manifold visualization of learned pathology embeddings

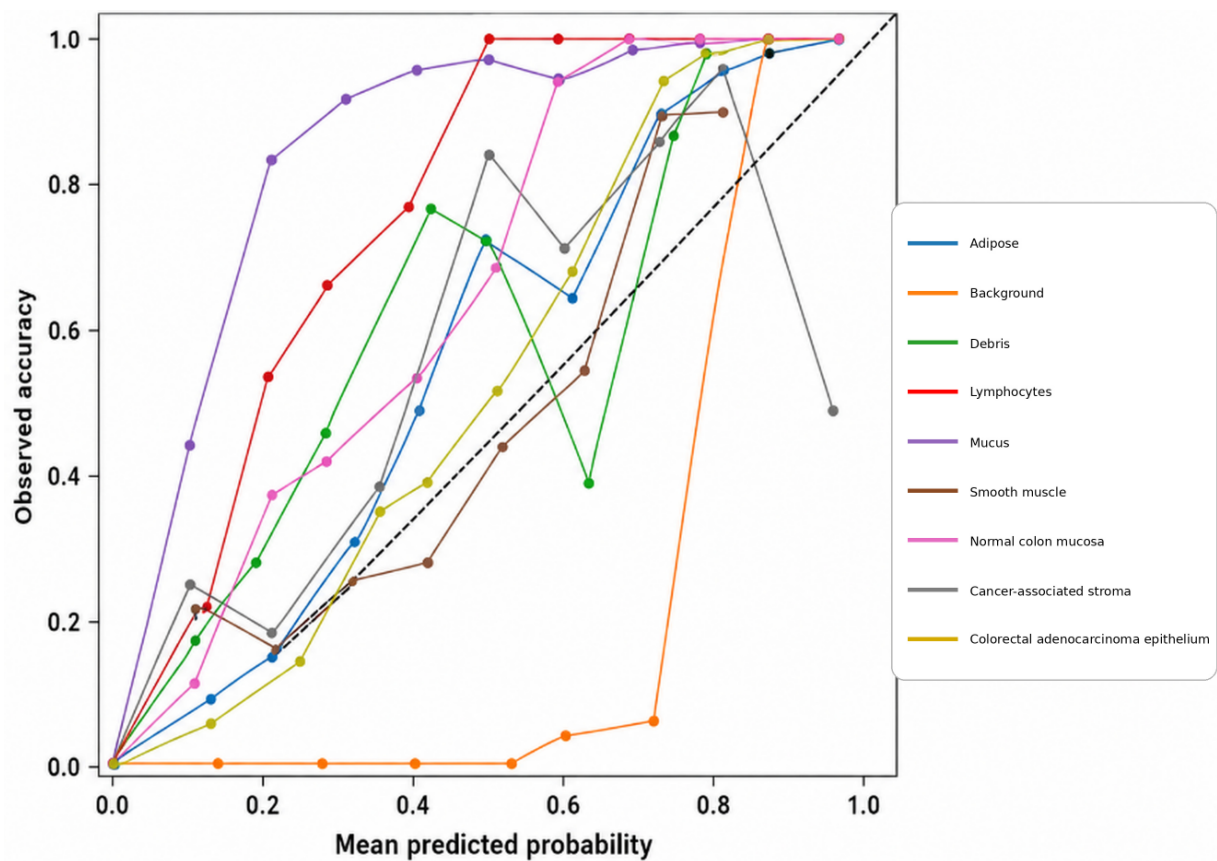


Figure 10. Reliability diagram/calibration analysis of PathRes_FM

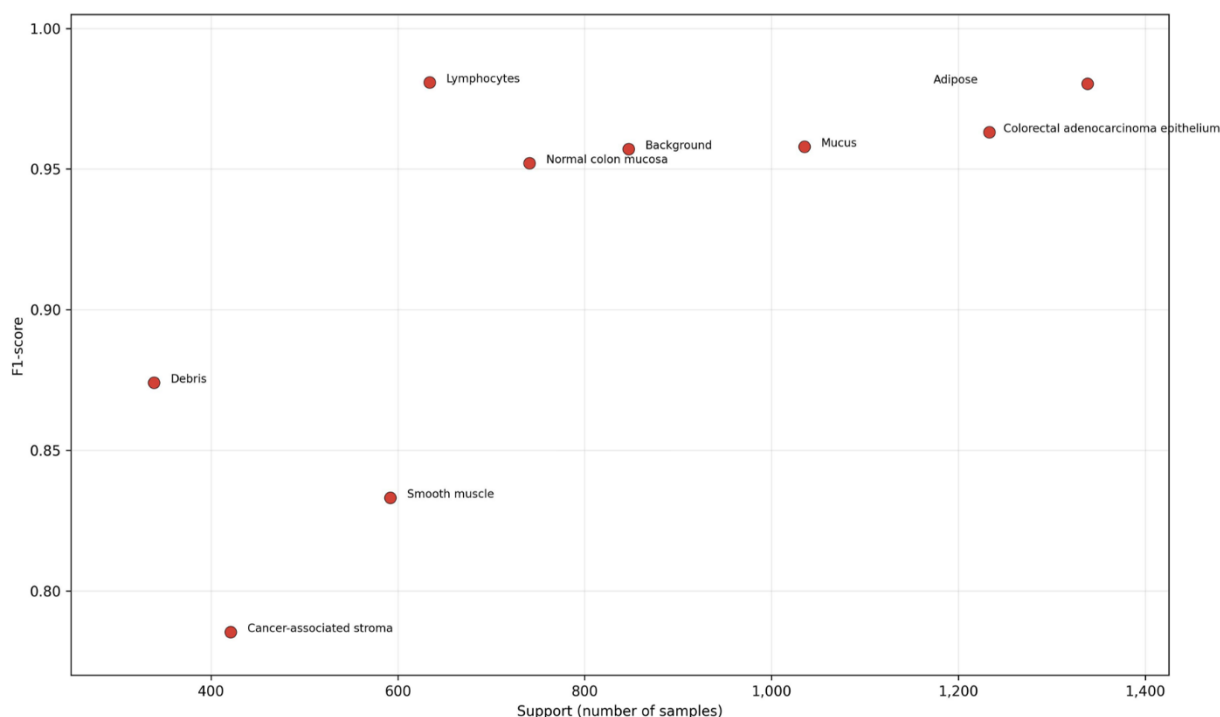


Figure 11. Relationship between class support and F1-score

that tissue morphology and inter-class similarity influence performance more strongly than sample count alone.

Figure 12 presents the average precision for each tissue class and complements the precision–recall curves by summarizing class-level retrieval quality. The highest average precision values are observed for visually distinctive categories such as lymphocytes, adipose, mucus, and colorectal adenocarcinoma epithelium, while lower values indicate classes that require further model refinement.

Figure 13 presents qualitative prediction examples from the test set. Each image is shown with the ground-truth label and the model-predicted label, allowing visual inspection of correct and incorrect classifications. This qualitative view helps determine whether errors occur in ambiguous, blurred, or mixed-tissue patches and complements the numerical evaluation.

Overall, the results demonstrate that the proposed PathRes_FM framework is a strong and competitive model for multi-class colorectal histopathology tissue classification. The global metrics confirm excellent overall predictive quality; the class-wise analysis shows particularly strong performance on diagnostically important tissue types, and the baseline comparison establishes clear superiority over the listed reference methods. At the same time, the figure-based analyses reveal that the remaining limitations are concentrated in a few morphologically

challenging categories rather than reflecting a general weakness of the model. Therefore, within the scope of the implemented image-level benchmark study, the results support the effectiveness of PathRes_FM as a reliable colorectal histopathology classification framework.

Although the results are promising, several limitations remain. The study used only the PathMNIST benchmark and did not include an external validation cohort, multi-center dataset, scanner-diverse cohort, or WSI-level testing. Therefore, the findings should be interpreted as benchmark-level evidence of image-level classification performance rather than evidence of clinical robustness or deployment readiness.

The current experiments also did not include repeated runs, confidence intervals, formal statistical significance testing, or a complete ablation study of dropout, MixUp, CutMix, label smoothing, AdamW, and OneCycleLR. These analyses are necessary to determine the independent contribution of each regularization component and to verify whether the observed performance gains are statistically stable. Similarly, the Grad-CAM results provide qualitative interpretability only; future validation by expert pathologists and quantitative localization assessment would be required for stronger explainability claims.

6. Future work

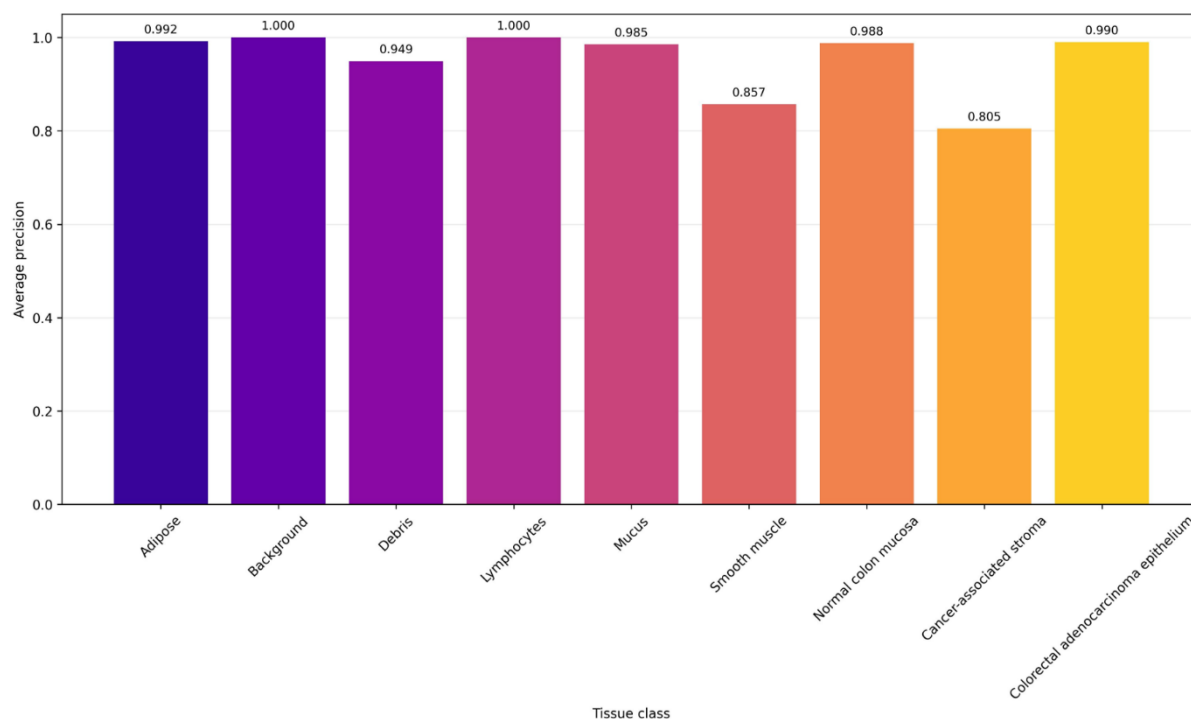


Figure 12. Average precision by tissue class

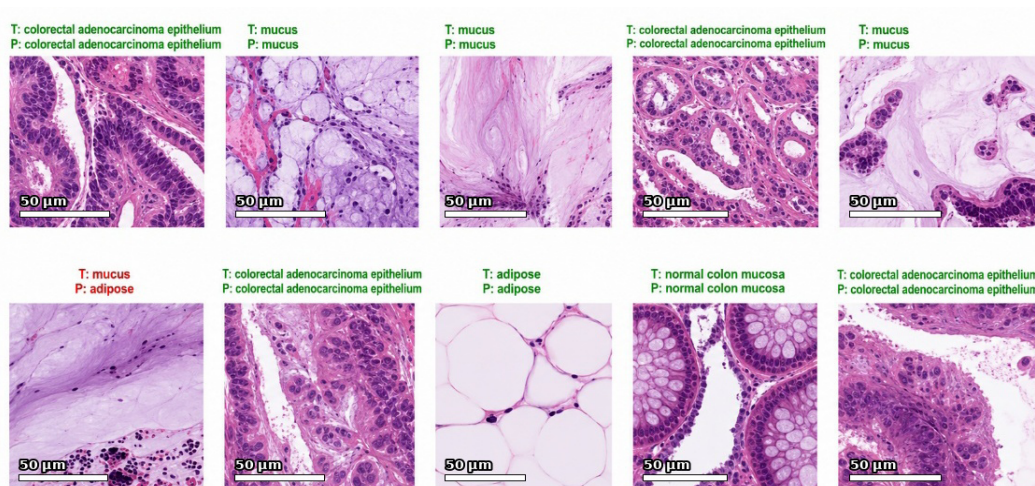


Figure 13. Qualitative prediction examples of the proposed PathRes_FM model (scale bar = 50 µm; magnification = 20×)

Although the present study demonstrates strong performance on the PathMNIST image-level colorectal histology dataset, several future directions remain important. First, future work should improve performance in the comparatively weaker categories, especially cancer-associated stroma and smooth muscle. Potential directions include hard-sample mining, attention-based feature learning, class-aware or difficulty-aware loss functions, and targeted augmentation strategies for morphologically

similar tissue classes.

Second, the current implementation is limited to image-level classification on PathMNIST. Future work should extend the framework toward WSI analysis, where local patch predictions can be combined with larger tissue context and patient-level information. This extension would better align the method with real pathology workflows.

Third, future studies should investigate pathology-specific self-supervised or weakly supervised pretraining using larger collections of unlabeled pathology images. The present model uses ImageNet-pretrained ResNet-50 weights, which provide strong performance; however, pathology-specific representations may further improve robustness and reduce dependence on labeled data. External validation on additional colorectal pathology datasets and multi-center cohorts is also necessary to assess generalization beyond PathMNIST.

Finally, explainability, calibration, and uncertainty estimation should be strengthened in future work. Although the current study includes Grad-CAM-style visualization, t-SNE analysis, and calibration assessment, stronger validation would require expert pathologist review, quantitative localization assessment, cross-dataset robustness testing, and more comprehensive uncertainty analysis. Multimodal extensions that integrate histology images with pathology reports or clinical context may also be explored as a future research direction beyond the scope of the present study.

7. Conclusion

This study presents PathRes_FM, a regularized ResNet-50 transfer-learning framework for nine-class colorectal histopathology tissue classification using the PathMNIST benchmark. The model achieved an overall accuracy of 94.04%, AUC-ROC of 0.9890, weighted F1-score of 0.9399, weighted precision of 0.9410, and weighted recall of 0.9404 on the official test split. Class-wise results showed strong performance for adipose, lymphocytes, mucus, normal colon mucosa, and colorectal adenocarcinoma epithelium, while cancer-associated stroma, smooth muscle, and debris remained more difficult because of morphological overlap.

The present study clarifies that PathRes_FM is a supervised benchmark-level image-classification framework rather than a new whole-slide foundation model or clinical deployment system. Its value lies in a reproducible regularized training pipeline and comprehensive evaluation using global metrics, class-wise analysis, ROC and precision-recall curves, calibration, feature-space visualization, and Grad-CAM-style interpretation. Future work should focus on external validation, direct baseline retraining, ablation studies, repeated experiments with confidence intervals, stain and scanner robustness, expert-validated explainability, and extension toward whole-slide and pathology-specific representation-learning systems.

Acknowledgments

None.

Funding

This research did not receive any specific grant from public, commercial, or not-for-profit funding agencies.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this manuscript.

Author contributions

Conceptualization: All authors

Formal analysis: All authors

Investigation: All authors

Methodology: Prakashini Ambati

Writing-original draft: Prakashini Ambati

Writing-review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The dataset used in this study is the publicly available PathMNIST subset of the MedMNIST benchmark. It can be accessed through the MedMNIST Python API using `data_flag = "pathmnist."` The study used the official PathMNIST training and test splits. Derived results, trained model outputs, and additional implementation details are available from the corresponding author upon reasonable request.

References

- Bousis, D., Verras, G. I., Bouchagier, K., Antzoulas, A., Panagiotopoulos, I., Katinioti, A., Kehagias, D., Kaplanis, C., Kotis, K., Anagnostopoulos, C. N., & Mulita, F. (2023). The role of deep learning in diagnosing colorectal cancer. *Gastroenterology Review*, 18(3), 266–273.
<https://doi.org/10.5114/pg.2023.129494>
- Campanella, G., Hanna, M. G., Geneslaw, L., Miralflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8),

1301–1309.

<https://doi.org/10.1038/s41591-019-0508-1>

Campanella, G., Chen, S., Singh, M., Verma, R., Muehlstedt, S., Zeng, J., Stock, A., Croken, M., Veremis, B., Elmas, A., Shujski, I., Neittaanmäki, N., Huang, K., Kwan, R., Houldsworth, J., Schoenfeld, A. J., & Vanderbilt, C. (2025). A clinical benchmark of public self-supervised pathology foundation models. *Nature Communications*, 16(1).

<https://doi.org/10.1038/s41467-025-58796-1>

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., & Mahmood, F. (2024). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3), 850–862.

<https://doi.org/10.1038/s41591-024-02857-3>

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A. L., Razavian, N., & Tsirigos, A. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10), 1559–1567.

<https://doi.org/10.1038/s41591-018-0177-5>

Davidson, K. W., Barry, M. J., Mangione, C. M., Cabana, M., Caughey, A. B., Davis, E. M., Donahue, K. E., Doubeni, C. A., Krist, A. H., Kubik, M., Li, L., Ogedegbe, G., Owens, D. K., Pbert, L., Silverstein, M., Stevermer, J., Tseng, C.-W., & Wong, J. B. (2021). Screening for colorectal cancer. *JAMA*, 325(19), 1965.

<https://doi.org/10.1001/jama.2021.6238>

Ding, T., Wagner, S. J., Song, A. H., Chen, R. J., Lu, M. Y., Zhang, A., Vaidya, A. J., Jaume, G., Shaban, M., Kim, A., Williamson, D. F. K., Robertson, H., Chen, B., Almagro-Pérez, C., Doucet, P., Sahai, S., Chen, C., Chen, C. S., Komura, D., ... Mahmood, F. (2025). A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11), 3749–3761.

<https://doi.org/10.1038/s41591-025-03982-3>

Hu, D., Jiang, Z., Shi, J., Xie, F., Wu, K., Tang, K., Cao, M., Huai, J., & Zheng, Y. (2025). Pathology report generation from whole slide images with knowledge retrieval and multi-level regional feature selection. *Computer Methods and Programs in Biomedicine*, 263, 108677.

<https://doi.org/10.1016/j.cmpb.2025.108677>

Lee, J., Lim, J., Byeon, K., & Kwak, J. T. (2025). Benchmarking pathology foundation models: Adaptation strategies and scenarios. *Computers in Biology and Medicine*, 190, 110031.

<https://doi.org/10.1016/j.compbiomed.2025.110031>

Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide

images. *Nature Biomedical Engineering*, 5(6), 555–570.

<https://doi.org/10.1038/s41551-020-00682-w>

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L. P., Gerber, G., Parwani, A. V., Zhang, A., & Mahmood, F. (2024). A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3), 863–874.

<https://doi.org/10.1038/s41591-024-02856-4>

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Zhao, M., Chow, A. K., Ikemura, K., Kim, A., Pouli, D., Patel, A., Soliman, A., Chen, C., Ding, T., Wang, J. J., Gerber, G., Liang, I., Le, L. P., Parwani, A. V., Weishaupt, L. L., & Mahmood, F. (2024). A multimodal generative AI copilot for human pathology. *Nature*, 634(8033), 466–473.

<https://doi.org/10.1038/s41586-024-07618-3>

Neidlinger, P., El Nahhas, O. S. M., Muti, H. S., Lenz, T., Hoffmeister, M., Brenner, H., van Treeck, M., Langer, R., Dislich, B., Behrens, H. M., Röcken, C., Foersch, S., Truhn, D., Marra, A., Saldanha, O. L., & Kather, J. N. (2025). Benchmarking foundation models as feature extractors for weakly supervised computational pathology. *Nature Biomedical Engineering*, 10(6), 1113–1123.

<https://doi.org/10.1038/s41551-025-01516-3>

Nicke, T., Schäfer, J. R., Höfener, H., Feuerhake, F., Merhof, D., Kießling, F., & Lotz, J. (2025). Tissue concepts: Supervised foundation models in computational pathology. *Computers in Biology and Medicine*, 186, 109621.

<https://doi.org/10.1016/j.compbiomed.2024.109621>

Pocevičiūtė, M., Ding, Y., Bromée, R., & Eilertsen, G. (2025). Out-of-distribution detection in digital pathology: Do foundation models bring the end to reconstruction-based approaches? *Computers in Biology and Medicine*, 184, 109327.

<https://doi.org/10.1016/j.compbiomed.2024.109327>

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., & Zhang, Y. (2021). TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34, 2136–2147.

Tran, M., Schmidle, P., Guo, R. R., Wagner, S. J., Koch, V., Lupperger, V., Novotny, B., Murphree, D. H., Hardway, H. D., D'Amato, M., Lefkes, J., Geijs, D. J., Feuchtinger, A., Böhner, A., Kaczmarczyk, R., Biedermann, T., Amir, A. L., Mooyart, A. L., Ciompi, F., ... Peng, T. (2025). Generating dermatopathology reports from gigapixel whole slide images with HistoGPT. *Nature Communications*, 16(1).

<https://doi.org/10.1038/s41467-025-60014-x>

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., Yang, E., Mathieu, P., van Eck, A.,

- Lee, D., Viret, J., Robert, E., Wang, Y. K., Kunz, J. D., Lee, M. C. H., ... Fuchs, T. J. (2024). A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10), 2924–2935.
<https://doi.org/10.1038/s41591-024-03141-0>
- Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., & Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81, 102559.
<https://doi.org/10.1016/j.media.2022.102559>
- Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., ... Poon, H. (2024). A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015), 181–188.
<https://doi.org/10.1038/s41586-024-07441-w>
- Xu, Y., Wang, Y., Zhou, F., Ma, J., Jin, C., Yang, S., Li, J., Zhang, Z., Zhao, C., Zhou, H., Li, Z., Lin, H., Wang, X., Wang, J., Han, A., Chan, R. C. K., Liang, L., Zhang, X., & Chen, H. (2025). A multimodal knowledge-enhanced whole-slide pathology foundation model. *Nature Communications*, 16(1).
<https://doi.org/10.1038/s41467-025-66220-x>
- Yan, F., Wu, J., Li, J., Wang, W., Chen, Y., Wei, L., Lu, J., Chen, W., Gao, Z., Li, J., Li, H., Yan, H., Ma, J., Chen, M., Lu, Y., Chen, Q., Wang, Y., Ling, X., Wang, X., ... Wang, Z. (2025). PathOrchestra: A comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks. *npj Digital Medicine*, 8(1).
<https://doi.org/10.1038/s41746-025-02027-w>
- Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2023). MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1).
<https://doi.org/10.1038/s41597-022-01721-8>