

Leakage-Resistant Colorectal Histopathology Classification Using ConvNeXt-Tiny and the Midnight-12k Pathology Foundation Model

Ahmed F. Fadhil* 

Technical management institute, Northern technical, Mosul, Iraq.

*Correspondence: Ahmed.fehr@ntu.edu.iq



Article – Peer Reviewed

Received: July 16, 2026

Accepted: September 02, 2026

Published: September 14, 2026

Copyright: © 2026 RAME Publishers

This is an open access article under the CC BY 4.0 International License.



<https://creativecommons.org/licenses/by/4.0/>

Cite this article: Ahmed F. Fadhil, "Leakage-Resistant Colorectal Histopathology Classification Using ConvNeXt-Tiny and the Midnight-12k Pathology Foundation Model", *International Journal of Computational and Electronic Aspects in Engineering*, RAME Publishers, vol. 7, issue 4, pp. 399-415, 2026.

<https://doi.org/10.26706/ijceae.7.4.20260808>

Abstract: The classification of colorectal histopathology images is crucial for computer-assisted pathological diagnosis, however the similarity of morphological structures between the categories of tissue makes multiclass classification challenging. This paper explores whether a Patho generative foundation model can better differentiate colorectal tissues in a leakage-safe evaluation protocol. The four-class dataset consisted of 6,160 images of histopathology tissue types — Normal, Serrated, Adenocarcinoma, and Adenoma — that were tested with ConvNeXt-Tiny, the first proposed framework and Midnight-12k, the last. After applying the official test benchmark, the overall accuracy and other metrics for Midnight-12k were: 81.39 % accuracy, 81.50 % weighted F1-score, 93.89 % macro specificity, 94.90 % macro-ROC-AUC, 83.17 % macro PR-AUC, and MCC 0.7366. The precision (95.69%), recall (97.86%), F1-score (96.76%) and ROC-AUC (99.60%) were well achieved for Adenocarcinoma. F1 score of 82.18% with normal tissue and precision of 92.20%, specificity of 99.41% with Adenoma achieved. Overall macro precision, recall, and F1-score were 79.73%, 73.47%, and 74.43%, respectively. Midnight-12k boosted the accuracy from 74.58% to 77.24% and macro F1-score from 70.38% to 73.18% compared to ConvNeXt-Tiny under strict five-fold evaluation. The results illustrate the promise of pathologists' foundation models in colorectal histopathology classification and underscore the difficulty of distinguishing between Serrated and Adenoma.

Keywords: Deep Learning; histopathology images; ConvNeXt-Tiny; Midnight-12k; Pathology foundation model.

1. Introduction

Colorectal cancer (CRC) is a significant health problem all over the world and is the third most prevalent cancer type in the world, accounting for the second cause of cancer related death. In 2022, the world was estimated to have seen about 1.9 million new CRC cases and over 900,000 deaths from CRC. A prompt and effective diagnosis is therefore crucial if treatment has to be planned as effectively as possible and outcomes are to be as good as possible. Histopathological examination of biopsy or surgical resected tissues is still critical in CRC diagnosis as it can help the pathologist to determine whether the tissue is not cancerous, benign or cancerous based on the features of the tissue seen in the microscope [1].

With a rise in the use of digital pathology, there has been a growing possibility to use artificial intelligence (AI) and deep-learning techniques to support pathologists in analyzing tissue images. There is a trend towards digitally processing tissue images creating opportunities for the application of artificial intelligence (AI) and deep-learning techniques in pathology. The deep neural networks can be used for automatic learning of discriminative morphological patterns in histopathological images and have been explored for applications such as tissue classification, cancer detection, cancer grading, biomarker prediction and even prognosis. However, computational histopathology is still a challenge as there are significant intra-class variations in tissue structure and similar morphological properties of different diagnostic classes. Moreover, sources of histopathology datasets are susceptible to the problem of class imbalance, labeling uncertainty, and the expensive collection of expert labels [2].

The issues are more pronounced in the Chaoyang colorectal histopathology dataset which was released by Zhu et al. to work with a robust histopathology classification in realistic annotation setting. This dataset includes 6,160 patches of colorectal tissue images from four diagnostic categories: (Normal, Serrated, Adenocarcinoma, and Adenoma). There are 4021 training images and 2139 official test images. Examples of these tissue classes are illustrated in Fig. 1. The figure shows an example of the large morphological diversity found in the classes themselves and also indicates how close together some classes can be, for example, Serrated lesions are visually similar to Adenoma. The similarities can make it difficult for reliable multiclass classification and encourage the employment of more discriminative feature representations. Three pathologists were reported to have independently annotated the images in the original study, with images being consistently annotated forming the test set, while there was disagreement in a large number of the training images. Therefore, this dataset is a challenging classification problem due to the similarity of morphology structure and the uncertainty of the annotations [3].

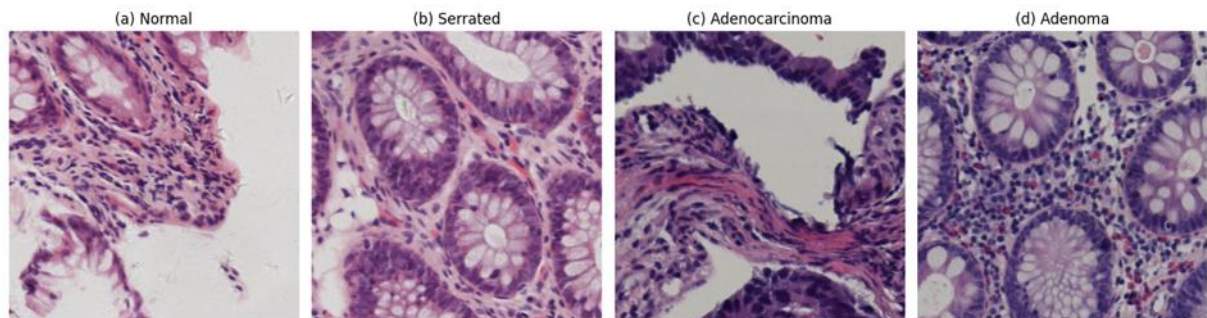


Figure 1. Representative colorectal histopathology images from the four classes used in this study: (a) Normal, (b) Serrated, (c) Adenocarcinoma, and (d) Adenoma.

Convolutional neural networks (CNNs) have been widely adopted in the development of many classification systems for medical images in the past. Recently, the modern CNN architectures like ConvNext have adopted the idea of designing CNN using concepts from the field of Vision Transformers without replacing convolution operations. Liu et al. showed that ConvNeXT is a powerful and general visual representation network for image recognition tasks. ConvNeXT-Tiny was selected, to this end, for the present study as the main conventional baseline, against which pathology-specific representations were assessed [4]. Existing models based on natural images are effective but may not best describe the specialized morphological structure of pathological tissues. However, this constraint has encouraged the new surge of the progressivity of pathology foundation models, trained using either self-supervised or related pretraining techniques on large sets of histopathology images to acquire a general representation. Lately, such models have been shown to be capable of generalizing to many different downstream pathology tasks and minimize the need for substantial amounts of task-specific labelled data. For instance, UNI presented its results on the general-purpose pathology representations that performed well on various computational pathology (CP) benchmarks, and Virchow used large-scale pathology-specific pretraining for downstream tasks such as cancer detection [5].

This trend has been continued by more recent pathology foundation models. For instance, Phikon-v2 was trained on millions of histology image tiles using the self-supervised learning paradigm and successfully transferred to a range of pathology tasks [6]. In this context, the family of pathology foundation models, Midnight', is of relevance to the present investigation, as it was introduced by Karasikov et al. The authors showed that they were able to learn strong pathology representations with significantly fewer whole slide images as compared to several previous large scale foundation models. Similar to prior approaches, the Midnight-12k variant, with about 12,000 whole slide images for pretraining, achieved competitive performance on a suite of downstream pathology tasks [7]. However, the performance of a powerful pretrained representation is not sufficient to guarantee reliable evaluation. Conventional image-level random splitting might lead to the inclusion of correlated images in both training and validation sets, especially for histopathology datasets with multiple images from different, but related, tissue sources. This can lead to overly optimistic estimates of generalization. The present study thus focuses on choosing an evaluation protocol that prioritizes leakage, by restricting images from the same identified source group to a single fold of the validation set. Important to note, these groups are not referred to as patient identities, but as root-source groups, since there's no patient-level information found in the metadata behind the dataset that was used in this work.

2. Objective and contribution

This study aims at determining if representations learned on a foundation model on a colorectal specialized task can be added to a strong conventional visual backbone for four-class colorectal histopathology classification, while strictly following an evaluation protocol where the models used as a source task are clearly defined without being generalized to the target task.

The main contributions of this work are:

- The implementation of a root-source-group-aware five-fold validation technique to minimize the likely root source leakage.
- A direct comparison between a strong conventional ImageNet baseline trained with ConvNeXt-Tiny and the representations learned with the Midnight-12k using the same source-group-aware validations folds.
- The implementation of a more efficient final framework where the pretrained Midnight-12k encoder is kept frozen, and the light weight classification head is trained.
- complete evaluation involving paired out-of-fold comparison, class specific performance, evaluation using statistics, and separately stored benchmark for the official test.

This framework is not meant to be a pathological alternative, but to evaluate the effect of domain-specific foundation representations and careful evaluation design in enhancing automated colorectal tissue classification. Special focus will be given to the difficult separation of Serrated and Adenoma tissue patterns, and if pathology-specific representations will help with classification across all four types of tissue.

3. Literature Review

In recent years, the study of colorectal histopathology has begun to shift from classical convolutional neural networks towards techniques like transfer learning, attention mechanism, and the "Transformers" as well as pathology-specific foundation models. While great progress has been made in the aforementioned strategies, there are still issues such as dataset variability, computational complexity for models, external validation, class imbalance, and potential information leakage to consider.

Neto et al. [8] proposed an interpretable deep-learning framework for the whole-slide image weakly and mixed supervised colorectal cancer diagnosis. The system achieved an accuracy 93.44% (using approximately 10,500 WSIs) and 84.91% (evaluated on the external TCGA cohort). The authors suggested that part of this reduction is due to the differences in tissue morphology between training data and TCGA. The research also cited the problem of the vast amount of WSI data, the need to have experts annotate the data, and the possible occurrence of false negatives during summarization at the patch level. The system has shown that it has a significant diagnostic capability however, the external performance of the system shows that the ability to hold the accuracy across the various pathology sources holds challenges.

Yengec-Tasdemir et al.[9] presented a colorectal polyp classification method based on supervised contrastive learning and Big Transfer (BiT). The method was designed to learn discriminative representations of morphologically similar polyp subtypes in both a custom dataset and the public UniToPatho dataset, for which the accuracy values of 87.1% and 70.3% were obtained respectively. One important drawback is the large gap between the results for the two datasets. The experiments showed the benefits of contrastive representation learning; however, there was a drop in performance on UniToPatho, reflecting the impact of dataset properties and domain variation. Furthermore, the research in the present work concerned with colorectal polyp classification, not broadly within Normal–Serrated–Adenoma–Adenocarcinoma classification problem as in the present study.

Fu et al.[10] reported that the transformer model is effective for the classification of colon cancer histopathological images. Their methodology included farthest-point sampling to choose representative WSI patches, multi-scale grouping to add information in the surrounding patches, Transformer-based modeling, and focal loss to handle class imbalance. The model was tested with a hospital sample and with the samples from TCGA and was superior to the compared methods. However, it was only tested on two WSI datasets, and the technique must carry out several processing steps: patch sampling, feature extraction, spatial grouping, and the construction of a Transformer model. This is more complicated than those methods that resort to a pretrained representation and require a lightweight downstream classifier. In addition, it has a WSI-level experimental setting that cannot be directly compared with patch-level colorectal tissue classification as performed in Chaoyang.

Ke et al.[11] recently proposed a framework combining domain-specific transfer learning, multiple pretrained CNNs, and multi-head self-attention feature fusion. The model was tested on EBHI, Chaoyang, and COAD with an accuracy of 86.72% when tested in Chaoyang - important for the present study. The authors utilized 70%-20%-10% training-validation-test set split, and fused multiple CNNs that were trained independently before the attention based feature fusion. The authors,

however, explicitly mentioned the need for higher computational demands: They claimed the pretraining and fine-tuning phases are required to be completed separately over several different networks, which increases the fine-tuning time for the target domain several folds from the computational load of each CNN model in isolation. Furthermore, the reported experimental protocol doesn't describe the root-source-group separation used in the present study. Therefore, the accuracy for Chaoyang should not be compared directly with those under strict source-group controlled cross validation.

Campanella et al.[12], carried out a clinical benchmark on a number of clinically relevant tasks using publicly available self-supervised pathology foundation models. Their results indicated that, in general, modern pathology-specific encoders outperformed ImageNet-pretrained representations, with differences among some of the more recent DINO/DINOv2 relatively small. They also noted that there were significant technical variations in the tissue fixation, staining and scanning performed across data from three institutions only, a focus on tile-level feature representations using a fixed aggregation strategy, and the lack of a scope study of more complex slide-level modeling. They also observed that the traditional tile-based size might not be suitable for certain prediction problems that involve more complex topographic data across the globe.

Neidlinger et al.[13], conducted an even wider independent performance, comparing 19 pathology foundation models across 13 patient cohorts, 6818 patients and 9528 slides for lung, colorectal, gastric and breast cancer. The study showed that there is no single “best” foundation model and that if its foundation is trained with different data, they can bring complementary representations. What's important is that performance was lower in low data, low prevalence scenarios. The authors noted that the comprehensive evaluation, while undertaken, was confined to four types of cancers only; however, model inclusion was limited by licensing and research accessibility. Therefore, even fairly large-scale independent benchmarks, cannot capture the behaviour of foundation models across all pathological tissue, diagnosis and clinical settings.

Karasikov et al.[7], proposed the Midnight family of pathology representation models, showing that competing representations of pathology could be achieved from very few WSIs when compared to some of the modern FMs. Specifically, our model Midnight-12k was trained over some 12,000 WSIs from the TCGA released dataset, and its downstream performance is competitive despite the smaller pretraining set it used. However, there are some identified limitations, which were published for evaluation. Reviewer analyses identified problems with task-specific degradation of some higher resolution settings, the lack of knowledge on the significance of individual training components and of the representativeness of proprietary parts of the larger pretraining sets. In addition, it was noted that the original experimental comparison did not employ a statistical significance test, which could be considered a limitation. These considerations highlight the need to evaluate these foundation-model tools independently on the downstream task as opposed to taking for granted that a good performance on the benchmark will automatically be reflected on any downstream histopathology task.

Very recent evidence further reinforces this point. Bareja et al.[14], benchmarked 32 general vision, vision-language, and pathology foundation models for patch- and slide-level computational pathology tasks. Pathology-specific vision models usually achieved good results, results in external and out-of-domain test sets differed and, sometimes, the difference in the top models was only moderate and was task dependent. Furthermore, there was no consistent correlation between the amount of pretraining data and downstream performance, nor was there between model size and performance. This benchmark is a large-scale test and shows an important current drawback of foundation model research: that a good model on one benchmark does not necessarily mean that the large-scale model is optimized for other pathology tasks.

Overall, recent studies show significant advancement in the area of automated colorectal histopathology classification and pathology foundation modeling, but some restrictions still exist. Conventional multi-network approaches are able to perform well but are often very complex, whereas foundation models can offer task-dependent performance and domain sensitivity and lack in robustness when implementing from institution to institution. In addition, there were significant variations in the validation strategy used amongst studies, leaving the comparison of reported accuracy values inconclusive. This recent independent benchmarking has explicitly pointed out contamination and optimistic evaluation in current cases of overlapping institutional or pretraining sources, which cause this if external validation is not adequately monitored[13]. Therefore, direct comparison of models has been investigated in the present work under a five-fold identical root-source-group-aware benchmark, and the statistically valid form of the final model has then been evaluated on a separately preserved official test benchmark where it is kept locked. This design aims at assessing whether the use of conventional image-level splitting is not enough to achieve a consistent improvement when working with pathology-specific foundation representations.

4. Material and Methods

A representation of the baseline and proposed classification architectures is depicted in Fig. 2. The components of the experimental pipeline include image preprocessing, a baseline feature learning network, ConvNeXt-Tiny, a pathology-specific feature extraction network, Midnight-12k, feature aggregation, multi-layer perception of the features, a lightweight classification network, and a thorough source-group-aware evaluation. The key goal was to study whether such

representations learned by a large pathology foundation model could yield more discriminative representations than a traditional ImageNet-trained architecture, with the requirement of a small downstream classifier.

The proposed Midnight-12k: framework has some significant differences with the baseline, especially in the representation stage. The pretrained Midnight-12k encoder was frozen instead of training and fine-tuning it over the target data set. Global information represented with CLS token, and distributed contextual information represented with patch token, were fused and fed to a LayerNorm-linear classification head to classify the 3072-dimensional.

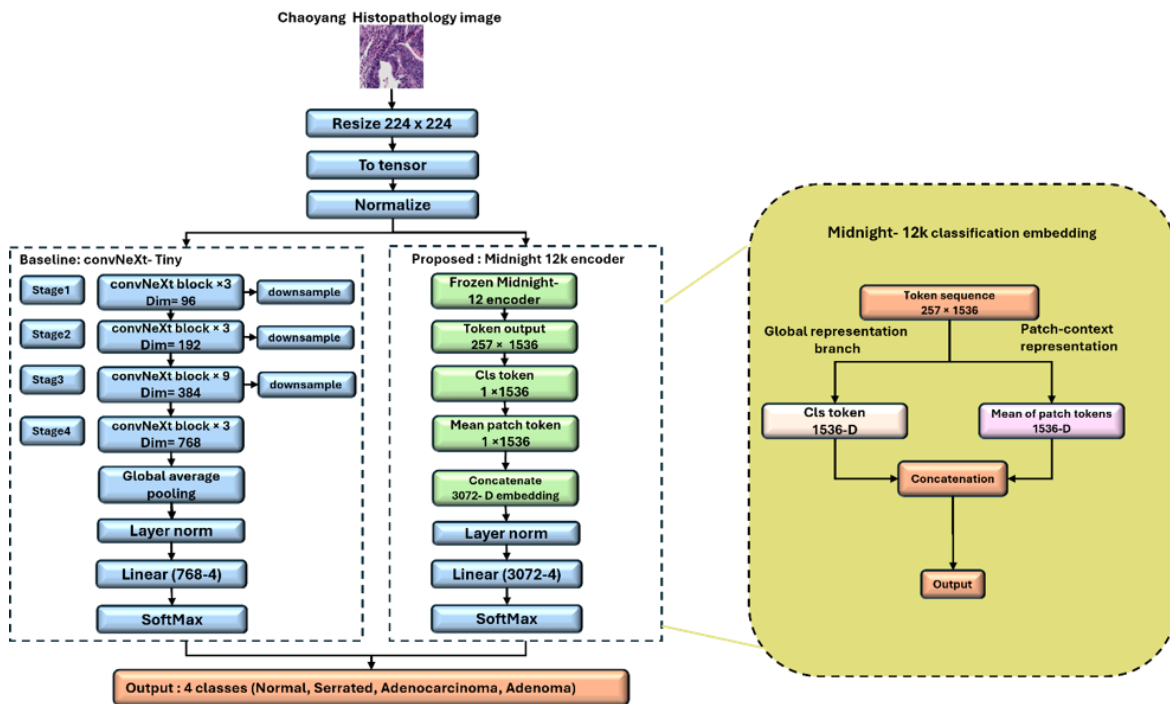


Figure 2 The block diagram of the baseline model ConvNeXt-Tiny and the proposed model Midnight-12k for four classes colorectal histopathology classification. The Midnight-12k encoder continues to be frozen, and the CLS/mean patch-token descriptions are combined to embed a 3072-D vector for lightweight classification.

4.1 Dataset Description

Experiments were carried out on colorectal histopathology images from a database of Chaoyang (6,160 histopathological images originally 512×512 pixel). The data set includes four tissue diagnostics: Normal, Serrated, Adenocarcinoma, Adenoma.

There are 4,021 images in the training partition and 2,139 images in the official test partition, which serves as the benchmark. The class distribution is shown in Table 1.

Table 1. Distribution of the Chaoyang Histopathology Dataset

Class	Training	Official Test	Total
Normal	1,111	705	1,816
Serrated	842	321	1,163
Adenocarcinoma	1,404	840	2,244
Adenoma	664	273	937
Total	4,021	2,139	6,160

Moderate class imbalance is observed with Adenocarcinoma being the largest class and Adenoma being the smallest class. Therefore, the overall accuracy alone was not used to assess the performance, and both the macro-averaged and class-specific metrics were added.

4.2 Hardware and Software Configuration

The experiments were performed using the Python programming language and deep-learning framework PyTorch in a Linux-embedded JupyterLab. The calculations were done on an NVIDIA (GB10) GPU using CUDA acceleration. The hardware and software configuration of the experiments is presented in Table 2.

Table 2. Hardware and software configuration used in the experiments.

Component	Configuration
Deep-learning framework	PyTorch 2.10.0+cu128
Torchvision	0.25.0+cu128
CUDA	12.8
GPU	NVIDIA GB10
Development environment	JupyterLab
Mixed precision	BF16 where applicable

To increase efficiency in both memory usage and computation, mixed-precision inference was applied when extracting the features of the foundational model. Extracted representations were then recorded in 32 bits floating point format to be used for classifier development. The original size of these images is 512×512 pixels. Images were then transformed to the model's required size of 224×224 for Midnight-12k.

The pre-processing of the images was performed as follows:

$512 \times 512 \rightarrow \text{resize} \rightarrow \text{center crop } (224 \times 224) \rightarrow \text{tensor} \rightarrow \text{Normalization}$.

Bicubic interpolation was employed during resizing. The RGB channels were normalized using $\mu = (0.5, 0.5, 0.5)$ and $\sigma = (0.5, 0.5, 0.5)$.

For an image channel IC, normalization is expressed as in (1):

$$I^*c = (Ic - \mu_c) / \sigma_c. \quad (1)$$

5. Leakage-Resistant Root-Source Grouping

A root-source-group-aware splitting strategy was used to minimize information leakage when cross-validation was conducted. Exploring the Chaoyang training-image filenames revealed that more than one image could have the same source identifier, suggesting that they were derived from the same source group. Images from the same source might have correlated histological structures, which could lead to over-optimistic performance estimates if images from the same source were used to both train and validate a model.

For every validation fold, $G_{\text{train}} \cap G_{\text{val}} = \emptyset$. Where G_{train} and G_{val} represent training and validation root-source groups. Thus, all images belonging to a particular root-source group remained together.

To compare the two models fairly, the same fold assignment was used for ConvNeXt-Tiny and Midnight-12k. Importantly, these root-source groups should not be read as identities of patients as patient-level identifiers are not provided in the dataset metadata. The official test set was not used in the construction of folds or in the selection of the model.

6. Baseline Model: ConvNeXt-Tiny

ConvNeXt-Tiny was employed as the main conventional baseline. The architecture contains four hierarchical stages with feature dimensions:

$96 \rightarrow 192 \rightarrow 384 \rightarrow 768$.

The four stages contain: 3, 3, 9, 3 ConvNeXt blocks, respectively. Following hierarchical feature extraction, global average pooling was applied, followed by LayerNorm and a linear classification layer: $768 \rightarrow 4$.

The resulting logits correspond to the four tissue categories: These are: Normal, Serrated, Adenocarcinoma, Adenoma. The baseline generated a standard image-level reconstruction and was the reference architecture for evaluating if pathology-domain specific pretraining could improve upon it.

7. Proposed Midnight-12k Framework

Proposed framework exploits in its framework a large pathologically specific foundation model pretrained on histopathological images called Midnight-12k. Midnight-12k contains approximately 1.136×10^9 parameters. However, the pretrained part of the encoder was kept frozen.

Therefore, θ_{midnight} is constant and no gradients were used to update the encoder parameters. Midnight-12k generates the token for an image I obtained as the input histopathology image.

$$Z = E(I), Z \in \mathbb{R}^{257 \times 01536}.$$

The token sequence consists of:

- one CLS token;
- 256 patch tokens.

7.2 Global Representation Using the CLS Token

The CLS token provides a global representation of the image: $Z_{\text{CLS}} \in \mathbb{R}_{1536}$. It summarizes information gathered through the Transformer encoder across the image.

7.3 Patch-Context Representation

The rest of the 256 tokens are localizers of image patches. Instead of discarding them, their average was calculated as (2):

$$Z_{\text{patch}} = 1/256 \sum_{i=1}^{256} Z_i. \quad (2)$$

Where $Z_{\text{patch}} \in \mathbb{R}^{1536}$.

7.4 Feature Concatenation

The global CLS and the mean patch representation are concatenated as in (3):

$$Z = Z_{\text{CLS}} \parallel Z_{\text{patch}}. \quad (3)$$

Where $Z \in \mathbb{R}^{3072}$. The resulting representation is integrated with global contextual information and relevant patch level information.

7.5 Lightweight Classification Head

The resulting 3072-dimensional feature vector was normalized using the LayerNorm as in (4).

$$z = \text{layernorm}(z). \quad (4)$$

Which was subsequently processed by a single trainable linear classifier as in (5):

$$O = Wz + b. \quad (5)$$

Where $W \in \mathbb{R}^{4 \times 3072}$.

Softmax converts the four logits into class probabilities as in (6):

$$P(y = k|x) = \frac{\exp(o_k)}{\sum_{j=1}^4 \exp(o_j)}. \quad (6)$$

The final prediction is shown in (7):

$$y^* = \arg \max [P(y = k|x)]. \quad (7)$$

A key benefit in this design is that the size of the foundation-model encoder is not optimized, whereas the lightweight classification head is optimized.

7.6 Training Configuration

A representative Midnight data were extracted once and stored as 3072-dimensional feature vectors for development experiments. The locked final classifier configuration is stated in Table 3.

The final classifier consisted of LayerNorm(3072) → linear(3072,4)

Table 3 The final Midnight-12k classification head's training configuration.

Parameter	Value
Optimizer	AdamW
Learning rate	1×10^{-3}
Weight decay	1×10^{-4}

Loss	Cross-Entropy
Epochs	5
Final classifier batch size	128
Encoder	Frozen

The cross-entropy loss is represented in (8):

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{ic} \log(p_{ic}) \quad (8)$$

Where C=4.

7.7 Strict Five-Fold Evaluation Protocol

The performance of the system developed was assessed through rigorous Five-fold Root-source group aware Cross validation method. Root-source groups were saved for each fold for a different purpose (validation), and the other groups were used for training. Each sample was used only once in a validation fold.

Then all five predictive partitions were concatenated POOF = P1 $\cup$ P2 $\cup$ P3 $\cup$ P4 $\cup$ P5 . Producing Out-of-Fold predictions for each set of the 4,021 training images. ConvNeXt-Tiny and Midnight-12k underwent an identical paired comparison because the exact same fold assignments were used for both models.

8. Evaluation Metrics

The dataset is multiclass and imbalanced, thus multiple complementary measures are used. The multiclass problem was treated as a one-versus-rest problem for all classes.

The accuracy is the overall percentage of correctly classified samples as in (9).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

Precision is the percentage of the samples correctly divided into the expected class, as in (10).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

In the one-versus-rest classification mode, it is the recall with respect to the target class, and in the multiclass classification mode, it is the meaning of the class recalls[15] as in (11).

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP+FN} \quad (11)$$

Specificity evaluates the ability of the classifier to correctly reject samples belonging to other categories[16] as in (12).

$$\text{specificity} = \frac{TN}{TN+FP} \quad (12)$$

The F1-score balances precision and recall as in (13).

$$F1 = 2 \times \frac{\text{PRECISION} \times \text{RECALL}}{\text{PRECISION} + \text{RECALL}} \quad (13)$$

The Matthews correlation coefficient (MCC) is a measure of the quality of binary (two-class) classifications, which ranges from -1 to +1 as in (14). The main advantage of using MCC is its ability to include all four factors of the confusion matrix, which makes it particularly useful for imbalanced classification.

$$MCC = \frac{Tp \times TN - Fp \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (14)$$

Macro F1-score for C class is represented in (15). Each tissue class therefore contributes equally regardless of class frequency.

$$\text{Macro - F1} = \frac{1}{C} \sum_{c=1}^C F1_c \quad (15)$$

The Receiver Operating Characteristic (ROC) curve is a plot of the ratio between sensitivity and false positive rate as the classification threshold changes [17]-[22]. The area under this curve is calculated as in (16).

$$\text{AUC}_{\text{ROC}} = \int_0^1 \text{TPR}(\text{FPR})d(\text{FPR}). \quad (16)$$

The Precision-Recall(PR-AUC) curve is used to evaluate how precision and recall relate to varying thresholds as shown in (17). PR-AUC is especially informative in the presence of class imbalance.

$$\text{AUC}_{\text{PR}} = \int_0^1 \text{Precision}(\text{Recall})d(\text{Recall}). \quad (17)$$

8. Results and Discussion

8.1 Overall Performance Under Strict Five-Fold Evaluation

The effect of the root-source-group-aware five-fold protocol was assessed before the use of the pathway-specific representation. For ConvNeXt-Tiny and Midnight-12k, one test set of fold assignments was used, so that any performance differences could be attributed to the learned representations and not due to differences in data partitioning. The performance of both the out-of-fold (OOF) methods is summarized in Table 4 using the 4,021 development images. The principal classification metrics were improved consistently by midnight-12k.

Table 4 Comparative performance of ConvNeXt-Tiny and Midnight-12k across classification metrics

Metric	ConvNeXt-Tiny	Midnight-12k	Improvement
Accuracy	74.58%	77.24%	+2.66 pp
Macro Precision	70.47%	73.63%	+3.16 pp
Macro Recall	70.43%	73.00%	+2.57 pp
Macro Specificity	91.58%	92.49%	+0.91 pp
Macro F1	70.38%	73.18%	+2.81 pp
Weighted F1	74.05%	76.84%	+2.79 pp
Balanced Accuracy	70.43%	73.00%	+2.57 pp
MCC	0.6509	0.6877	+0.0368

While the absolute improvements seem fairly small, these were reported based on a highly constrained evaluation setting in which related root-source groups are not allowed to be present in the training and validation partitions at the same time. The former is more relevant because the measure of improved MCC takes all of the components of the confusion matrix into consideration, which is not the case for raw accuracy. Fig. 3 shows the comparison of the two models. On all reported pooled measures, Midnight-12k performed better than the baseline.

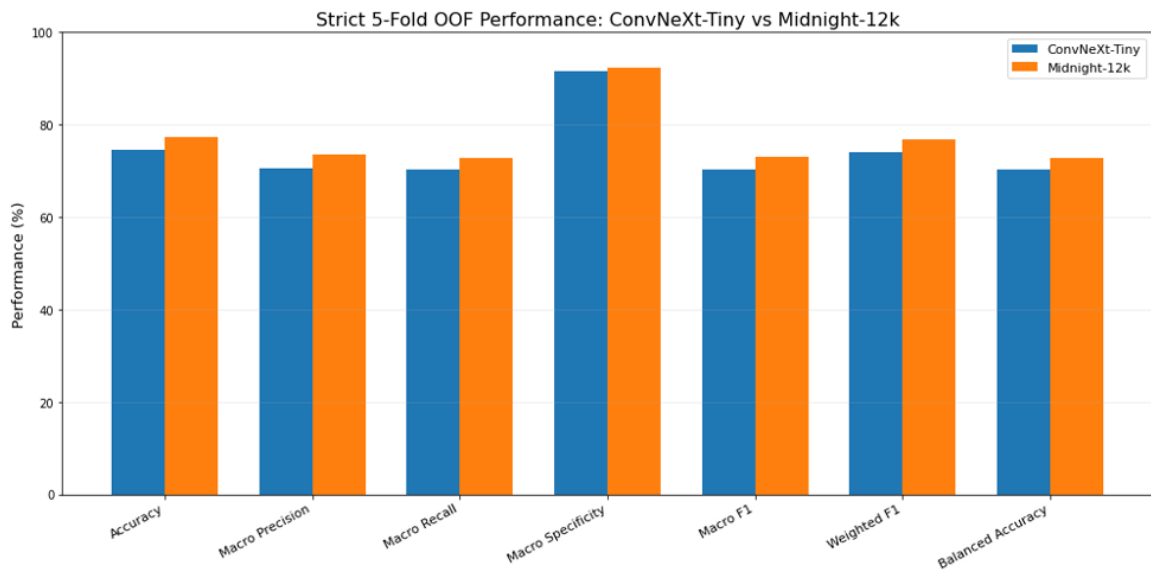


Fig. 3. Comparisons of the ConvNeXt-Tiny model with the Midnight-12k model under the same strict 5-fold root-source-group-aware evaluation using pooled out-of-fold predictions.

8.2 Fold-Level Consistency

Analysis of performance by fold provides a better indication of model improvement than simply an overall look at the pooled performance. As can be observed in Table 5, the Macro-F1 value of Midnight-12k is better than ConvNeXt-Tiny in each and every fold.

Table 5 Five-fold cross-validation accuracy of ConvNeXt-Tiny and Midnight-12k

Fold	ConvNeXt-Tiny	Midnight-12k	Improvement
Fold 1	65.72%	67.14%	+1.42 pp
Fold 2	74.30%	79.28%	+4.98 pp
Fold 3	68.66%	74.92%	+6.26 pp
Fold 4	68.89%	69.68%	+0.79 pp
Fold 5	68.85%	71.69%	+2.84 pp
Mean $\pm$ SD	69.28 $\pm$ 3.11%	72.54 $\pm$ 4.72%	+3.26 pp

The largest gains were in Fold 3 (+6.26 %) and Fold 2 (+4.98 %). Midnight-12k had a positive improvement in all the 3 folders, although least significant in Fold 4. The repeated direction of improvement is critical to note as this suggests that the improvement was not achieved by a single positive partition. In contrast, however, the advantage of the pathology-specific representation was found in all source-group distributions. Additionally, the fold-wise behavior of both models is demonstrated in Fig. 4 where Midnight-12k consistently outperformed the other model in each of the validation folds.

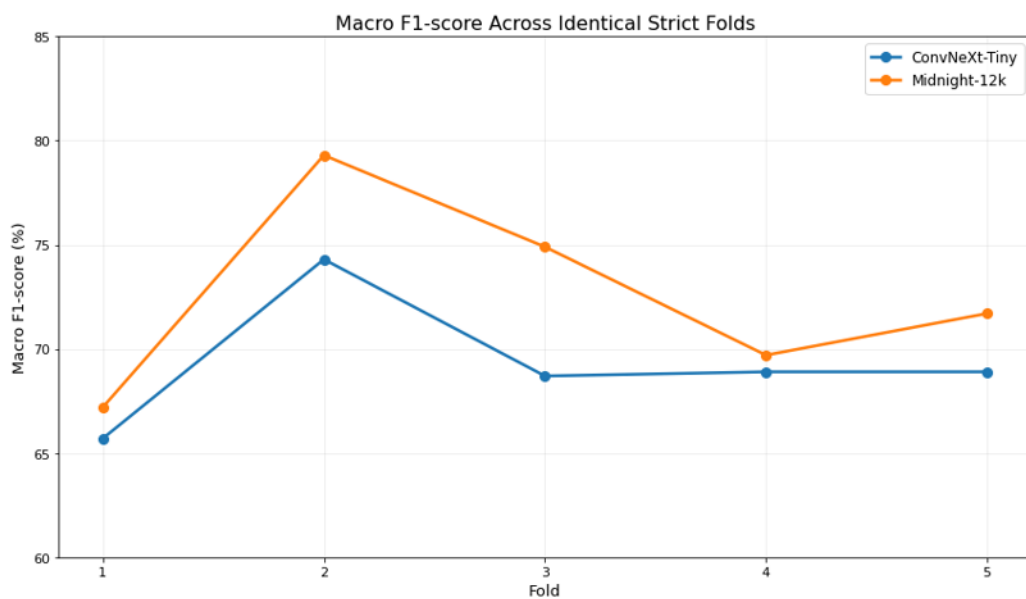


Fig. 4 The folded Macro-F1 scores under identical strict root-source-group-aware validation folds, between ConvNeXt-Tiny and Midnight-12k.

8.3 Class-Wise Performance Improvement

The impact of the foundation-model representation was also assessed independently for every histopathological category. Table 6 shows the F1-scores of these pooled OOF predictions for each class.

Table 6 Class-wise comparison of ConvNeXt-Tiny and Midnight-12k performance, showing relative and absolute improvements across the four histopathology classes

Histopathology Class	ConvNeXt-Tiny	Midnight-12k	Absolute Change	Relative Improvement
Normal	77.05%	78.80%	+1.75 pp	+2.27%
Serrated	48.99%	54.39%	+5.40 pp	+11.02%
Adenocarcinoma	91.47%	94.15%	+2.68 pp	+2.93%
Adenoma	63.99%	65.40%	+1.41 pp	+2.20%

The largest improvement was noticeable within Serrated class where F1 increased from 48.99% to 54.39%. This is equivalent to 5.40 percentage points absolute improvement and an approximately 11.0% relative boost compared to the ConvNeXt baseline. This result is significant as Serrated tissue is one of the most challenging categories in the data set. Adenocarcinoma presented consistently very good classifications for both architectures with an improvement from 91.47% to 94.15% for the Midnight-12k architecture. There was also improvement for tissues in Normal and Adenoma cases. Thus, the overall benefit of Midnight-12k did not limit to the prevalent Adenocarcinoma class. F1 benefited for all four diagnostic categories with the highest gain in the more challenging “Serrated” category.

Although the absolute performance of the Serrated and Adenoma classes improved after Midnight-12k, it remained lower than that of the Normal and Adenocarcinoma classes. There may be a part of this reason because there is a significant morphological overlap between the Serrated tissues and Adenoma tissues, where some of them may have similar glandular structures and architectural pattern. Furthermore, the moderate class imbalance and the uncertainty of the annotations in the Chaoyang dataset may affect the difficulty. Hence, pathology-specific representations enhanced discrimination among all four classes, but failed to fully address the difficulty of discriminating between morphologically similar Serrated and Adenoma tissues.

F1 improvements per class are shown graphically in Fig. 5.

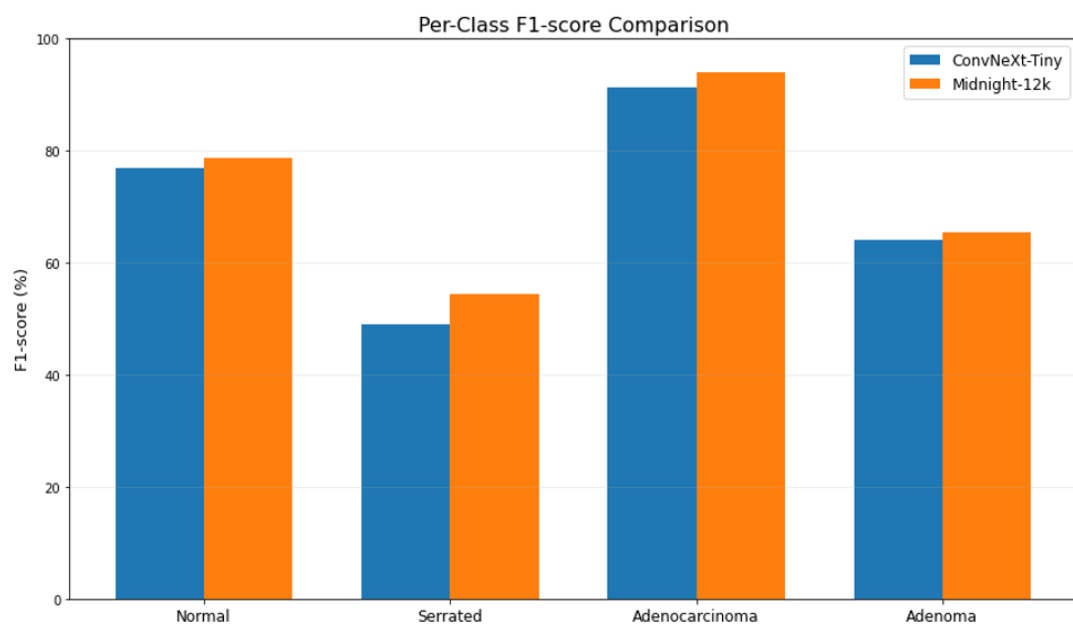


Fig. 5. Class-wise F1-score comparison between ConvNeXt-Tiny and Midnight-12k using pooled strict five-fold OOF predictions.

8.4 Detailed Class-Wise Comparison

A detailed one versus rest analysis gives more insight into the source of the improvement.

Table 6 Class-wise comparison of ConvNeXt-Tiny and Midnight-12k across precision, recall, specificity, F1-score, and PPV, with the corresponding performance change for each histopathology class.

Class	Metric	ConvNeXt-Tiny	Midnight-12k	Change
Normal	Precision	76.88%	75.16%	−1.72 pp
	Recall	77.23%	82.81%	+5.58 pp
	Specificity	91.13%	89.55%	−1.58 pp
	F1-score	77.05%	78.80%	+1.75 pp
	NPV	91.29%	93.17%	+1.88 pp
Serrated	Precision	52.14%	57.12%	+4.98 pp
	Recall	46.20%	51.90%	+5.70 pp
	Specificity	88.77%	89.68%	+0.91 pp
	F1-score	48.99%	54.39%	+5.40 pp
	NPV	86.17%	87.56%	+1.39 pp
Adenocarcinoma	Precision	88.43%	93.11%	+4.68 pp
	Recall	94.73%	95.23%	+0.50 pp
	Specificity	93.35%	96.22%	+2.87 pp

	F1-score	91.47%	94.15%	+2.68 pp
	NPV	97.06%	97.41%	+0.35 pp
Adenoma	Precision	64.43%	69.13%	+4.70 pp
	Recall	63.55%	62.05%	-1.50 pp
	Specificity	93.06%	94.52%	+1.46 pp
	F1-score	63.99%	65.40%	+1.41 pp
	NPV	92.81%	92.64%	-0.17 pp

The results of Midnight-12k were not always a uniform improvement across all class-specific metrics as shown in Table 6. For instance, at Normal recall, there was a large increase, but a small decrease in precision and specificity. Likewise, the precision of Adenoma was enhanced from 64.43% to 69.13% and specificity was enhanced from 93.06% to 94.52% with a small drop in recall. For the Serrated class, precision increased by 4.98 %, recall by 5.70 % and F1 by 5.40 %.

Serrated class showed a clear balanced increase with precision pointing up 4.98 points, recall up 5.70 points and F1 up 5.40 points. The simultaneous improvements show that Midnight-12k successfully lowered false-positive and false-negative from this challenging category.

The highest degree of improvement was seen for Adenocarcinoma, with an improvement of 88.43% to 93.11% precision and an improvement in specificity from 93.35% to 96.22%. This indicates that the specificity of the diagnosis on pathology not only maintained a high sensitivity but was also significantly improved to avoid false Adenocarcinoma diagnosis.

8.5 Statistical Assessment of the Improvement

Paired statistical analysis was applied to see if the improved position occurred consistently across all folds. The differences between the various folds were: +0.0142, +0.0498, +0.0626, +0.0079, +0.0284. Each of the five values was positive.

This finding shows that it was Midnight-12k that was preferred, given improvement in the direction of the folds. The test has limited statistical resolution but, as only five fold pairs were available. So an additional root-source-group-bootstrap was also conducted where the resampling unit was the 68 source groups. The results of the root-source-group bootstrap analysis are summarized in Table 7.

Table 7. Root-source-group bootstrap analysis of the performance difference between Midnight-12k and ConvNeXt-Tiny

Metric	Mean Improvement	95% Bootstrap CI	(P(\Delta>0))
Accuracy	+0.0273	[-0.0068, +0.0629]	0.9376
Macro F1	+0.0301	[-0.0086, +0.0725]	0.9294
MCC	+0.0386	[-0.0081, +0.0897]	0.9400

At the source-group level, the accuracy of Midnight-12k ranged from highest to lowest, with 35 groups achieving higher accuracy, 12 groups achieving the same accuracy, and 21 groups achieving lower accuracy. Therefore, with all three measures, Midnight-12k was supported with probability at or above 0.92 by the bootstrap distributions. However, each 95% confidence interval contained zero. This leads to the consistent observed improvement being substantiated, yet the evidence can't be considered statistically conclusive according to source-group bootstrap.

8.6 Training Behavior of the Final Classification Head

Following development-stage model selection, Midnight-12k was locked and its encoder remained frozen. Only the lightweight classification head was trained on all 4,021 training images for five epochs. Table 8 presents the optimization trajectory.

Table 8. Training behavior of the final Midnight-12k classification head

Epoch	Cross-Entropy Loss	Training Macro F1
1	0.6958	67.24%
2	0.5169	75.94%
3	0.4644	78.54%
4	0.4403	80.10%
5	0.4577	77.92%

The results indicate a significant reduction in training loss from 0.6958 at Epoch 1 to 0.4403 at Epoch 4, while the training Macro-F1 score rose from 67.24% to 80.10%. A slight degradation was noticed at Epoch 5; the loss rises to 0.4577 while the Macro-F1 was reduced to 77.92%. The five epochs were not chosen by looking at the official test set though. It matched the previously locked config. development. For this reason, the minimal Epoch-5 variation was not applied to update the final model. The performance of the last classification head is shown in Figs. 6 and 7. The training cross-entropy loss is shown in Fig. 6, and the training Macro-F1 loss trajectory is shown in Fig. 7.

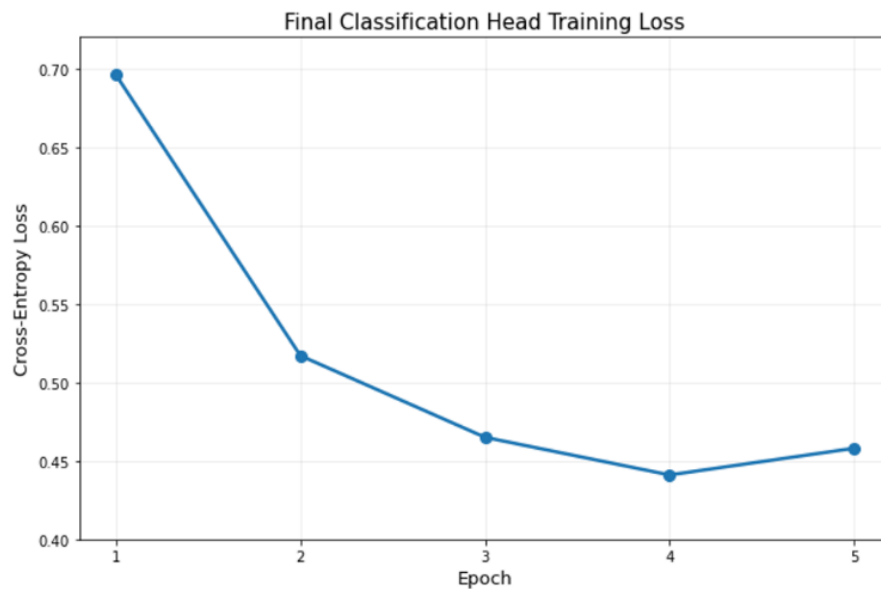


Fig. 6. Cross-entropy training loss of the final lightweight classification head trained on the complete 4,021-image training set. The Midnight-12k encoder remained frozen.

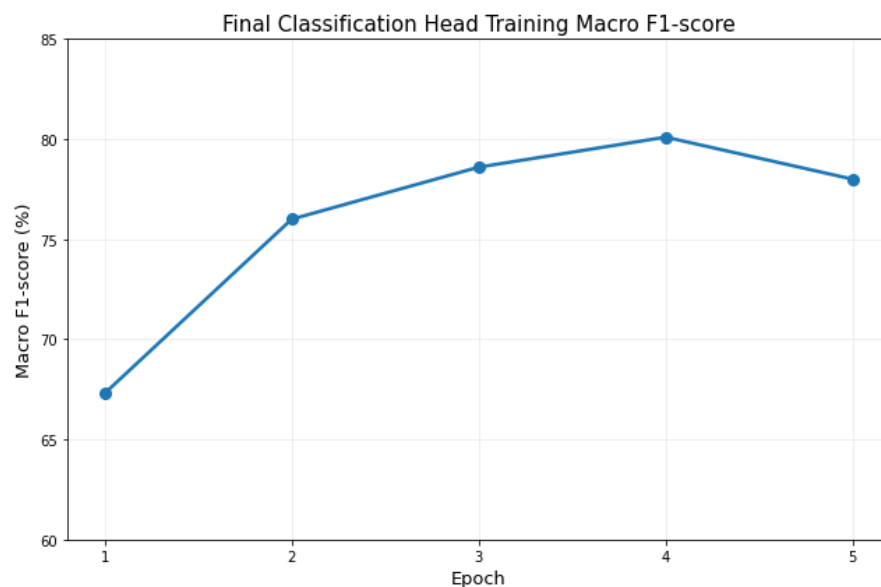


Fig. 7. Training Macro-F1 of the lightweight Midnight-12k classification head during the fixed five-epoch final training stage.

8.7 Performance on the Official Test Benchmark

Once all development phase experiments, model comparisons and configuration decisions were completed, the configuration set for the Midnight-12k pipeline was “locked”. The whole rest of the encoder was frozen while the whole classification head was trained upon the whole 4,021 training images, and the classification model was evaluated once upon the official test images with a test benchmark of 2,139 images. The overall test performance are reported in Table 9.

Table 9. Overall performance on the official test benchmark

Metric	Value
Accuracy	81.39%
Macro Precision	79.73%
Macro Recall	73.47%
Macro F1-score	74.43%
Weighted F1-score	81.50%
Balanced Accuracy	73.47%
Macro Specificity	93.89%
MCC	0.7366
Macro ROC-AUC	94.90%
Macro PR-AUC	83.17%

8.8 Per-Class Official-Test Performance.

Table 10 provides the full class-specific analysis. The best results were seen with Adenocarcinoma. The model had 97.86% accuracy in correctly identifying Adenocarcinoma samples and 95.69% precision and 97.15% specificity. The F1 measured 96.76%, and the ROC-AUC and PR-AUC were 99.60% and 99.28%, respectively. These results show that the remaining categories are nearly completely distinguishable from Adenocarcinoma. The balanced behavior in classification was also noticed in normal tissue with precision of 82.95%, recall of 81.42% and f1-scores of 82.18%. Adenoma showed an interesting 'uneven' pattern. The results were highly accurate (92.20%) and highly specific (99.41%) showing that predictions placed under the label Adenoma were in general reliable. But only 47.62% were retrieved, meaning that many actual samples of Adenoma were put into different categories. The Serrated class was the most challenging overall with F1-score of 55.99%, and PR-AUC of 58.90%. The ROC-AUC however (88.28%) indicates that despite the residual imperfect final multiclass decision boundaries, the learned representations still carried a lot of information class-discriminative.

Table 10. Per-class performance of Midnight-12k on the official test benchmark

Class	Precision	Recall	Specificity	F1	ROC-AUC	PR-AUC
Normal	82.95%	81.42%	91.77%	82.18%	94.86%	87.51%
Serrated	48.10%	66.98%	87.24%	55.99%	88.28%	58.90%
Adenocarcinoma	95.69%	97.86%	97.15%	96.76%	99.60%	99.28%
Adenoma	92.20%	47.62%	99.41%	62.80%	96.86%	86.99%

Overall performance on the class level is illustrated in Fig. 8, with the focus on the particularly strong discrimination of Adenocarcinoma and the lower recall for Adenoma.

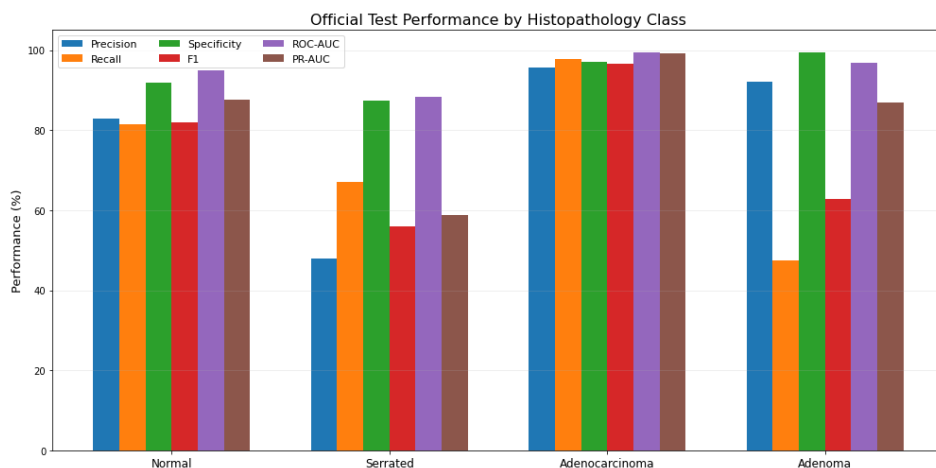


Fig. 8. Precision, recall, specificity, F1-score, ROC-AUC, and PR-AUC for each histopathological class on the locked official test benchmark.

8.9 Confusion-Matrix Analysis

The official-test confusion matrix is shown in Table 11 and Fig. 9. In this test, the observed recall is 97.86%, with 822 Adenocarcinoma samples classified correctly out of 840 cases. Across all the other categories only 18 Adenocarcinoma samples were misclassified. The normal and serrated tissues, however, showed considerable backwards confusion. Of the 705 Normal samples, 115 (16.3%) were foreseen as Serrated. Likewise, 100 out of 321 Serrated samples (31.2%) were determined for Normal. The most noticeable error pattern for Adenoma is that 107/273 (39.2%) samples were erroneously classified as Serrated. The relatively low Adenoma recall (47.62%) can be attributed to only 130 Adenoma samples being correctly recognized. The main limitation does not seem to be the recognition of malignant tissues; that's why this error structure designates this error. Instead, the latter remains as the challenge to differentiate between groups of morphologically similar nonmalignant and precancerous lesions – Serrated and Adenoma.

Table 11. Confusion matrix of Midnight-12k on the official Chaoyang test benchmark.

True \ Predicted	Normal	Serrated	Adenocarcinoma	Adenoma
Normal	574	115	11	5
Serrated	100	215	2	4
Adenocarcinoma	6	10	822	2
Adenoma	12	107	24	130

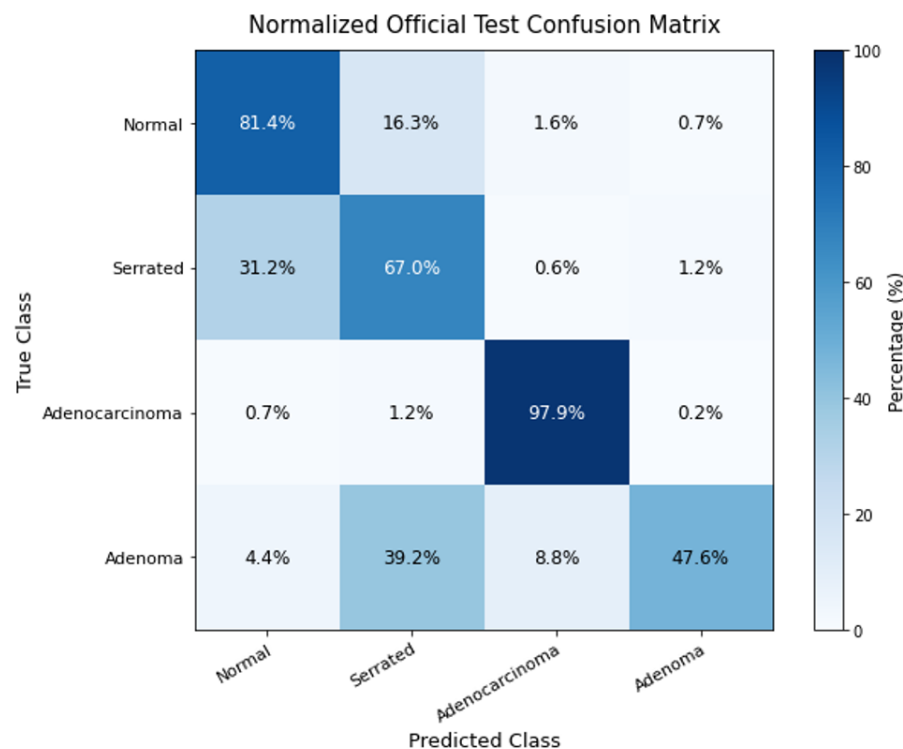


Fig. 9. Normalized confusion matrix of Midnight-12k on the official Chaoyang test benchmark. Values are row-normalized percentages of each true histopathological class. The lower macro-average is explained by the confusion matrix above.

9. Limitations

There are some limitations in this study. First, 4021 training images were available but were from only 68 root-source groups, reducing the number of distinct root-source groups for effective statistical analysis. In addition, these groups were created from file names and are not realized identities of patients. The independence of the official test set could also only be partially verified; thus the results report can only be interpreted as an official test value; external generalization is not possible. This dataset is moderately imbalanced and has morphological overlap between the two tissues (Serrated & Adenoma), causing decreased performances for these two classes. Furthermore, the presented system had a fixed encoder (Midnight-12k), and a full detailed tuning of the encoder was not explored. Another constraint is the lack of external validation with an independent multi-institutional data set. The Chaoyang test partition was not included in the model development, but does not represent clinical generalizability between institutions, scanners, staining protocols or patient populations. As such, the

results reported should be considered as results of a given dataset and not as evidence of clinical deployment readiness. Lastly, one colorectal histopathology dataset was evaluated. Future research should therefore include larger independent multi-institutional samples to further demonstrate the validity of the generalizability and statistical strength of the improvements observed.

10. Conclusion and Future work

This paper presents a study to assess the efficacy of learning pathology-specific baseline representations of four classes of colorectal adherence histopathology in the leakage-resistant root-source-group-aware evaluation protocol. The final framework used a low complexity classification head combined with a frozen Midnight-12k encoder, which enables the use of pathology representations at higher levels without full end-to-end fine-tuning. Overall, Midnight-12k achieved the better performance than the ConvNeXt-Tiny baseline under the same stringent 5-fold testing across all 5 folds. Pooled accuracy increased from 74.58% to 77.24%, and Macro-F1 from 70.38% to 73.18. The results of the fixed model in the official test benchmark were 81.39% accuracy, 81.50% weighted F1-score, 93.89% macro specificity, 94.90% macro-ROC-AUC and 83.17% macro-PR-AUC. Notably, a high F1 score of 96.76% and an ROC-AUC of 99.60% were achieved for Adenocarcinoma. However, the primary challenge left was the discrimination of Serrated and Adenoma tissues. The focus of future research will be to improve the separability of these classes by using more sophisticated feature aggregation, class-aware learning strategies and meticulous fine-tuning of the pathway foundational models. External generalizability will also be important through evaluation of larger, standalone, multi-institutional colorectal histopathology cohorts. Other research opportunities include exploring additional methods for the calibration of models, interpretability, uncertainty estimation, and computationally efficient model adaptation strategies for reliable utilization of foundation models in computer-assisted pathological analysis.

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] World Health Organization, "Colorectal cancer," WHO, Feb. 13, 2026. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer>.
- [2] D. Komura, M. Ochi, and S. Ishikawa, "Machine learning methods for histopathological image analysis: Updates in 2024," *Comput. Struct. Biotechnol. J.*, vol. 27, pp. 383–400, 2025, doi: 10.1016/j.csbj.2024.12.033.
- [3] C. Zhu, W. Chen, T. Peng, Y. Wang, and M. Jin, "Hard sample aware noise robust learning for histopathology image classification," *IEEE Trans. Med. Imaging*, vol. 41, no. 4, pp. 881–894, 2022, doi: 10.1109/TMI.2021.3125459.
- [4] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 11966–11976, doi: 10.1109/CVPR52688.2022.01167.
- [5] R. J. Chen, T. Ding, M. Y. Lu, et al., "Towards a general-purpose foundation model for computational pathology," *Nat. Med.*, vol. 30, pp. 850–862, 2024, doi: 10.1038/s41591-024-02857-3.
- [6] A. Filiot, P. Jacob, A. Mac Kain, and C. Saillard, "Phikon-v2, a large and public feature extractor for biomarker prediction," *arXiv preprint arXiv:2409.09173*, Sep. 2024, doi: 10.48550/arXiv.2409.09173.
- [7] M. Karasikov, J. van Doorn, N. Känzig, M. E. Cesur, H. M. Horlings, R. Berke, F. Tang, and S. Otálora, "Training state-of-the-art pathology foundation models with orders of magnitude less data," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025, Lecture Notes in Computer Science*, vol. 15967. Cham, Switzerland: Springer Nature Switzerland, 2025, pp. 573–583.
- [8] P. C. Neto, D. Montezuma, S. P. Oliveira, et al., "An interpretable machine learning system for colorectal cancer diagnosis from pathology slides," *npj Precis. Oncol.*, vol. 8, Art. no. 56, 2024, doi: 10.1038/s41698-024-00539-4.
- [9] S. B. Yengce-Tasdemir, Z. Aydin, E. Akay, S. Dogan, and B. Yilmaz, "An effective colorectal polyp classification for histopathological images based on supervised contrastive learning," *Comput. Biol. Med.*, vol. 172, Art. no. 108267, Apr. 2024, doi: 10.1016/j.combiomed.2024.108267.
- [10] Z. Fu, Q. Chen, M. Wang, and C. Huang, "Transformer based on multi-scale local feature for colon cancer histopathological image classification," *Biomed. Signal Process. Control*, vol. 100, pt. A, Art. no. 106970, Feb. 2025, doi: 10.1016/j.bspc.2024.106970.
- [11] Q. Ke, Y. C. Hum, W.-S. Yap, T. S. Tan, H. Nisar, H. Mokayed, A. Li, R. Gao, and Y. Gan, "Histopathological classification of colorectal cancer based on domain-specific transfer learning and multi-model feature fusion," *Sci. Rep.*, vol. 15, Art. no. 35155, Oct. 2025, doi: 10.1038/s41598-025-19134-z.

- [12] G. Campanella, S. Chen, M. Singh, et al., “A clinical benchmark of public self-supervised pathology foundation models,” *Nat. Commun.*, vol. 16, Art. no. 3640, 2025, doi: 10.1038/s41467-025-58796-1.
- [13] P. Neidlinger, O. S. M. El Nahhas, H. S. Muti, et al., “Benchmarking foundation models as feature extractors for weakly supervised computational pathology,” *Nat. Biomed. Eng.*, vol. 10, pp. 1113–1123, 2026, doi: 10.1038/s41551-025-01516-3.
- [14] R. Bareja, F. Carrillo-Perez, Y. Zheng, et al., “A benchmark study of vision and pathology foundation models for computational pathology,” *Nat. Commun.*, 2026, doi: 10.1038/s41467-026-76004-6.
- [15] R. Albasrawi, M. Ilyas, and Q. Althebyan, “Modified YOLOv8x model for coronary stenosis detection and troponin risk stratification,” *Discov. Comput.*, vol. 29, Art. no. 34, 2026, doi: 10.1007/s10791-026-09913-1.
- [16] R. Albasrawi, M. Ilyas, and O. Bayat, “A modified YOLO-based deep learning model for accurate ST-elevation myocardial infarction detection in ECG images,” *Neural Comput. Appl.*, vol. 38, Art. no. 617, 2026, doi: 10.1007/s00521-026-12347-x.
- [17] A. B. Salih, R. B. Mohammed, A. A. Othman, S. A. Mostafa, and M. Kamalrudin, “YOLOv8-based real-time pedestrian detection model for urban traffic surveillance systems,” *Int. J. Inform. Vis.*, vol. 10, no. 2, pp. 556–564, Mar. 2026.
- [18] S. I. Hamad, “Utilizing convolutional neural networks for the identification of lung cancer,” *International Journal of Computational and Electronic Aspects in Engineering*, vol. 6, no. 1, pp. 35–41, Mar. 2025, doi: 10.26706/ijceae.6.1.20250206.
- [19] N. Waisi, “Comparative evaluation of ResNet-50 and EfficientNet-B1 for pneumonia detection in chest X-ray images using transfer learning,” *International Journal of Computational and Electronic Aspects in Engineering*, vol. 6, no. 2, pp. 89–97, Jun. 2025, doi: 10.26706/ijceae.6.2.20250405.
- [20] M. K. Yousif, “Enhanced diabetic retinopathy classification using hybrid EfficientNetB0 with CBAM,” *International Journal of Computational and Electronic Aspects in Engineering*, vol. 7, no. 2, pp. 176–188, May 2026, doi: 10.26706/ijceae.7.2.20260307.
- [21] N. Waisi, “Deep feature fusion method for images classification,” *International Journal of Computational and Electronic Aspects in Engineering*, vol. 5, no. 4, pp. 148–153, Dec. 2024, doi: 10.26706/ijceae.5.4.20241103.
- [22] I. S. Seger, “Evaluating and selecting optimal CNN architectures for accurate pneumonia detection in chest X-rays,” *Int. J. Comput. Electron. Aspects Eng.*, vol. 5, no. 4, pp. 183–193, Dec. 2024, doi: 10.26706/ijceae.5.4.20241106. The study compares CNN configurations for medical-image classification.