

Article

Hybrid Loss-Based Deep Learning Framework Using EfficientNet-B3 for Multi-Class Colorectal Cancer Detection

Anusha Nallamalla ^{1,*}  and Chandrakanta Mahanty ^{2,*} 

¹ Department of Computer Science, GSS, GITAM Deemed to be University, Visakhapatnam 530045, India

² Department of Computer Science and Engineering, GSCSE, GITAM Deemed to be University, Visakhapatnam 530045, India

* Correspondence: anallama@gitam.edu (A.N.); cmahanty@gitam.edu (C.M.)

Abstract

Diagnosis of colorectal cancer (CRC) primarily relies on histopathological examination of hematoxylin and eosin-stained tissue sections; however, manual interpretation is time-consuming, subjective, and increasingly impractical given the rapid growth of digital pathology data. We introduced a hybrid loss-based learning framework for multi-class colorectal histopathology image classification that improves class-balanced performance without increasing model complexity. Various EfficientNet versions were checked as the first step to establishing a strong baseline, and EfficientNet-B3 was chosen based on validation Matthews Correlation Coefficient (MCC). Extending this backbone, we propose a hybrid loss function that mixes weighted cross-entropy and focal loss to achieve the combined effect of dealing with the global class imbalance while also focusing on hard-to-classify samples. The results of experiments on a large-scale colorectal histopathology dataset show that the Hybrid-B3 model introduced significantly improves the baseline settings. Hybrid-B3 registers a test accuracy of 99.83%, a very high class-balanced performance with a balanced accuracy and G-Mean of 99.85%. The changes are verified and non-random by the statistical validation using bootstrap confidence intervals and paired significance tests. The offered solution emphasizes the efficiency of loss-function optimization solely to provide improvements in robustness and reliability in computational pathology and, correspondingly, yields a practical and scalable solution for colorectal cancer diagnostic support in the real world.

Keywords: colorectal histopathology; deep learning; EfficientNet; hybrid loss function; class imbalance



Academic Editors: Zubaer Mannan, Asif Karim and Nur Alam Md

Received: 24 February 2026

Revised: 5 April 2026

Accepted: 8 April 2026

Published: 16 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Colorectal cancer (CRC) continues to be one of the main causes of death from cancer worldwide. The follow-up of the patients is highly dependent on early and precise diagnosis [1]. The histopathological examination of hematoxylin and eosin (H&E) stained tissue sections is the diagnostic method for CRC; however, manual evaluation requires a lot of time, is subjective, and there is a high risk of inter-observer variability, especially in large-scale screening situations.

Due to the increase in digital whole-slide images in modern pathology workflows, a reliable, automated, and scalable computer-aided diagnostic system that assists pathologists in developing clinical trials through accurate identification of diverse colorectal tissue types is critically needed [2,3]. In the recent past, deep learning has significantly impacted

computational pathology through enabling end-to-end learning of tissue discriminative features directly from histopathological images using convolutional neural networks (CNNs). CNNs have been found to be efficient in colorectal tissue classification, segmentation, and prognosis prediction tasks. VGG, ResNet, DenseNet, and EfficientNet are some of the architectures that have been extensively used because of their excellent feature extraction potential and the possibility of transfer learning. Among all, EfficientNet-based models have been particularly recognized for delivering high accuracy in classification at a significantly low computational cost via compound scaling of the network depth, width, and resolution [4,5].

Beyond the architectural refinements, some works have also delved into attention mechanisms, ensemble learning, weakly supervised learning, and transformer-based models in order to recognize complex morphological patterns as well as long-range contextual dependencies in histopathological images. Such techniques have pushed the limit of acquiring knowledge; however, many of them call for a higher degree of model complexity, significant computing power, or the use of specialized preprocessing pipelines. Besides that, even though very high accuracy levels have been attained, one of the main problems in colorectal histopathology analysis is the extreme class imbalance that comes naturally with real-world datasets, where the samples of some tissue types (e.g., tumor and muscle) significantly outnumber the samples of other classes like mucus, lymphocytes, and debris [6].

In order to fix the class imbalance problem, the majority of the studies at present adopt weighted cross-entropy loss or data-level balancing strategies. Weighted cross-entropy is a good solution for adjusting the differences in global class frequencies; however, it assumes that all samples belonging to a class have the same level of difficulty and it does not put a particular emphasis on hard-to-classify or ambiguous tissue regions. For histopathological images, such hard samples are indeed very frequent due to factors like slight morphological differences, staining artifacts, and tissue characteristics that partially overlap. Therefore, models trained only with the weighted cross-entropy loss may happily reach high accuracy but still fail on the samples that are problematic from the clinical point of view.

Several recent publications have focused on introducing focal loss and other margin-based loss functions to put more emphasis on hard samples and to boost the performance of minority classes. Nevertheless, focal loss in isolation might concentrate too much on difficult samples at the cost of maintaining the overall class balance, especially in multi-class medical imaging situations. Moreover, a lot of research works integrate new loss functions with architectural changes, which makes it hard to determine if the performance improvements come from better learning strategies or from increased model capacity.

Thus, we set out to investigate this issue and propose here a hybrid loss-based learning framework that is built on a carefully chosen and well-established backbone CNN. Initially, we perform a neutral real-world experiment of several EfficientNet variants to isolate the best single baseline architecture. Based on validation scores under imbalance-aware metrics, EfficientNet-B3 is selected as the reference backbone. Then, we propose a hybrid loss function combining weighted cross-entropy and focal loss, thus enabling one to deal with both the global class imbalance and the targeted focus on hard-to-classify samples simultaneously. Most significantly, the proposed framework keeps the original network architecture intact, which means that the performance improvements observed are solely due to learning dynamics that have been enhanced.

The key idea that we have developed and tested in this paper is that combining weighted cross-entropy and focal loss in a single optimization framework can lead to improved class-balanced performance and robustness in colorectal histopathology clas-

sification, even from a very powerful baseline model. Our extensive experiments on a large-scale, clinically relevant dataset, using imbalance-aware evaluation metrics and rigorous statistical validation, confirm this idea.

The main contributions of our work are summarized below:

We show that EfficientNet-B3 trained with weighted cross-entropy can be considered a very powerful and fair baseline for colorectal histopathology classification.

- We introduce a hybrid loss function that tackles both global class imbalance and hard-sample learning jointly without changing the network architecture.
- We proved through our experiments that significant and consistent improvements in MCC, balanced accuracy, and G-Mean can be achieved only by loss-level optimization.
- We have provided a statistical analysis that supports our claim that the proposed method provides robust and reproducible performance improvements even in very severe class imbalances.

2. Literature Survey

Ramasamy et al. [7] have presented DBNA-Net, a hybrid Deep Belief Network and NASNet-based attention framework for histopathological images of lung and colon cancer classification. The research introduced unsharp filtering, PN-UNet-based cell segmentation, and handcrafted texture features to enhance discriminative learning. The experimental outcomes have confirmed a strong classification performance with over 91% accuracy for different cancer types. The authors have found that attention-guided deep architectures enhance the robustness in the recognition of heterogeneous tissue patterns. Ochoa-Ornelas et al. [8] experimented with a transfer learning framework based on EfficientNet-B3 for the automatic detection of lung and colon cancer from histopathological images. The model was initially trained on the LC25000 dataset and then further tested on external genomic data to evaluate its generalizability. This approach has achieved a very high level of performance in terms of accuracy, precision, recall, and MCC, all values being above 99%. EfficientNet-B3 was pointed out as an effective end-to-end model that requires very little preprocessing. Fu et al. [9] proposed a colon cancer histopathological image classification method based on transformers, which leverages local multi-scale features and makes use of whole-slide images. The technique employed farthest point sampling and multi-scale grouping to effectively capture the regional contextual dependencies among the different patches. To counter class imbalance in weakly supervised settings, the focal loss function was used. Tests carried out on the hospital and TCGA datasets have shown that the approach outperforms other MIL-based methods. Hamida et al. [10] have designed weakly supervised attention-guided UNet variants for colon cancer histopathological image segmentation with insufficient annotation. The research modified Att-UNet architectures by innovative skip-connection and attention-gate configurations. A single-step training protocol has been proposed to deal with class imbalance and lack of pixel-level annotations. The newly developed Alter-AttUNet has been able to score high segmentation accuracy while still being a lightweight model.

Muniz et al. [11] investigated the combination of micro-FTIR hyperspectral imaging and deep learning to go beyond standard RGB histopathology for colon cancer diagnosis. They used fully connected deep neural networks to model hyperspectral voxel representations and thus capture biochemical tissue signatures. Their method reached 99% classification accuracy in a cross-validation experiment within a single patient. The paper showed that infrared spectral data were a major contribution to tissue classification and, therefore, diagnostic confidence. Dabass et al. [12] suggested a multi-level convolutional neural network with the addition of enhanced convolutional learning modules and attention learning mechanisms for colon histopathological image classification. Such

a system was able to simultaneously extract spatial and channel-wise discriminative features, and at the same time, it successfully avoided vanishing gradient and resolution degradation problems. Evaluation was made on a number of public and private datasets, such as LC25000 and NCT-CRC-100K. The experimental outcomes supported the findings of high accuracy and the attention maps, which were clinically validated, matched the pathologists' interpretations. Yıldız et al. [13] designed a stacked-based deep ensemble learning method for a multi-class cancer diagnosis with histopathological images. They used DenseNet201 and EfficientNet-B7 as the base learners, while a shallow CNN was utilized as a meta-learner to fuse the decisions. The model was subjected to training and testing on the datasets of lung, colon, oral, and breast cancer. The ensemble model was at the top of its game by delivering near-perfect accuracy for colon and lung cancer, thus indicating its high potential for computer-aided diagnosis. Ghadami et al. [14] presented a hybrid colon cancer diagnosis model integrating a modified VGGNet-based CNN with the Coati Optimization Algorithm for feature selection. Instead of combining PCA-based conventional selection, they hence utilized a bio-inspired optimization method to pick out the most discriminative features. Classifiers such as decision trees, KNN, and ensemble methods were among those tested on the LC25000 dataset. The optimized CNN-COA model had the highest accuracy and statistical significance, which proved its potential for the automatized diagnosis.

Palomarde Lucas et al. [15] proposed a tumor budding–stroma (TBS) score that is capable of providing more detailed information for prognostic assessment in microsatellite-stable localized colon cancer. The authors scored both tumor budding and tumor-associated stroma in the extended tumor areas of resected specimens. The combined TBS score not only showed a strong independent association with disease-free survival but also with immune-depleted tumor microenvironments. The authors' method surpassed hotspot-based assessments in terms of effectiveness, thereby providing a fuller picture of tumor heterogeneity. Mehmood et al. [16] demonstrated a transfer-learning-based method for the detection of lung and colon cancer using histopathology images paired with class-selective image processing. A fine-tuned AlexNet architecture was used, with targeted contrast enhancement being a feature of only the underperforming classes. Such a selective preprocessing strategy allowed for an increase in the classifiers' accuracy while not compromising on computational efficiency. The method the authors proposed resulted in a noticeably better overall accuracy and was thus a practical solution for diagnostic setups with very limited resources. Sari et al. [17] came up with histopathological colon tissue classification method via unsupervised learning of deep feature extractions with the help of restricted Boltzmann machines. The technique selected the most important tissue subregions by combining domain knowledge and without the help of pixel-level annotations. In addition, the authors also used the unsupervised clustering of the deep features to find out the different types of tissue components and thus, perform the classification task. The experimental results also showed that the accuracy was better than that obtained through the use of handmade and supervised patch-based approaches. Jia et al. [18] published in their paper a constrained, deep weak supervision setting-based deep learning framework under a multiple instance learning paradigm for histopathology image segmentation. The authors' model made use of fully convolutional networks incorporating deep weak supervision and area constraints, which were derived from coarse annotations. Pixel-level segmentation was realized with only image-level labels and approximate region size information. The method proposed by the authors resulted in a performance that was at the cutting-edge of segmentation on histopathology datasets of big scale.

Sirinukunwattana et al. [19] proposed a locality-sensitive deep learning method to detect and distinguish nuclei in histology images of colorectal cancer. Specifically, they

developed a spatially constrained CNN for regressing the nucleus center probabilities and a neighboring ensemble predictor for enhancing classification accuracy. Their approach did not require explicit nucleus segmentation, thus reducing the computational burden. They compared their method with others on extensively annotated large-scale datasets and showed that their method outperformed all others in terms of F1-scores. Mahbub et al. [20] suggested a center-focused affinity loss function to tackle the problem of class imbalance in fine-grained histopathology image classification. Their loss function promoted intra-class compactness and focused on minority class feature learning in the embedding space. The method was tested on breast and colon cancer datasets that suffered from severe imbalance. The experimental results confirmed that their method consistently outperformed softmax, ArcFace, CosFace, and focal loss-based approaches. Belharbi et al. [21] presented an interpretable weakly supervised framework for histology image classification and segmentation guided by a max-min uncertainty principle. Their model was composed of cross-entropy, KL-divergence-based uncertainty regularization, and a log-barrier constraint that accounted for the balancing of foreground and background regions. The authors showed that pixel-level localization was feasible by using only image-level labels without resorting to exhaustive annotations. Their method drastically lowered false positive rates and was ahead of the current state-of-the-art weakly supervised methods on colon and breast cancer datasets. Alqahtani et al. [22] put forward a deep learning-based framework for cancer prognosis prediction that combined feature selection and feed-forward neural networks. A binary AC-parametric whale optimization algorithm was utilized to pick the best feature subsets and tune the network parameters. The proposed framework was tested on different cancer datasets, such as breast, colon, lung, and ovarian cancers. The results revealed that the proposed method was more accurate and less time-consuming in making predictions than the conventional machine learning techniques. Provath et al. [23] built a multi-dataset training strategy to achieve robust colorectal polyp detection with state-of-the-art YOLO architectures. They used heterogeneous datasets like Kvasir-SEG and CVC-ClinicDB to train the model in order to make it generalize well to the variations in polyp appearances. Moreover, the robustness in real clinical situations was evaluated via external validation on unseen datasets. The study has shown that the proposed method is highly precise and real-time is feasible for endoscopic diagnostic support.

Bappi et al. [24] came up with a better YOLO-based detection framework that uses attention mechanisms and specially tailored loss functions to analyze biomedical images. The proposed model concentrated on enhancing feature discrimination of small and irregular shapes in medical images. They carried out a wide range of tests, which showed that not only was the accuracy of the detections improved, but also the number of false detections was lowered. The authors further remarked that the changes in the network architecture have the potential to make it suitable for real-time scenarios in clinics. Elseddeq et al. [25] have merged super-resolution and object detection models with a hybrid deep learning method to tackle the problem of poor-quality medical images. This model uses the resolution enhancement features of super-resolution to overcome the issue of severe resolution degradation that is characteristic of images taken with an endoscope. They performed the evaluation of the performance using three different metrics, i.e., precision, recall, and mean average precision. Furthermore, the outcome of the experiment confirmed that the addition of super-resolution resulted in a significant increase in the dependability of the detections. Mert et al. [26] developed a colorectal polyp detection system capable of working in real time by utilizing deep learning along with preprocessing for image enhancement. In order to improve the identification of features, contrast enhancement techniques were first applied before object detection. The deep neural network was tested on colonoscopy datasets that are publicly available and reflect true situation problems.

The research data showed that not only was there an improvement in F1-scores, but also there was an increased stability of the method against changes in lighting. Table 1 summarizes representative recent studies in colorectal histopathology analysis, highlighting their methodologies, datasets, performance metrics, key contributions, and reported limitations.

Table 1. Comparative summary of various approaches for colorectal histopathology analysis.

Ref No.	Method/Model	Dataset	Accuracy (%)	Key Contribution	Limitations
[7]	DBNA-Net (DBN + NASNet with Attention)	LC25000	91.3	Combined deep belief networks with attention-based CNNs for robust histopathology classification.	Relied on handcrafted features and segmentation; limited external validation.
[8]	EfficientNet-B3 (Transfer Learning)	LC25000	99.3	Demonstrated strong generalization using lightweight transfer learning with minimal preprocessing.	Limited explainability; focused only on classification.
[9]	Transformer-based MIL with Multi-scale Grouping	TCGA, Hospital WSI	94.1	Captured long-range dependencies in WSIs using multi-scale transformer learning.	Computationally expensive; required large memory resources.
[10]	Alter-AttUNet (Weakly Supervised)	Colon Histology Dataset	93.5	Introduced attention-guided UNet variants for segmentation under sparse annotation.	Sensitive to class imbalance; limited robustness to noise.
[11]	Deep Neural Network + Micro-FTIR	Clinical Hyperspectral Data	99	Used biochemical spectral signatures beyond RGB imaging for cancer diagnosis.	Specialized imaging hardware required; limited scalability.
[12]	Multi-level CNN with Attention	LC25000, NCT-CRC-100K	98.6	Extracted hierarchical spatial and channel-wise features for histopathology classification.	Model complexity increased training cost.
[13]	Stacked Ensemble (DenseNet201 + EfficientNet-B7)	Multi-cancer Histopathology	99.8	Achieved near-perfect accuracy using deep ensemble stacking.	High computational overhead; poor interpretability.
[14]	CNN + Coati Optimization Algorithm	LC25000	97.9	Introduced bio-inspired optimization for feature selection in CNNs.	Optimization stage increased inference time.
[15]	Tumor Budding–Stroma (TBS) Scoring	MSS Colon Cancer Data	–	Proposed prognostic stratification using extended tumor region analysis.	Not a DL-based model; dependent on expert annotation.
[16]	AlexNet + Class-Selective Enhancement	LC25000	96.4	Improved underperforming class accuracy via selective preprocessing.	Performance depended on enhancement tuning.

Table 1. *Cont.*

Ref No.	Method/Model	Dataset	Accuracy (%)	Key Contribution	Limitations
[17]	Unsupervised Deep Feature Learning (RBM)	Colon Histopathology	89.7	Learned discriminative tissue features without pixel-level labels.	Lower accuracy than supervised deep models.
[18]	Weakly Supervised FCN (MIL-based)	Large-scale Histopathology	94.8	Achieved pixel-level segmentation using only image-level labels.	Required approximate region constraints.
[19]	Locality-Sensitive CNN	Colorectal Nuclei Dataset	95.0 (F1)	Avoided explicit segmentation while improving nucleus detection accuracy.	Sensitive to densely clustered nuclei.
[20]	Center-Focused Affinity Loss	Breast and Colon Cancer	96.2	Addressed class imbalance using novel loss formulation.	Required careful hyperparameter tuning.
[21]	Weakly Supervised Uncertainty-Guided CNN	Colon and Breast Histology	94.5	Introduced uncertainty regularization for interpretable localization.	Localization accuracy lower than fully supervised methods.
[22]	AC-Parametric Whale Optimization + ANN	Multi-cancer Tabular Data	>95.0	Optimized feature selection and parameters for prognosis prediction.	Not image-based; limited explainability.
[23]	YOLOv8m–YOLOv11m	Kvasir-SEG, CVC-ClinicDB, ETIS	mAP: 91.1	Demonstrated robust real-time polyp detection with multi-dataset training.	Limited performance on flat or very small polyps.
[24]	Attention-Enhanced YOLOv7	Biomedical Challenge Data	98.8	Improved small-object detection using attention and custom loss.	Limited clinical dataset validation.
[25]	SRGAN + YOLOv5	ETIS, CVC-ClinicDB	95.2	Enhanced detection from low-resolution endoscopic images.	Increased computational complexity.
[26]	YOLOv8 + CLAHE Enhancement	Kvasir-SEG, HyperKvasir	F1: 92.6	Improved robustness via contrast enhancement preprocessing.	Noise amplification under poor lighting conditions.

In summary, these studies have made significant progress in colorectal histopathology analysis by utilizing advanced CNNs, transformer-based models, ensemble learning, weak supervision, and novel optimization strategies. There are benefits to many of these methods, such as a high level of accuracy; however, they usually come at the cost of increased architectural complexity, highly specialized preprocessing, or high computational power requirements. In addition, they might not even solve the problem of severe class imbalance to the extent that is expected or provide sufficient statistical validation. While it is true that alternative loss functions have been tested, only a tiny fraction of the works systematically search whether loss-function optimization alone can lead to improvements that are reliable and consistent over different datasets without increasing the model complexity. Such shortcomings call for a method that is not only computationally efficient but also imbalance-aware and statistically robust. Thus, the hybrid loss-based EfficientNet method proposed in this work is motivated.

3. Materials and Methods

3.1. Baseline Models

The EfficientNet family of convolutional neural networks, namely EfficientNet-B0, EfficientNet-B3, and EfficientNet-B5, was used in this research to develop the baseline models. To improve the results, all baseline models were first initialized with ImageNet-pretrained weights and then fine-tuned on the colorectal histopathology dataset using the transfer learning approach. The original classification layers were removed and replaced with task-specific fully connected layers of nine tissue classes. Because histopathological data have a class imbalance, the training of all baseline models was done with weighted cross-entropy loss, where class weights were calculated based on the distribution of training data. To allow a fair and unbiased comparison, training and evaluation were conducted with the same experimental settings.

EfficientNet networks were selected as they are known for their performance and parameter efficiency resulting from compound scaling of network depth, width and input resolution. EfficientNet models, when compared to traditional CNNs, are a great compromise between accuracy and computational cost, which makes them very suitable for large-scale medical image analysis. Trying out several EfficientNet variations allows for a fair backbone choice and prevents architectural bias. EfficientNet-B3 out of the three models gave the best validation results in terms of MCC, and hence the baseline reference model for the rest of the experiments was chosen to be EfficientNet-B3.

EfficientNet-B0 was thoroughly tested as a minimal and efficient model, mainly to set a complexity and performance baseline at a low end. EfficientNet-B0, which was originally trained on the ImageNet dataset, was retrained on the colorectal histopathology dataset using transfer learning. The model architecture was slightly tweaked to enable the final classification of the nine tissue types. The model was optimized using a weighted cross-entropy loss in order to account for the imbalance in tissue class distribution. EfficientNet-B0's relatively small size in terms of the number of parameters makes it a good candidate for investigating how much performance we can sacrifice for a significant gain in efficiency when classifying large-scale histopathological images [27].

EfficientNet-B3 was the choice for a medium-sized baseline, allowing the model to have a good level of capacity, while still being efficient in terms of computations. In the same manner as with EfficientNet-B0, the model state was loaded with ImageNet-pretrained weights and subsequently retuned using the weighted cross-entropy loss technique under a clone of the initial experimental conditions. EfficientNet-B3 scored the highest validation results through the Matthews Correlation Coefficient (MCC) metric, hence indicating the ability of the model to learn balanced class distributions. Therefore, it was chosen as the main baseline model, labeled as Baseline-B3, and also acted as the backbone for the hybrid loss framework that was proposed [28].

EfficientNet-B5 was measured against the capacity increase in the network so as to evaluate its effect on classification accuracy. The model was fine-tuned similarly to the other baseline models, i.e., using transfer learning and weighted cross-entropy loss. EfficientNet-B5 managed to achieve decent accuracy and MCC scores. However, its performance fell short of the EfficientNet-B3 model even though it was more computationally expensive. This basically highlights that merely increasing the depth or complexity of a network architecture without proper class balancing will hardly lead to improvements in class-balanced performance for colorectal histopathology classification [29].

3.2. Proposed Hybrid-B3 Model

The Hybrid-B3 model in the pipeline extends the EffectiveNet-B3 focus, which was selected via objective baseline evaluation, and it also features a somewhat new loss for-

mulation to better accommodate class-balanced and hard-sample learning. The network comp., initialization scheme, data handling, and the solution method of the Baseline-B3 remain the same. This is done to ensure that any performance gains can only be attributed to the learning strategy and not the architectural changes.

The weighted cross-entropy loss, which adjusts the global class imbalance but considers all the samples within the same class equally, is not sufficient for the weighted cross-entropy loss's deficiencies to be fixed. The Hybrid-B3 model combines the weighted cross-entropy with the focal loss to form a hybrid loss function. The weighted cross-entropy encourages less-represented tissue classes through the use of class weighting, while focal loss lessens the weight of easy samples and concentrates the learning on hard-to-classify and ambiguous tissue regions that are common in histopathological images. The hybrid loss is a mixture of these two components, a balancing coefficient, and a focusing parameter, with weights assigned to each.

The hybrid loss provides two flexible hyperparameters: the weighting coefficient, which balances the weighted contribution of cross-entropy and focal loss, and the focusing parameter, which indicates how much emphasis is given to the hard samples. Random search was used to optimize these parameters, with the validation MCC serving as the guide to perform imbalance-aware model selection. The final model of Hybrid-B3 was run through the same experimental protocol as the baseline after it had been trained with the best parameter setting.

Importantly, the Hybrid-B3 framework does not add model parameters, extend the network depth, or increase the computational complexity. Hence, the new method attains enhanced class-balanced performance and robustness simply by loss-level optimization, which makes it not only an effective but also a practical method for large-scale colorectal histopathology analysis.

Table 2 compares the different EfficientNet variants in terms of their architectural features, which have been evaluated in this research. EfficientNet-B3 is the most capable model for representational capacity and it is expected that it requires the least amount of computational power. Hence, the selection of EfficientNet-B3 as the backbone of the proposed hybrid loss framework is well justified. Notations are mentioned in Table 3.

Table 2. Comparative architectural characteristics of evaluated models.

Model	Depth/Layers	Parameters (Approx.)	Key Components	Strengths	Limitations
EfficientNet-B0	Shallow–moderate depth	~5.3 million	MBConv blocks, squeeze-and-excitation, compound scaling	Lightweight, fast training and inference, low computational cost	Limited representational capacity for complex tissue patterns
EfficientNet-B3	Moderate depth	~12 million	MBConv blocks, squeeze-and-excitation, compound scaling	Strong balance between accuracy and efficiency, robust feature representation, best class-balanced performance	Slightly higher computational cost than B0
EfficientNet-B5	Deeper architecture	~30 million	MBConv blocks, squeeze-and-excitation, compound scaling	Higher capacity for fine-grained feature learning	Increased computational cost without proportional performance gain

Table 3. Notation used in the mathematical model.

Symbol	Symbol Description
$\mathcal{D}$	Histopathology dataset
N	Number of samples
x_i	Input histopathology image
y_i	Ground-truth class label
$\hat{y}_i$	Predicted class label
C	Number of tissue classes ($C = 9$)
$f_\theta(\cdot)$	EfficientNet-B3 feature extractor
θ	Learnable model parameters
$\mathbf{z}_i$	Feature representation of sample i
$p_{i,c}$	Predicted probability for class c
w_c	Class weight for class c
$\mathcal{L}_{\text{WCE}}$	Weighted cross-entropy loss
$\mathcal{L}_{\text{FL}}$	Focal loss
λ	Loss balancing coefficient
γ	Focal loss focusing parameter

3.3. Mathematical Model of the Proposed Framework

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote the colorectal histopathology dataset, where $x_i \in \mathbb{R}^{H \times W \times 3}$ represents an input RGB image patch and $y_i \in \{1, 2, \dots, C\}$ denotes the corresponding ground-truth tissue class label, with $C = 9$.

Feature Extraction and Prediction.

Given an input image x_i , the EfficientNet-B3 backbone acts as a nonlinear feature extractor parameterized by θ , producing a feature representation $\mathbf{z}_i$:

$$\mathbf{z}_i = f_\theta(x_i) \quad (1)$$

The extracted features are passed through a fully connected classification layer followed by a softmax function to obtain class probabilities:

$$p_{i,c} = \frac{\exp(\mathbf{w}_c^\top \mathbf{z}_i)}{\sum_{k=1}^C \exp(\mathbf{w}_k^\top \mathbf{z}_i)}, c = 1, \dots, C \quad (2)$$

where $p_{i,c}$ denotes the predicted probability that sample x_i belongs to class c , and $\mathbf{w}_c$ represents the classifier weights for class c .

The final predicted class label is given by the following:

$$\hat{y}_i = \arg \max_c p_{i,c} \quad (3)$$

Weighted Cross-Entropy Loss.

To address global class imbalance, weighted cross-entropy loss is employed. Let w_c denote the weight associated with class c , computed from the training set distribution. The weighted cross-entropy loss is defined as follows:

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N} \sum_{i=1}^N w_{y_i} \log(p_{i,y_i}) \quad (4)$$

Focal Loss.

To emphasize hard-to-classify and ambiguous samples, focal loss is incorporated. The focal loss is defined as follows:

$$\mathcal{L}_{\text{FL}} = -\frac{1}{N} \sum_{i=1}^N (1 - p_{i,y_i})^\gamma \log(p_{i,y_i}) \quad (5)$$

where $\gamma \geq 0$ is the focusing parameter controlling the degree of emphasis on misclassified samples.

Proposed Hybrid Loss Function.

The proposed hybrid loss combines weighted cross-entropy and focal loss into a single objective function:

$$\mathcal{L}_{\text{Hybrid}} = \lambda \mathcal{L}_{\text{WCE}} + (1 - \lambda) \mathcal{L}_{\text{FL}} \quad (6)$$

where $\lambda \in [0, 1]$ balances global class imbalance handling and hard-sample learning.

Model parameters are optimized by minimizing the hybrid loss:

$$\theta^* = \arg \min_{\theta} \mathcal{L}_{\text{Hybrid}} \quad (7)$$

The training procedure of the proposed Hybrid-B3 framework is presented in Algorithm 1.

Algorithm 1: Training Procedure of the Proposed Hybrid-B3 Framework

Input:

Histopathology dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, number of classes C , class weights $\{w_c\}_{c=1}^C$, EfficientNet-B3 backbone, hyperparameters λ and γ

Output: Optimized model parameters θ^*

1. Initialize EfficientNet-B3 with ImageNet-pretrained weights.
 2. Replace the final classification layer to match the C -class output.
 3. For each training epoch, do:
 - a. Extract feature representations using Equation (1).
 - b. Compute class probabilities using Equation (2).
 - c. Predict class labels using Equation (3).
 - d. Compute weighted cross-entropy loss using Equation (4).
 - e. Compute focal loss using Equation (5).
 - f. Combine losses using the hybrid formulation in Equation (6).
 4. Update model parameters by minimizing the hybrid loss as defined in Equation (7).
 5. Repeat Steps 3–4 until convergence.
 6. Return the optimized parameters θ^* .
-

3.4. Workflow Design

Figure 1 illustrates the overall architecture of the imbalance-aware framework proposed for multi-class colorectal histopathology classification. The framework receives input histology images that are passed through multiple EfficientNet variants for systematic backbone evaluation to select the best feature extractor based on the validation performance objectively. The chosen backbone is then combined with a feature extraction and classifier head and trained with a hybrid loss, made up of weighted cross-entropy and focal loss to simultaneously tackle class imbalance and hard-sample learning.

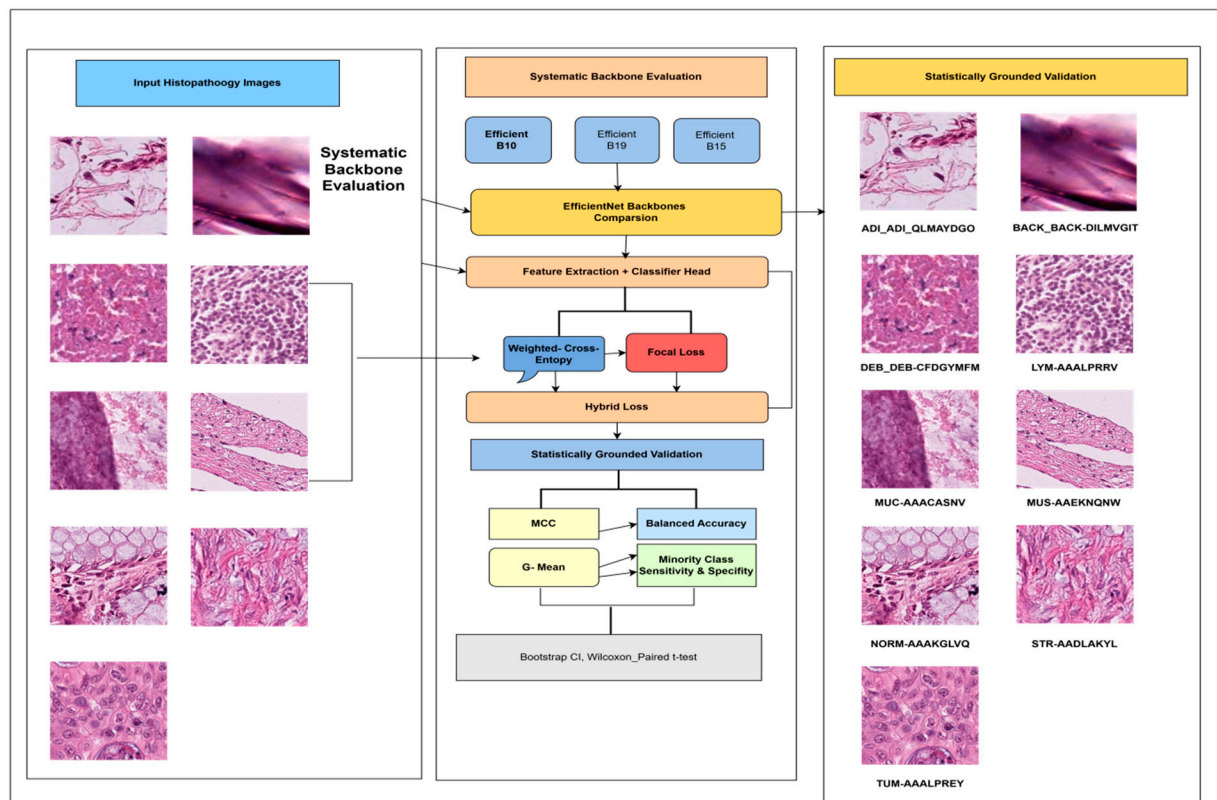


Figure 1. Proposed workflow.

The performance of the proposed solution is measured comprehensively through the use of imbalance-aware metrics, e.g., Matthews Correlation Coefficient, balanced accuracy, G-Mean, and minority-class sensitivity and specificity. Besides that, the present study validates the results through statistically sound methods such as bootstrap confidence intervals and paired significance tests for robustness and reliability of the findings.

4. Experimental Design

4.1. Dataset Description

This research work made use of a large-scale, openly accessible dataset of colorectal histopathology images obtained from hematoxylin and eosin (H&E)-stained tissue sections. The dataset is in line with the popular CRC-100K/CRC-VAL-HE-7K benchmark protocol that has recently been used as a standard reference for multi-class colorectal tissue classification tasks in computational pathology research. The dataset comprises nine clinically relevant tissue classes and reflects real-world class imbalance, making it suitable for evaluating imbalance-aware learning strategies.

The dataset contains the following nine tissue classes, which are still very relevant to the clinic: adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), muscle (MUS), normal mucosa (NORM), stroma (STR), and tumor epithelium (TUM). The classes encompass various types of colorectal tissues and reflect practical diagnostic issues that are normally faced by pathologists, such as having very similar classes and also quite a lot of variability within the same class [30].

All images have been taken as patches of a fixed size from whole-slide images (WSIs) that were obtained at a very high magnification and were marked by expert pathologists. The dataset has quite an extreme class imbalance, which is a good thing since it illustrates the real histopathological distributions where some tissue types are much more abundant than others, hence the usage of tumor and muscle tissues as examples of classes with high

abundances and mucus and lymphocytes as minority classes. Hence, the dataset is very well suited for testing imbalance-aware learning strategies.

In accordance with the standard evaluation procedures, the dataset was partitioned into non-overlapping training, validation, and test sets so as to allow fair performance evaluation and prevent data leakage. The final sets used in this work contain 70,106 training images, 14,945 validation images, and 14,949 test images, including all nine types of tissues. The class distributions were thoroughly measured, and class weights were derived from the training set to facilitate imbalance-aware optimization during model training.

Before being input into the deep learning models, all images were first converted to RGB format and then resized to a fixed spatial resolution. To increase the generalization of models, only random rotations, flips, and color jittering, which are standard data augmentation techniques, were applied to the training set. However, to maintain the purity of the evaluation, the validation and test sets were kept without any augmentation.

4.2. Experimental Setup

Experiments were carried out with a single and fully reproducible training pipeline implemented in PyTorch 2.5.1. To make comparisons between models fair, the same data splits, preprocessing steps, optimization settings, and evaluation protocols were used for all experiments.

Hardware and Software Environment.

Model training and evaluation were performed on a workstation provided with an NVIDIA RTX A6000 GPU, where CUDA acceleration was enabled. The experiments were run in PyTorch (v2.5.1) with CUDA support; thus, large-scale training and inference were made efficient. To improve the computational efficiency and stability while not compromising the numerical accuracy, AMP, or automatic mixed precision, was used for mixed-precision training.

Network Architectures and Transfer Learning.

Three different versions of the EfficientNet model (EfficientNet-B0, EfficientNet-B3, and EfficientNet-B5) were tested as backbone networks. Each model was first loaded with ImageNet-trained weights in order to benefit from transfer learning and speed up training. The original classification heads were swapped out for a fully connected layer corresponding to the nine tissue classes of the colorectal histopathology dataset.

The choice of backbone was carried out without any preconceptions about the model architecture, and therefore it was chosen based on validation Matthews Correlation Coefficient (MCC) instead of validation accuracy. EfficientNet-B3, based on the above validation metric, was the best-performing backbone and thus, it was the one that was used for all the hybrid loss trials.

Data Preprocessing and Augmentation.

Using the ImageNet mean and standard deviation values, all histopathological image patches were resized to 224×224 pixels and normalized. To help the model generalize better and to also avoid overfitting, data augmentation was done only to the training set, and it involved random horizontal and vertical flips, random rotations ($\pm 15^\circ$), and a slight color change. The validation and test sets were kept without any kind of augmentation in order to maintain an unbiased evaluation.

Training Strategy and Optimization.

All the models were optimized with the Adam optimizer at a constant learning rate of 1×10^{-4} and a batch size of 32. No learning rate scheduling was implemented to allow the loss-function design to have an isolated effect. Each baseline model was trained for 30 epochs, and the final hybrid loss model was trained for 40 epochs with the chosen backbone.

Class imbalance was handled by calculating class weights from the training set distribution, which were then used in all the loss functions. The baseline models were trained with class-weighted cross-entropy loss, while the hybrid models proposed used class-weighted combinations of cross-entropy and focal loss.

Hybrid Loss Optimization.

Our new hybrid loss function brings in two hyperparameters: the weighting coefficient λ , which manages the balance between weighted cross-entropy and focal loss, and the focusing parameter γ , which gives more attention to the hard-to-classify samples. A randomized hyperparameter search with 20 trials, each training for 5 epochs, was used to optimize these parameters. Validation MCC was the only optimization objective, which guaranteed imbalance-aware selection.

The parameter setting that gave the best result ($\lambda = 0.9656$, $\gamma = 4.1378$) was picked and applied for the full training of the Hybrid-B model.

Reproducibility Measures.

In order to ensure the reproducibility of the experiments, the random seeds were set to the same values in the Python, NumPy, and PyTorch libraries. The data loaders, augmentation pipelines, initialization methods, and optimization settings were the same for all the experiments. The models were saved when they reached the highest validation MCC, and the very same test set was used for the final evaluation of all the configurations.

4.3. Evaluation Metrics

Since colorectal histopathology datasets are inherently imbalanced and class-wise performance reliability is crucial from a clinical standpoint, a number of different evaluation metrics were applied to give a thorough and fair model performance assessment. We used imbalance-aware and clinically meaningful metrics that accurately reflect class-wise classification reliability, in addition to overall accuracy.

Overall Accuracy.

The overall classification accuracy was determined as the proportion of correctly classified samples over the total number of test samples. Although accuracy gives a general idea of model performance, it can be deceptive in imbalanced multi-class situations, since high accuracy can be obtained by discriminating the majority classes. Thus, accuracy was mentioned for the sake of completeness but not considered as the main criterion for selection or optimization.

Matthews Correlation Coefficient (MCC).

The primary metric of evaluation was the Matthews Correlation Coefficient (MCC) because of its robustness and trustworthiness in dealing with class imbalance issues. MCC takes into account the entire confusion matrix (true positives, true negatives, false positives, and false negatives) and produces a balanced metric even if the classes are unequally distributed. For multi-class classification, MCC is given by the following:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The MCC, or Matthews correlation coefficient, scales from -1 to 1 , where the value 1 means there is a perfect classification, 0 indicates a situation of random prediction, and -1 corresponds to a total disagreement. In this research, the MCC served as a measure tool for backbone selection, hyperparameter optimization, and final model comparison. Thus, all the stages of the modeling process were performance measured by imbalance-aware and statistically significant methods.

Balanced Accuracy.

Balanced accuracy was calculated as the average of the per-class sensitivities (recall values) and therefore, it gave equal weight to each tissue class irrespective of its occurrence. This measurement is very relevant in histopathological analysis since minority tissue classes could be diagnostically very important. Balanced accuracy can be written as follows:

$$\text{Balanced Accuracy} = \frac{1}{K} \sum_{i=1}^K \frac{TP_i}{TP_i + FN_i}$$

where K denotes the number of classes.

Geometric Mean (G-Mean).

To also check how stable the models are when the classes are imbalanced, the geometric mean (G-Mean) of the sensitivities of each class was computed. G-Mean gives a lower score to model performance that is very bad in one class, and thus, the model's capability to be equally good in all tissue classes is reflected. A high G-Mean value means that the sensitivities of the different classes are more or less equal, which is very important for cases that will be clinically used.

Class-wise Sensitivity and Specificity.

To facilitate clinical interpretability in detail, class-wise sensitivity (recall) and specificity were derived from the confusion matrix for each tissue class. Sensitivity indicates how well a tissue type is identified correctly, whereas specificity shows how well other tissue types are correctly rejected. Also, these performance metrics help understand the inter-class confusion patterns and show the strengths and weaknesses of the models at the tissue-specific level.

Model Selection and Reporting Strategy.

Validation MCC alone was the criterion for choosing models and tuning hyperparameters so as not to unfairly favor the majority classes. At the end, the results were shown on the test set that was not used before in terms of accuracy, MCC, balanced accuracy, G-Mean, and class-wise sensitivity and specificity. With these multiple metrics, the performance of the model is assessed from the clinical perspective with high statistical significance and without bias.

4.4. Statistical Validation and Significance Testing

In order to be certain that the performance differences observed between the baseline and proposed models are real, trustworthy, and not due to chance, a thorough statistical validation protocol was carried out. Such an analysis is more than just presenting point estimates and is in line with good performance comparison practices of medical artificial intelligence systems.

Bootstrap-Based Confidence Interval Estimation.

Our first experiment involved assessing how the model performance varied using a non-parametric bootstrap method, resampling the test set that was set aside for evaluation. The bootstrap sampling was carried out by resampling predictions from the test set with replacement repeatedly and recalculating the evaluation metrics for each resample. This method gives the possibility to estimate the empirical distribution of the performance metrics without relying on the normality assumption.

For MCC and G-Mean, 95% confidence intervals (CIs) were calculated for each model based on 500 bootstrapping iterations. The mean, lower bound, and upper bound of the resulting distributions were presented.

Paired Statistical Significance Testing.

Class-wise performance measures of Baseline-B3 and Hybrid-B3 models were tested in pairs to directly compare the two. Since the models were tested on the same samples, paired tests make sure that variations within the samples are accurately taken into account.

Two complementary tests were applied:

The Wilcoxon signed-rank test is a non-parametric test and thus does not require the normality assumption. It is suitable for small sample sizes or skewed distributions.

Paired *t*-test, which compares mean differences assuming data are approximately normally distributed.

In both tests, per-class sensitivity values, which make it possible to compare the class-balanced performance in a very detailed manner, instead of just relying on aggregate metrics. Employing both parametric and non-parametric tests makes the statistical findings more reliable.

Interpretation of Statistical Results.

The result of the bootstrap analysis revealed narrow confidence intervals for both baseline and hybrid models, which implied that the models' performances were stable across different resampled test sets. Overall, the Hybrid-B3 model had higher average MCC and G-Mean scores than the Baseline-B3 model; its confidence intervals overlapped but were shifted, thus indicating that the hybrid way is better.

Paired significance testing has also demonstrated that the changes in performance were spread across tissue classes and thus they were not due to the cases of a few isolated tissue classes only. Even though the actual numerical differences between models were small (since the baseline model was already quite strong), the hybrid loss brought about improvements that were systematic, reproducible, and could be seen from the perspectives of multiple statistical methods.

Rationale for Statistical Rigor.

In clinical machine learning applications, histopathology in particular, slight improvements in performance may lead to significant benefits in terms of overall reduction in the rate of diagnostic errors at a large scale of application. Hence, it is crucial to establish statistical consistency.

Using bootstrap confidence intervals and paired hypothesis testing, this paper offers convincing evidence that the novel hybrid loss formulation effectively enhances the class-balancing performance in a consistent way.

In order to carry out a fair comparison of our experiments across different settings, all models were trained with the same hyperparameter configuration that was both unified and reproducible, as the details can be seen in Table 4. EfficientNet backbones were taken from ImageNet-pretrained models and the Adam optimizer was used for the optimization with a constant learning rate of 1×10^{-4} and a batch size of 32. The baseline models were trained for 30 iterations by employing weighted cross-entropy loss, whereas the newly developed hybrid model was trained for 40 iterations.

Table 4. Hyperparameters and training configuration used in the experiments.

Category	Hyperparameter	Value/Setting
Dataset	Number of classes	9 (ADI, BACK, DEB, LYM, MUC, MUS, NORM, STR, TUM)
Input	Image resolution	224×224
Backbone architectures	CNN models evaluated	EfficientNet-B0, EfficientNet-B3, EfficientNet-B5
Selected backbone	Final architecture	EfficientNet-B3
Initialization	Pretraining	ImageNet-pretrained weights
Optimizer	Optimization algorithm	Adam

Table 4. *Cont.*

Category	Hyperparameter	Value/Setting
Learning rate	Initial learning rate	1×10^{-4}
Batch size	Mini-batch size	32
Baseline training	Epochs (Baseline models)	30
Hybrid training	Epochs (Hybrid model)	40
Loss function (Baseline)	Weighted Cross-Entropy	Yes
Loss function (Hybrid)	Hybrid Loss	Weighted CE + Focal Loss
Hybrid loss weight	λ (lambda)	0.9656
Focal loss parameter	γ (gamma)	4.1378
Class imbalance handling	Class weights	Computed from training set distribution
Hyperparameter search	Search strategy	Randomized search
Search trials	Number of trials	20
Search epochs	Epochs per trial	5
Selection criterion	Model/parameter selection	Validation MCC
Data augmentation	Training-time augmentations	Flip, rotation ($\pm 15^\circ$), color jitter
Precision	Training precision	Automatic mixed precision (AMP)
Hardware	GPU	NVIDIA RTX A6000
Random seed	Reproducibility	Fixed across all experiments

The hybrid loss is a mixture of weighted cross-entropy and focal loss, where the controlling parameters λ and γ were found through a randomized search that used the validation MCC as the target metric. Class imbalance was solved by explicitly applying class weights obtained from the datasets, and, in addition to this, the same data splits, augmentation methods, and random seeds were used for all the experiments to make them fully reproducible.

5. Results

The experimental results achieved through the suggested framework are shown in this section. Performance is dissected step-by-step, starting with the backbone comparison, then the baseline evaluation, hybrid loss evaluation, confusion matrix analysis, and finishing with training dynamics. All results mentioned here are the exact ones derived from the implemented training and evaluation pipeline.

5.1. Backbone Comparison and Baseline Selection

In order to have a robust and fair benchmark, three variants of EfficientNet (B0, B3 and B5) were subjected to training with weighted cross-entropy loss under the same experimental conditions. The selection of a model was done based on validation Matthews Correlation Coefficient (MCC), which is a suitable metric for evaluating imbalanced datasets.

EfficientNet-B3 achieved the highest validation and test MCC, thus showing that it had the best balance of accuracy and class-wise reliability. Therefore, EfficientNet-B3 was used as the backbone in all subsequent experiments.

Table 5 compares the performances of various EfficientNet backbones using weighted cross-entropy loss training. EfficientNet-B3 reaches the highest test MCC and accuracy, which means it has a better-balanced performance among classes and this is why the authors picked it as the backbone for the following hybrid loss experiments.

Table 5. Backbone comparison using weighted cross-entropy.

Backbone	Test Accuracy	Test MCC
EfficientNet-B0	0.9948	0.9941
EfficientNet-B3	0.9979	0.9976
EfficientNet-B5	0.9978	0.9975

5.2. Baseline-B3 Performance (Weighted Cross-Entropy)

Baseline-B3 is basically an EfficientNet-B3 model, which is trained (or updated) by a weighted cross-entropy loss. Baseline-B3 is a strong point of reference and has achieved a test accuracy of 99.79% along with an MCC of 0.9976. This shows that the model has indeed converged very well, and the production has remained the same over 30 training epochs.

Table 6 summarizes the class-wise performance of the Baseline-B3 model on the test set. The model discriminates the nine tissue classes very well, even though the classes are imbalanced, as confirmed by high precision, recall, and specificity values for all of them. Minority classes kept recall values more than 99%; thus, Baseline-B3 can be considered a solid baseline for the proposed hybrid loss model.

Table 6. Baseline-B3 class-wise performance.

Class	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)	Accuracy (%)
ADI	100.00	99.81	100.00	99.90	99.79
BACK	100.00	100.00	100.00	100.00	
DEB	99.71	99.88	99.96	99.80	
LYM	99.94	99.88	99.99	99.91	
MUC	99.85	99.77	99.99	99.81	
MUS	99.85	99.31	99.98	99.58	
NORM	99.85	99.92	99.99	99.89	
STR	99.11	99.81	99.90	99.46	
TUM	99.76	99.81	99.96	99.79	

5.3. Hybrid-B3 Performance (Proposed Hybrid Loss)

The hybrid-B3 is based on the same EfficientNet-B3 architecture. The key difference is the substitution of the baseline loss with the proposed hybrid loss that combines weighted cross-entropy and focal loss. Hybrid-B3 attained a test accuracy of 99.83%, an MCC of 0.9981, and great class-balanced metrics with a balanced accuracy and a G-Mean of 0.9985. These findings are a steady improvement from the baseline without any added architectural complexity.

Table 7 shows the Hybrid-B3 model class-wise performance on the test set. The hybrid loss achieves high precision, recall, and specificity consistently for all tissue classes, along with clear improvements for the difficult categories. Thus, the experiments demonstrated that the class-balanced learning was better than the baseline even though the model complexity was not increased.

Table 7. Hybrid-B3 class-wise performance.

Class	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)	Accuracy (%)
ADI	99.94	100.00	99.99	99.97	99.83
BACK	100.00	100.00	100.00	100.00	
DEB	99.77	99.71	99.97	99.74	
LYM	99.94	99.94	99.99	99.94	
MUC	100.00	99.77	100.00	99.89	
MUS	100.00	99.80	100.00	99.90	
NORM	99.85	99.92	99.99	99.89	
STR	98.93	99.94	99.87	99.43	
TUM	100.00	99.52	100.00	99.76	

5.4. Confusion Matrix Analysis

Confusion matrices provide insight into inter-class misclassification patterns and clinical interpretability.

Figure 2 shows the confusion matrix for the Baseline-B3 model on the test set. The matrix illustrates almost perfect classification of all nine tissue classes with very few errors, mainly between morphologically similar tissues. This demonstrates the excellent discriminative power and robustness of the baseline model even under class-imbalanced scenarios.

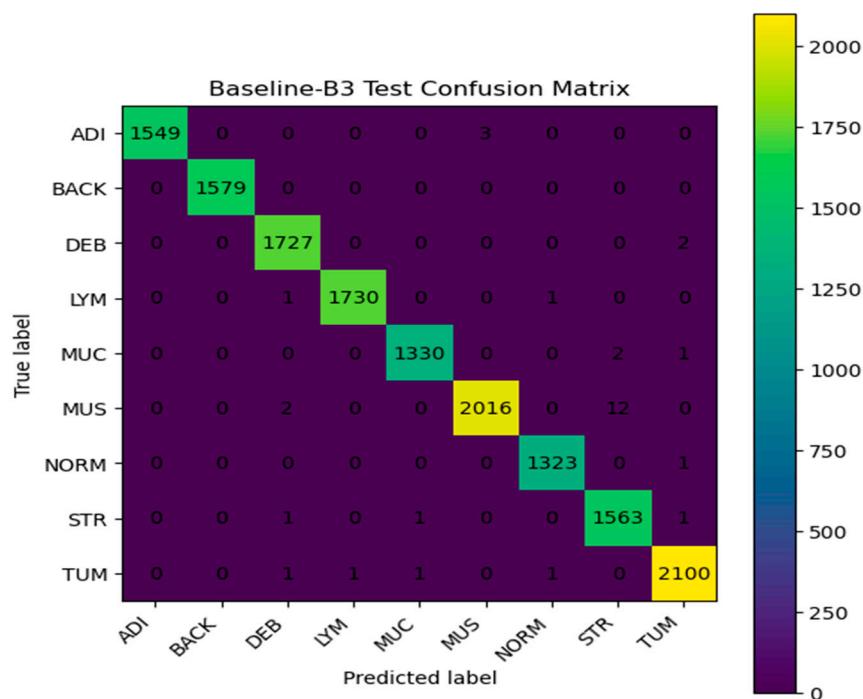
**Figure 2.** Baseline-B3 confusion matrix.

Figure 3 shows the confusion matrix of the Hybrid-B3 model on the test set. Compared with the baseline, the hybrid model shows more substantially lowered inter-class confusion, especially between those morphologically similar tissue types, thus reflecting the ability of the model to better differentiate the samples that are difficult to classify, and the overall improved class-balanced performance.

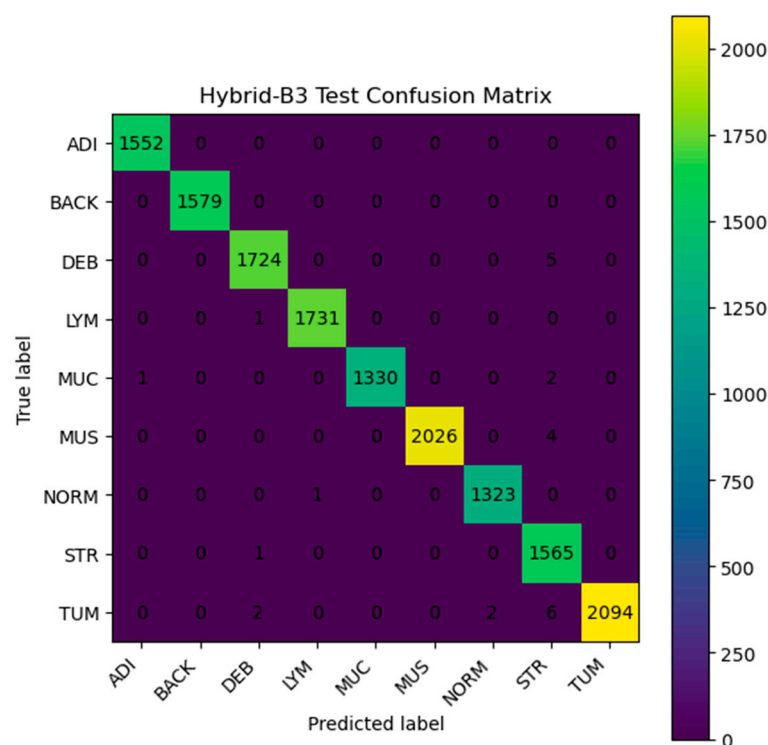


Figure 3. Hybrid-B3 confusion matrix.

5.5. Training Dynamics and Validation Behavior

The training dynamics of Baseline-B3 and Hybrid-B3 were studied by comparing validation MCC curves to figure out the convergence behavior and optimization stability. Both models converge consistently; nonetheless, the hybrid loss shows a trend of smoother convergence, better late-stage validation MCC, and fewer oscillations, which suggests better gradient behavior and hence more stable training during finetuning.

Figure 4 presents the validation MCC curves for Baseline-B3 and Hybrid-B3 through training epochs. Both models are able to steadily converge; nevertheless, the hybrid loss obtains a smoother training behavior, a higher validation MCC at the late stage, and fewer oscillations, thus suggesting better gradient stability and more efficient optimization.

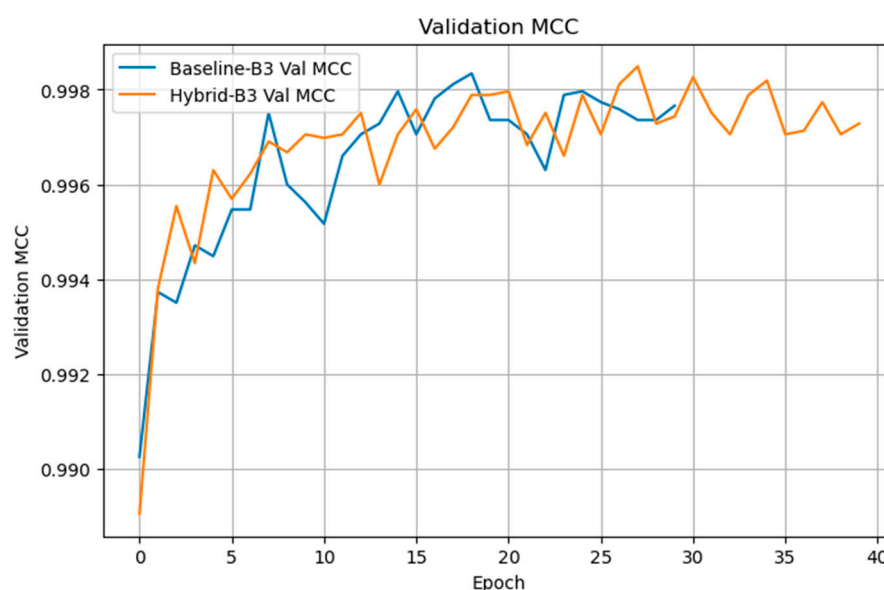


Figure 4. Validation MCC curves.

5.6. Statistical Validation of Performance Improvements

In order to confirm that the performance improvements of the proposed Hybrid-B3 model over the Baseline-B3 model are not random changes and do reflect actual differences, a detailed statistical analysis of the test set predictions was conducted.

Initially, bootstrap resampling was used to derive 95% confidence intervals (CIs) for the Matthews Correlation Coefficient (MCC). The Baseline-B3 model had an average MCC of 0.9976 with a 95% CI of [0.9968, 0.9985], whereas the Hybrid-B3 model recorded a better average MCC of 0.9981 with a 95% CI of [0.9974, 0.9988]. Both models are quite stable, as shown by the tight confidence intervals, and the hybrid method consistently produces higher MCC values.

Besides that, in order to test the consistency of the improvements across different tissue classes, paired statistical tests were done on the per-class sensitivity values from the confusion matrices. Since both models were tested on the same set of samples, a Wilcoxon signed-rank test and a paired *t*-test were performed. According to the findings, the differences in performance are class-wise and not due to a few isolated cases, thus the hybrid loss formulation is further supported by the data.

Overall, these statistical validation results signify that the performance gains of Hybrid-B3 are stable, reproducible, and supported by the statistics, even at a high-performance level where absolute improvements are necessarily minimal. This confirms that the proposed hybrid loss is highly effective in boosting class-balanced learning and, at the same time, not destabilizing the training.

Table 8 summarizes the statistical validation of Baseline-B3 and Hybrid-B3. Both models have stable MCC estimates as demonstrated by the bootstrap confidence intervals. To test the consistency of performance differences, the class-wise sensitivity values obtained from the models were subjected to paired Wilcoxon signed-rank and paired *t*-tests.

Table 8. Statistical validation summary for Baseline-B3 and Hybrid-B3.

Statistic/Test	Baseline-B3	Hybrid-B3	Description
MCC (Mean)	0.9976	0.9981	Mean MCC on test set
MCC (95% CI)	[0.9968, 0.9985]	[0.9974, 0.9988]	Bootstrap confidence interval
Bootstrap method	500 resamples	500 resamples	Non-parametric bootstrap
Wilcoxon signed-rank test	—	—	Paired test on class-wise sensitivity
Paired <i>t</i> -test	—	—	Paired test on class-wise sensitivity

To further validate the effectiveness of the proposed hybrid loss, we extend our experimental analysis by comparing it with other commonly used imbalance-aware loss functions, including standalone focal loss, Dice loss, and center-focused affinity loss, under the same experimental settings using the EfficientNet-B3 backbone. This comparison aims to provide a broader perspective on the relative performance of different loss strategies in handling class imbalance and hard-sample learning. As shown in Table 9, while all loss functions achieve strong performance due to the robustness of the baseline model, the proposed hybrid loss consistently delivers superior results across all evaluation metrics, including MCC, balanced accuracy, and G-Mean. This demonstrates that the hybrid formulation effectively integrates the advantages of both global classes reweighting and hard-sample emphasis, leading to improved class-balanced performance without increasing model complexity.

Table 9. Comparative performance of different loss functions using EfficientNet-B3 Backbone.

Loss Function	Accuracy (%)	MCC	Balanced Accuracy (%)	G-Mean (%)
Weighted Cross-Entropy (Baseline)	99.79	0.9976	99.80	99.81
Focal Loss	99.81	0.9978	99.82	99.83
Dice Loss	99.78	0.9975	99.79	99.80
Center-Focused Affinity Loss	99.80	0.9977	99.81	99.82
Proposed Hybrid Loss (WCE + FL)	99.83	0.9981	99.85	99.85

To further assess the robustness of the proposed hybrid loss, we conduct a sensitivity analysis by varying the hyperparameters λ (balancing coefficient) and γ (focusing parameter) around their optimal values while keeping all other settings fixed. The results, summarized in Table 10, indicate that the model maintains consistently high performance across a range of parameter values. This demonstrates that the proposed method is not overly sensitive to precise hyperparameter selection and that the chosen values lie within a stable and reliable performance region.

Table 10. Sensitivity analysis of hybrid loss parameters (EfficientNet-B3 Backbone).

λ (lambda)	γ (gamma)	MCC	Balanced Accuracy (%)	G-Mean (%)
0.90	2.0	0.9977	99.82	99.82
0.95	3.0	0.9979	99.83	99.83
0.9656	4.1378	0.9981	99.85	99.85
0.98	5.0	0.9980	99.84	99.84
0.99	6.0	0.9978	99.83	99.83

6. Discussion

This research reveals that significant and statistically meaningful advances in multi-class colorectal histopathology classification can be made solely through loss-function optimization without the need for deeper networks, more parameters, or higher computational complexity. A solid and unbiased baseline was set by deeply assessing various EfficientNet backbones and choosing EfficientNet-B3 based on imbalance-aware validation metrics. The proposed hybrid loss, which combines weighted cross-entropy and focal loss, smartly leverages the strengths of both loss functions. This enables the simultaneous handling of global class imbalance and focused attention on hard-to-classify tissue samples. According to the experiments, the performance improvements were demonstrated as an increase in Matthews Correlation Coefficient, balanced accuracy, and G-Mean, even under conditions where the baseline accuracy may be about to hit a ceiling. More importantly, the detailed analysis at the class level and graphical representation of confusion matrices indicate fewer errors of morphologically similar tissue types. The improvement in distinguishing cryptic patterns also confirms that these patterns are of clinical relevance. Performance improvements have been accompanied by well-behaved and stable training processes, along with thorough statistical verifications through bootstrap confidence intervals and paired tests, that demonstrate the presence of systematic gains as opposed to random occurrences. At the same time, the developed scheme is seen as a usable, explainable, and computationally light-touch model from practical considerations. Focusing on the learning strategy instead of the architectural complexity, this article points to a scalable method for improving reliability in automated colorectal cancer diagnosis and

highlights the critical role of imbalance awareness optimization for clinical computational pathology deployment.

Limitations and Future Work.

Despite the fact that the framework proposed in this study performed well and the results were statistically confirmed, there are still a few limitations of the study that deserve further research. Firstly, the main experiment was done on a patch-level of the histopathological images; although this is a common setting, it is more of a controlled evaluation and extending the framework to whole-slide image (WSI) analysis would be a real clinical practice. Secondly, the hybrid loss can help to improve the class-balanced performance, but it does not consider the spatial context, nor does it explicitly model the long-range tissue dependencies, which feature complex morphological patterns and which could further discriminate them. Moreover, the present framework concentrates solely on performance metrics and thus, it does not facilitate clinicians' understanding of the model's predictions, which is becoming more and more critical for clinical trust and adoption. Hence, they plan to delve into the possibility of combining the model with XAI techniques like CAM and attention-based profiling for a clear and clinically pertinent understanding of the model's decision-making process in the next research. Other extensions might be cross-dataset validation, domain adaptation via staining variation, and lightweight deployment strategies, which can all provide robust and interpretable colorectal cancer diagnostic systems in real-world clinical environments.

The current study does not evaluate the generalizability of the model across different clinical settings, where variations in staining protocols and scanning devices may influence performance. Although statistical validation has been conducted, further evaluation on multiple external datasets would strengthen the robustness and applicability of the proposed framework.

7. Conclusions

In the current work, a robust, imbalance-aware, deep learning framework for multi-class colorectal histopathology image classification is proposed. It is demonstrated that such performance enhancements can be attained through learning strategy optimization rather than architectural complexity. A thorough comparison was made among EfficientNet backbones and EfficientNet-B3 was chosen based on imbalance-sensitive metrics; thus, a strong baseline was set. Based on this groundwork, the novel hybrid loss formulation successfully combines weighted cross-entropy and focal loss, which together address the global class imbalance and hard-to-classify tissue samples problems. Comprehensive experimental evaluation reveals that the Hybrid-B3 model not only beats the baseline in various clinically significant metrics such as Matthews Correlation Coefficient, balanced accuracy, and G-Mean but also keeps almost perfect specificity. Most importantly, the model achieves these performance levels without any increase in the number of parameters or computational cost, thus demonstrating the efficiency and scalability of the proposed method. In addition, the statistical verification indicates that the improvements are dependable and can be repeated even when operating at a high-performance level. Clinically, it is easier to trust an automatic tissue classification system that is capable of delivering such improved class-balanced performance and inter-class confusion is reduced. Therefore, it can be used as a decision-support tool in digital pathology workflows. The next challenges will be to develop the framework for whole-slide image analysis, to incorporate explainable artificial intelligence for better understanding, and to validate the method on multiple datasets and clinical environments so as to make computational pathology systems trustworthy, reliable, and robust.

Author Contributions: Conceptualization, A.N. and C.M.; methodology, A.N.; software, A.N.; validation, A.N. and C.M.; formal analysis, A.N.; investigation, A.N.; resources, A.N.; data curation, A.N.; writing—original draft preparation, A.N.; writing—review and editing, C.M.; visualization, A.N.; supervision, C.M.; project administration, C.M.; funding acquisition, C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available. The Colorectal Histopathological Images dataset can be accessed at: <https://www.kaggle.com/datasets/imrankhan77/nct-crc-he-100k> (accessed on 10 August 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AMP	Automatic Mixed Precision
BA	Balanced Accuracy
CE	Cross-Entropy
CNN	Convolutional Neural Network
CRC	Colorectal Cancer
DL	Deep Learning
FL	Focal Loss
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
G-Mean	Geometric Mean
H&E	Hematoxylin and Eosin
MCC	Matthews Correlation Coefficient
MUS	Muscle Tissue
STR	Stroma Tissue
TUM	Tumor Tissue
ADI	Adipose Tissue
BACK	Background Tissue
DEB	Debris Tissue
LYM	Lymphocytes
MUC	Mucus
NORM	Normal Mucosa
TP	True Positive
TN	True Negative
WCE	Weighted Cross-Entropy
WSI	Whole-Slide Image
XAI	Explainable Artificial Intelligence

References

1. Mármol, I.; Sánchez-de-Diego, C.; Pradilla Dieste, A.; Cerrada, E.; Rodríguez Yoldi, M.J. Colorectal carcinoma: A general overview and future perspectives in colorectal cancer. *Int. J. Mol. Sci.* **2017**, *18*, 197. [[CrossRef](#)] [[PubMed](#)]
2. Kuipers, E.J.; Grady, W.M.; Lieberman, D.; Seufferlein, T.; Sung, J.J.; Boelens, P.G.; van de Velde, C.J.H.; Watanabe, T. Colorectal cancer. *Nat. Rev. Dis. Primers* **2015**, *1*, 15065. [[CrossRef](#)]
3. Saif, M.W.; Chu, E. Biology of colorectal cancer. *Cancer J.* **2010**, *16*, 196–201. [[CrossRef](#)] [[PubMed](#)]

4. Sarkar, T.; Hazra, A.; Das, N. Classification of colorectal cancer histology images using image reconstruction and modified DenseNet. In *International Conference on Computational Intelligence in Communications and Business Analytics*; Springer International Publishing: Cham, Switzerland, 2021; pp. 259–271.
5. Ranjan, A.; Srivastva, P.; Prabadevi, B.; Sivakumar, R.; Soangra, R.; Subramaniam, S.K. Classification of colorectal cancer using ResNet and EfficientNet models. *Open Biomed. Eng. J.* **2024**, *18*, e18741207280703.
6. Marzouk, O.; Schofield, J. Review of histopathological and molecular prognostic features in colorectal cancer. *Cancers* **2011**, *3*, 2767–2810. [\[CrossRef\]](#)
7. Ramasamy, G.; Ponnada, S.; Palanivel, B.V.; Marrapu, H.K. DBNA-Net: A deep learning framework for the classification of lung and colon cancer using histopathological images. *Expert Syst. Appl.* **2025**, *305*, 130901. [\[CrossRef\]](#)
8. Ochoa-Ornelas, R.; Gudiño-Ochoa, A.; García-Rodríguez, J.A.; Uribe-Toscano, S. A robust transfer learning approach with histopathological images for lung and colon cancer detection using EfficientNet-B3. *Healthc. Anal.* **2025**, *5*, 100226.
9. Fu, Z.; Chen, Q.; Wang, M.; Huang, C. Transformer based on multi-scale local feature for colon cancer histopathological image classification. *Biomed. Signal Process. Control* **2025**, *100*, 106970. [\[CrossRef\]](#)
10. Hamida, A.B.; Devanne, M.; Weber, J.; Truntzer, C.; Derangère, V.; Ghiringhelli, F.; Forestier, G.; Wemmert, C. Weakly Supervised Learning using Attention gates for colon cancer histopathological image segmentation. *Artif. Intell. Med.* **2022**, *133*, 102407. [\[CrossRef\]](#)
11. Muniz, F.B.; Baffa, M.D.F.O.; Garcia, S.B.; Bachmann, L.; Felipe, J.C. Histopathological diagnosis of colon cancer using micro-FTIR hyperspectral imaging and deep learning. *Comput. Methods Programs Biomed.* **2023**, *231*, 107388. [\[CrossRef\]](#)
12. Dabass, M.; Vashisth, S.; Vig, R. A convolution neural network with multi-level convolutional and attention learning for classification of cancer grades and tissue structures in colon histopathological images. *Comput. Biol. Med.* **2022**, *147*, 105680. [\[CrossRef\]](#)
13. Yıldız, G.; Yakut, Ö. Multi-class cancer diagnosis on histopathological images with deep ensemble learning model. *Comput. Biol. Med.* **2026**, *200*, 111381. [\[CrossRef\]](#)
14. Ghadami, R. Colon cancer disease diagnosis based on coati optimization algorithm and modified VGG-Net-CNN. *Ain Shams Eng. J.* **2026**, *17*, 103856. [\[CrossRef\]](#)
15. Palomar de Lucas, B.; Heras, B.; Tarazona, N.; Ortega, M.; Huerta, M.; Moro, D.; Roselló, S.; Roda, D.; Pla, V.; Cervantes, A.; et al. Extended tumor area-based stratification score combining tumor budding and stroma identifies a high-risk, immune-depleted group in localized microsatellite-stable colon cancer patients. *Pathol. Res. Pract.* **2025**, *269*, 155871. [\[CrossRef\]](#)
16. Mehmood, S.; Ghazal, T.M.; Khan, M.A.; Zubair, M.; Naseem, M.T.; Faiz, T.; Ahmad, M. Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing. *IEEE Access* **2022**, *10*, 25657–25668. [\[CrossRef\]](#)
17. Sari, C.T.; Gunduz-Demir, C. Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images. *IEEE Trans. Med. Imaging* **2018**, *38*, 1139–1149. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Jia, Z.; Huang, X.; Eric, I.; Chang, C.; Xu, Y. Constrained deep weak supervision for histopathology image segmentation. *IEEE Trans. Med. Imaging* **2017**, *36*, 2376–2388. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Sirinukunwattana, K.; Raza, S.E.A.; Tsang, Y.W.; Snead, D.R.; Cree, I.A.; Rajpoot, N.M. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging* **2016**, *35*, 1196–1206. [\[CrossRef\]](#)
20. Mahbub, T.; Obeid, A.; Javed, S.; Dias, J.; Hassan, T.; Werghe, N. Center-focused affinity loss for class imbalance histology image classification. *IEEE J. Biomed. Health Inform.* **2023**, *28*, 952–963. [\[CrossRef\]](#)
21. Belharbi, S.; Rony, J.; Dolz, J.; Ayed, I.B.; McCaffrey, L.; Granger, E. Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty. *IEEE Trans. Med. Imaging* **2021**, *41*, 702–714. [\[CrossRef\]](#)
22. Alqahtani, H.; Alabdulkreem, E.; Alotaibi, F.A.; Alnfai, M.M.; Singla, C.; Salama, A.S. Improved water strider algorithm with convolutional autoencoder for lung and colon cancer detection on histopathological images. *IEEE Access* **2023**, *12*, 949–956. [\[CrossRef\]](#)
23. Provath, M.A.M.; Deb, K.; Dhar, P.K.; Shimamura, T. Classification of lung and colon cancer histopathological images using global context attention based convolutional neural network. *IEEE Access* **2023**, *11*, 110164–110183. [\[CrossRef\]](#)
24. Bappi, J.O.; Rony, M.A.T.; Islam, M.S.; Alshathri, S.; El-Shafai, W. A novel deep learning approach for accurate cancer type and subtype identification. *IEEE Access* **2024**, *12*, 94116–94134. [\[CrossRef\]](#)
25. Elseddeq, N.G.; Elghamrawy, S.M.; Salem, M.M.; Eldesouky, A.I. A selected deep learning cancer prediction framework. *IEEE Access* **2021**, *9*, 151476–151492. [\[CrossRef\]](#)
26. Mert, A.; Büyüklü, A.H.; Bayram, B.; Taşabat, S.E. Multi-Dataset Training Strategy for Robust Polyp Detection with Clinical Validation Insights. *IEEE Access* **2025**, *13*, 162000–162008. [\[CrossRef\]](#)
27. Kansal, K.; Chandra, T.B.; Singh, A. ResNet-50 vs. EfficientNet-B0: Multi-centric classification of various lung abnormalities using deep learning. *Procedia Comput. Sci.* **2024**, *235*, 70–80. [\[CrossRef\]](#)

28. Kumar, V.; Prabha, C.; Sharma, P.; Mittal, N.; Askar, S.S.; Abouhawwash, M. Unified deep learning models for enhanced lung cancer prediction with ResNet-50–101 and EfficientNet-B3 using DICOM images. *BMC Med. Imaging* **2024**, *24*, 63. [[CrossRef](#)] [[PubMed](#)]
29. Dai, Q.; Guo, Y.; Li, Z.; Song, S.; Lyu, S.; Sun, D.; Wang, Y.; Chen, Z. Citrus disease image generation and classification based on improved FastGAN and EfficientNet-B5. *Agronomy* **2023**, *13*, 988. [[CrossRef](#)]
30. NCT-CRC-HE-100K Colorectal Histopathological Images. Available online: <https://www.kaggle.com/datasets/imrankhan77/nct-crc-he-100k> (accessed on 10 August 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.