Original Article

Advanced deep learning for multi-class colorectal cancer histopathology: integrating transfer learning and ensemble methods

Qi Ke[1,2], Wun-She Yap[2], Yee Kai Tee[2]^, Yan Chai Hum[2], Hua Zheng[3], Yu-Jian Gan[4]

[1]School of Big Data and Artificial Intelligence, Guangxi University of Finance and Economics, Nanning, China; [2]Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Kajang, Selangor, Malaysia; [3]Guangxi Key Laboratory of Seaward Economic Intelligent System Analysis and Decision-making, Guangxi University of Finance and Economics, Nanning, China; [4]Department of Information Studies, University College London, London, UK

Contributions: (I) Conception and design: Q Ke, YK Tee, WS Yap; (II) Administrative support: YC Hum, WS Yap, YJ Gan; (III) Provision of study materials or patients: Q Ke, WS Yap, H Zheng; (IV) Collection and assembly of data: Q Ke, YC Hum, YJ Gan; (V) Data analysis and interpretation: Q Ke, YK Tee, H Zheng; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Yee Kai Tee, DPhil. Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Jalan Sungai Long, Bandar Sungai Long, Cheras 43000, Kajang, Selangor, Malaysia. Email: teeyeekai@gmail.com.

Background: Cancer is a major global health threat, constantly endangering people's well-being and lives. The application of deep learning in the diagnosis of colorectal cancer can improve early detection rates, thereby significantly reducing the incidence and mortality of colorectal cancer patients. Our study aims to optimize the performance of deep learning model in the classification of colorectal cancer histopathological images to assist pathologists in improving diagnostic accuracy.

Methods: In this study, we developed ensemble models based on deep convolutional neural networks (CNNs) for the classification of colorectal cancer histopathology images. The method first involved data preprocessing techniques such as patch cropping, stain normalization, data augmentation and data balancing on histopathology images with different magnifications. Subsequently, the CNN models were fine-tuned and pre-trained using transfer learning methods, and models with superior performance were then selected as the base classifiers to build the ensemble models. Finally, the ensemble models were used to predict the final classification outcomes. To evaluate the effectiveness of the proposed models, we tested their performance on a publicly available colorectal cancer dataset, Enteroscope Biopsy Histopathological Hematoxylin and Eosin Image (EBHI) dataset.

Results: Experimental results show that the proposed ensemble model, composed of the top five classifiers, achieved the promising classification accuracy across sub-databases with four different magnification factors. Specifically, on the 40× magnification subset, the highest classification accuracy reached 99.11%; on the 100× magnification subset, it reached 99.36%; on the 200× magnification subset, it was 99.29%; and on the 400× magnification subset, it was 98.96%. Additionally, the proposed ensemble model achieved exceptional results in recall, precision, and F1 score.

Conclusions: The proposed ensemble models obtained good classification performance on the EBHI dataset of histopathological images for colorectal cancer. The findings of this study may contribute to the early detection and accurate classification of colorectal cancer, thereby aiding in more precise diagnostic analysis of colorectal cancer.

^ ORCID: 0000-0002-0263-6358.

Keywords: Histopathological image; colorectal cancer; deep learning (DL); transfer learning (TL); ensemble learning

Submitted Aug 14, 2024. Accepted for publication Jan 21, 2025. Published online Feb 26, 2025.
doi: 10.21037/qims-24-1641
View this article at: https://dx.doi.org/10.21037/qims-24-1641

Introduction

Cancer is a fatal disease caused by a variety of biochemical abnormalities and genetic disorders, where abnormal cells begin to develop uncontrollably in any tissue or organ of the body (1). According to the latest assessment by cancer statistics for the year 2020 (2), there were an estimated 19.3 million new cancer cases and nearly 10.0 million cancer deaths globally. Colorectal cancer is the third most common cause of cancer deaths worldwide, with more than 1.85 million cases and approximately 850,000 deaths per year. Approximately 3.2 million colorectal cancer patients are projected to be living in 2040, with China and the United States expected to have the highest incidence rates over the next 20 years (3). Besides obesity, smoking, alcohol consumption and unhealthy lifestyle, the incidence of colorectal cancer is also related to gender, heredity and family factors (4). Benign polyps begin in the lining of the colon or rectum and can develop into malignant tumors, which are the primary cause of cancer deaths.

Histopathological images serve as the gold standard for cancer diagnosis, offering higher resolution, sharpness, stable imaging, and facilitating easy observation and analysis (5). Histopathologic examination of the intestinal tract is a prerequisite for the diagnosis and treatment of colorectal cancer. Efficient and accurate cancer diagnosis is crucial for timely detection and subsequent treatment of patients. Due to the subtle differences between histopathologic images, the relatively concentrated color distribution, and the significant variations in cell characteristics under different microscopic magnifications, pathologists face a non-trivial task in performing this manual inspection, leading to visual fatigue and reduced diagnostic efficiency (6). Moreover, the manual inspection relies on the subjective experience and knowledge of the pathologists, introducing the risk of intra- and inter-observation variations during the analysis of histopathologic images (7). Addressing these challenges is crucial to improving the accuracy and efficiency of cancer diagnosis.

In recent years, the significance of deep learning (DL)-based models for analyzing histopathological images has been demonstrated in colorectal cancer research (8,9). These models can help pathologists overcome bias and error, and automate the diagnosis of colorectal cancer, facilitating timely treatment for patients (10). Many researchers have shown that DL methods have strong capabilities and effectiveness in the classification, identification, and detection of various cancers (11-13). Convolutional neural networks (CNNs) are powerful tools for feature extraction in cancer classification, utilizing their hierarchical structure, adaptive filtering, and ability to capture complex patterns to significantly improve classification accuracy (14). However, these models are often trained from scratch, requiring significant training time and a large number of training samples, which is a major challenge to be overcome in current medical DL research (15,16).

Transfer learning (TL) methods have played a vital role in the field of DL-based medical diagnostics. If the dataset at hand does not have an adequate number of images, then the CNN model will not achieve a good training performance. TL helps to adapt large pre-trained networks to downstream tasks, reducing the need for extensive training data, and is well-suited to address the problem of scarce medical image data (17). In the medical field, TL is often used due to the lack of large, widely annotated medical datasets comparable to the ImageNet database (18). In studies (19,20), researchers applied TL with pre-trained CNN models to classify colorectal cancer, achieving a remarkable classification accuracy 0.98. TL is an effective approach to improve the model's performance and adapt it to new and related tasks (21).

Each network model has its own unique strengths and weaknesses in feature extraction due to differences in their architectures (22). However, most medical image classification studies rely on a single architecture for feature extraction, which often fails to capture a wide range of effective feature. Some studies have also recognized the limitations of using a single model and have proposed hybrid classification models (23). Therefore, combining deep features from multiple networks with different structures is becoming more popular, as it usually improves classification performance (24).

Ensemble learning involves aggregating the output decisions of multiple base classifiers and combining pre-trained networks through an ensemble strategy to make final classification predictions (25). The core idea of ensemble learning is that the integration of multiple models can be more powerful than a single model. It can effectively reduce the bias and variance of individual models, thereby enhancing the accuracy and stability of predictions (22). Ensemble methods in medical imaging have wide-ranging applications across multiple domains, particularly in disease detection, segmentation, classification, and prediction, where they have achieved significant results (26). References (27,28) utilized ensemble learning for classification of cervical tissue histopathological images and malaria cell images, respectively. In reference (29), an ensemble learning model using a reproducible classification pipeline technique is employed to classify images from four datasets of varying complexity: colorectal histological images, coronavirus disease, skin lesion images, and diabetic retinopathy images.

There have been some studies on the application of ensemble learning in colorectal cancer pathology images. In reference (30), Khazaee Fadafen *et al.* constructed an ensemble model using multiple Support Vector Machine (SVM) classifiers, achieving high classification accuracies of 98.75% and 99.76% on the CRC-5000 and NCT-CRC-HE-100K colorectal cancer datasets, respectively. Yengec-Tasdemir *et al.* proposed a method for classifying colon tissue pathology images using a combination of ensemble learning and stain normalization techniques. Fine-tuned ConvNeXt-Tiny and ConvNeXt-Base were used as classifiers, achieving the highest classification accuracy of 95.00% across three different datasets (31). Ghosh *et al.* constructed an ensemble model using DenseNet121, Xception, Inception-ResNet V2, and a custom CNN architecture for multi-class colorectal histopathology image classification. By assigning nonlinear weights to individual architectures and updating the weights using the Adam optimizer, they achieved a maximum accuracy of 99.13% on the combined dataset (32). These ensemble models leverage the advantages of different networks and combine handcrafted features or those automatically extracted from DL architectures, resulting in more accurate and reliable classification outcomes for colorectal cancer histopathological images (33).

Numerous DL models have been explored for colorectal cancer detection. However, several critical issues remain unresolved. Many existing studies employ single-model TL techniques to address specific target tasks, without adequately considering and comparing the performance of different models in TL. The effects of various base classifiers and ensemble strategies on the performance of ensemble models have not been thoroughly explored or validated. Importantly, there is a noticeable gap in research on classification models for complex, multi-class colonoscopy biopsy histopathology image datasets. These datasets pose challenges such as varying magnification levels, high resolution, class imbalance, and the necessity to crop images into smaller patches for accurate classification.

To address these gaps, we conducted a systematic study on image classification for imbalanced multi-class colorectal cancer datasets with different magnification levels. This work proposed hybrid DL models that integrate the advantages of TL and ensemble learning for more accurate classification of colorectal cancer histopathological images. We used TL methods to tackle the challenge of limited public data availability. By integrating multiple CNNs, the models overcome the limitations of a single network in feature extraction. To validate the effectiveness of our proposed models, we evaluated performance on a new publicly available dataset, the Enteroscope Biopsy Histopathological Hematoxylin and Eosin Image (EBHI) dataset (34). We present this article in accordance with the TRIPOD+AI reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1641/rc).

Methods

Data pre-processing

Dataset

The dataset utilized in this study is a new publicly available dataset denoted as EBHI. The dataset comprises a total of 5,532 electron microscopy images, each with a resolution of 2,048×1,536 pixels, derived from colonoscopic biopsy samples. These images were annotated at the image level by two pathologists from the Cancer Hospital of China Medical University. It was jointly released by Northeastern University, China Medical University, Liaoning Provincial Cancer Hospital, and Liaoning Provincial Cancer Research Institute in China in 2023 (34). The dataset contains five tumor differentiation stages: adenocarcinoma, high-grade intraepithelial neoplasia (high-grade IN), low-grade intraepithelial neoplasia (low-grade IN), polyp and normal, and was divided into four magnifications, 40×, 100×, 200×,

Table 1 Details of the EBHI dataset

Classes	Dataset types	Magnification				
		40×	100×	200×	400×	Total
Adenocarcinoma	Original	205	471	790	812	2,278
	Patches cropping	1,779	4,927	8,599	8,975	24,280
	Data augmentation	1,781	4,927	8,599	8,975	24,282
High-grade IN	Original	47	80	130	161	418
	Patches cropping	411	855	1,447	1,819	4,532
	Data augmentation	1,781	4,927	8,599	8,975	24,282
Low-grade IN	Original	204	341	603	660	1,808
	Patches cropping	1,781	3,498	6,586	7,099	18,964
	Data augmentation	1,781	4,927	8,599	8,975	24,282
Polyp	Original	119	165	254	304	842
	Patches cropping	972	1,706	2,839	3,374	8,891
	Data augmentation	1,781	4,927	8,599	8,975	24,282
Normal	Original	17	29	61	79	186
	Patches cropping	145	311	685	878	2,019
	Data augmentation	1,781	4,927	8,599	8,975	24,282
Total	Original	592	1,086	1,838	2,016	5,532
	Patches cropping	5,088	11,297	20,156	22,145	58,686
	Data augmentation	8,905	24,635	42,995	44,875	121,410

EBHI, Enteroscope Biopsy Histopathological Hematoxylin and Eosin Image; IN, intraepithelial neoplasia.

and 400×. The data distribution in the original dataset is unbalanced across categories. The details of the dataset are shown in *Table 1*. *Figure 1* displays the five categories of histopathological images in the EBHI dataset at various magnifications, with each row representing the same magnification and each column representing the same category.

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Patches cropping

The images in the EBHI dataset are 2,048×1,536 pixels, which are too large for direct input into the network. In this study, we performed patch classification by extracting patches from the original images. A Patch-Cropping Algorithm was designed to crop the original histopathological images. The main ideas of the algorithm are: firstly, converting the original images into patches in a format suitable for patch classification. Secondly, removing empty patches to generate a more optimized and well-distributed dataset. The details of the Patch-Cropping Algorithm are as below:

(I) The original images are converted to grayscales by diminishing noise to enhance distinction between foreground and background.

(II) Edge detection is performed, and a grid of mask map is created according to the requested resolution.

(III) The mask map is cropped to generate the patch dataset. A threshold is set to evaluate whether each patch meets the minimum content requirement to be considered as foreground.

(IV) Empty patches that do not contain useful categorical information are automatically removed.

We cropped the original image into initial patches with dimensions of 380×380 pixels to meet the largest image input size of our tested models before resizing them to smaller patch size requirements by other networks to

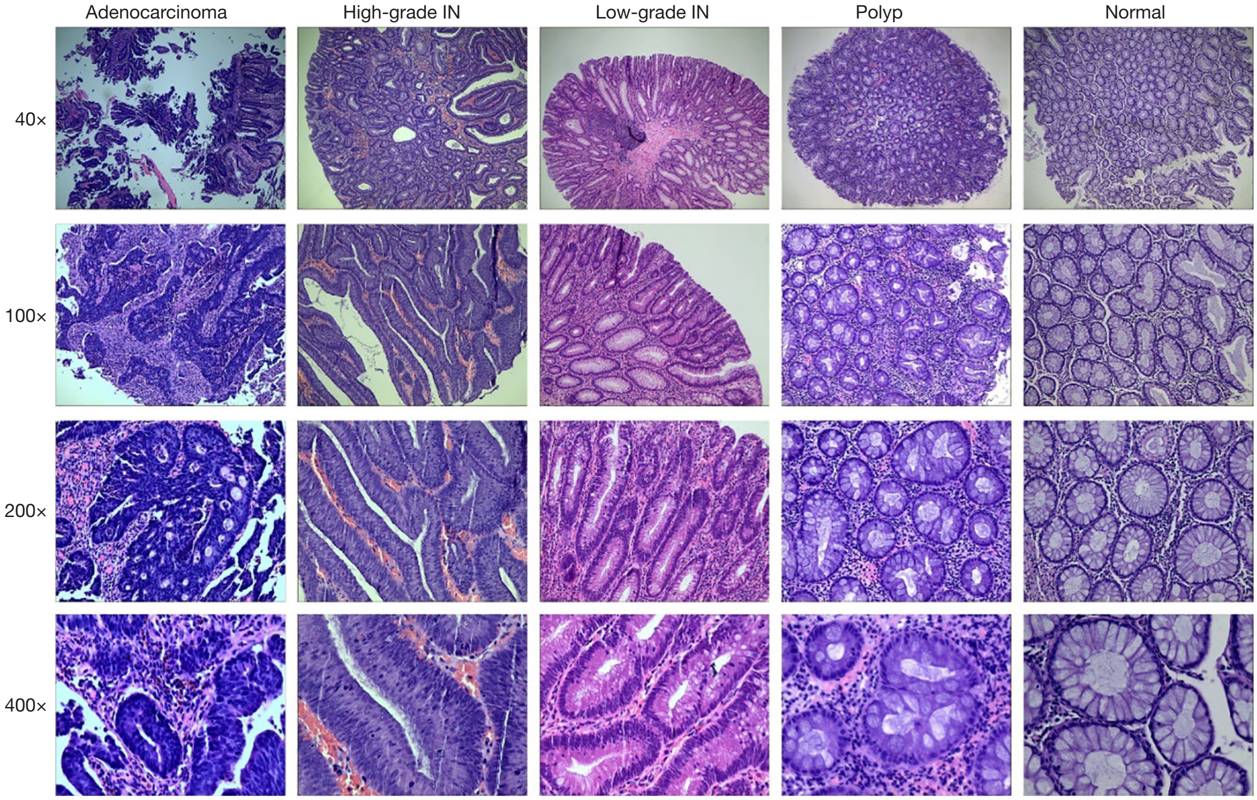

Figure 1 Representative EBHI dataset samples. EBHI, Enteroscope Biopsy Histopathological Hematoxylin and Eosin Image; IN, intraepithelial neoplasia.

maintain the same number of patches investigated by each model. The threshold value was set to 0.2. Patches in which the proportion of foreground pixels does not satisfy the 20% threshold are automatically removed. Due to tissue variations in different images, the number of patches extracted per image varies.

In *Table 1*, the row labeled "Patches cropping" indicates the data distribution after each category of patch cropping. After cropping and removing empty patches, the dataset contained a total of 58,686 patches. *Figure 2* shows the process of the Patch-Cropping Algorithm used in this paper.

Stain normalization

Stain normalization is a critical preprocessing step for pathology images, particularly in classification tasks involving tissue sections like hematoxylin and eosin (H&E)-stained images. Its primary purpose is to reduce the impact of variations in staining protocols, microscopes, or scanning equipment on image analysis. Without stain normalization, models may overfit to irrelevant staining artifacts and fail to capture essential tissue structures, thereby reducing generalization performance and classification accuracy (31). Given the color variability between staining batches in the EBHI dataset, we applied stain normalization to some images in the dataset.

Among the available normalization techniques, we used the Macenko method, which is widely regarded as one of the most effective methods for H&E-stained pathology images. The Macenko method converts RGB images into optical density (OD) space and employs singular value decomposition (SVD) for color deconvolution. This approach is highly adaptable, making it well-suited to handling staining variations across different laboratories or imaging systems. *Figure 3* provides a comparison of image samples before and after stain normalization on EBHI dataset. From *Figure 3*, we can observe that after stain normalization, the originally differently stained images have been unified, resulting in a more consistent staining appearance.

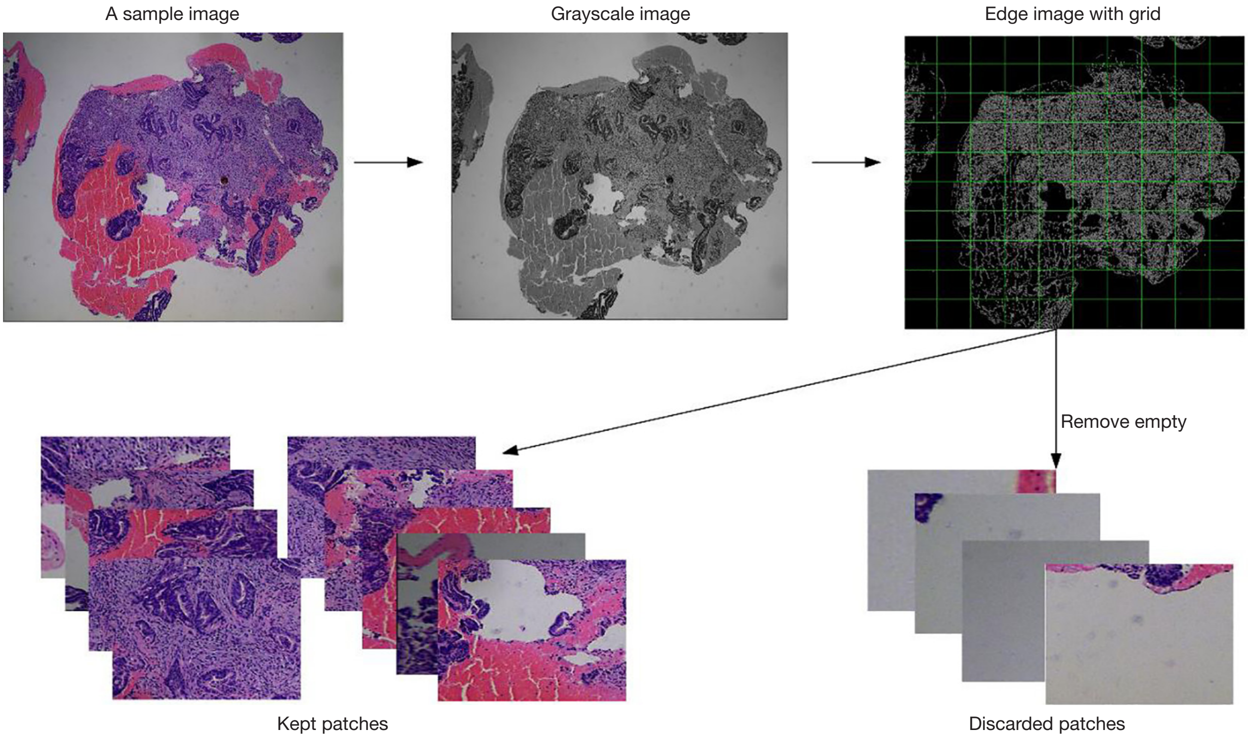

Figure 2 The workflow of the Patch-Cropping Algorithm processing.

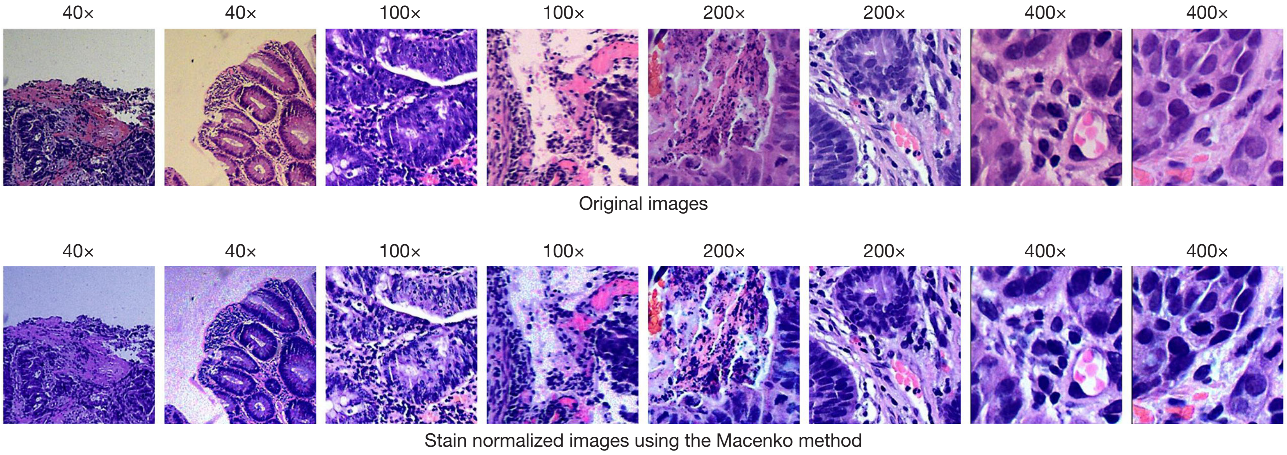

Figure 3 Original and stain normalized image samples.

Data augmentation

The generated patch dataset exhibits extreme class imbalance, with significant disparities between categories. This imbalance can introduce skewness or bias into the DL training model. To mitigate this issue, we employed various data augmentation methods to balance the class distribution within the training set.

We utilized data augmentation methods to generate a sufficiently large number of unique samples, including horizontal flip, vertical flip, random rotation, random scaling, random cropping, translation, random brightness adjustment, random contrast adjustment, random hue adjustment, Gaussian noise, and multiple combinations of variations. The extent of augmentation was determined based on the need to balance the classes within the dataset. Each category randomly applied multiple augmentation techniques according to the size of its data.

In this study, we performed data augmentation and balancing on the EBHI patch dataset to ensure that each class was balanced at different magnifications. The "Data augmentation" row in *Table 1* shows the size of each class after these processes. Overall, the size of the patch dataset was doubled through augmentation.

Classification model development

Pre-train networks

In an ensemble learning architecture, multiple pre-trained networks are required as base classifiers. To ensure the highest possible classification performance in ensemble learning, we conducted extensive evaluations of commonly used CNNs that have emerged in recent years. The chosen architectures represent popular and widely used DL models in medical image analysis. Different classification architectures are trained to ensure reliable performance. The selected architectures include ResNet (35), Xception (36), DenseNet (37) and EfficientNet (38). Finally, we selected the pre-trained networks with the excellent performance to serve as the base classifiers for the ensemble model.

TL

TL involves pre-training a model on a large-scale dataset like ImageNet, which contains over 1.2 million natural images spanning 1,000 different classes. The parameters learned during pre-training are used as the initial weights for the model, which is then fine-tuned and applied to the new target dataset. The advantage of TL is that it overcomes the challenges of limited dataset and scarcity of labeled samples (39). Given that the EBHI dataset was small and lacked sufficient data samples, the use of TL technology greatly reduced the training needs and workload.

During the fine-tuning process, only the last fully connected layer in the network was replaced with a new fully connected layer, whose size was adjusted to fit the number of categories in the new dataset. Specifically, the classification layer containing 1,000 nodes, used in the model trained on ImageNet, was replaced with a new layer corresponding to the five categories in the EBHI dataset. The other convolutional layers were frozen and kept unchanged to retain the excellent feature extraction capabilities of the various pre-trained models from the large ImageNet dataset.

Ensemble model architecture

Ensemble learning involves integrating multiple learning classifiers to accomplish classification or prediction tasks. By training different classifiers and using appropriate ensemble strategies, their predictions can complement each other to achieve high classification performance. Ensemble learning always outperforms a single classifier in terms of classification performance.

Each network structure varies, thus extracting different important features and interpreting the images differently. In this study, we evaluated the performance of several CNNs after pre-training and fine-tuning. Pre-trained CNNs with outstanding performance were selected to be combined into classifiers for the ensemble model, avoiding the limitations of any single network.

Figure 4 shows the structure of the proposed ensemble model. Initially, we fine-tuned and pretrained *n* CNNs, using TL for classification testing. After evaluating the models' classification performance, we selected the top *k* pretrained models with the best performance to serve as classifiers in the ensemble learning model. The ensemble model combines different integration strategies to make final classification predictions on histopathological images.

The other crucial aspect of ensemble learning lies in selecting an effective and appropriate ensemble strategy. Without a well-designed ensemble strategy, the full potential of ensemble learning cannot be realized. In our study, we chose three commonly used ensemble strategies to evaluate the performance of histopathology image classification for colorectal cancer, including majority voting, unweighted averaging, and stacking. We thoroughly tested each ensemble model to achieve more accurate and robust classification performance. Here are the details of each ensemble strategy:

- ❖ Majority voting: each classifier predicts a category based on its own classification results and casts one vote for that category. The class with the highest number of votes is selected as the ensemble prediction's final result.
- ❖ Unweighted averaging: the probability values of each category predicted by all the classifiers are summed and averaged. The category with the highest averaged probability is selected as the ensemble prediction category. Each classifier has an equal influence on the final result.
- ❖ Stacking: this strategy uses a meta-learner to integrate the outputs of multiple base classifiers. A new meta-learner is trained using the predictions of

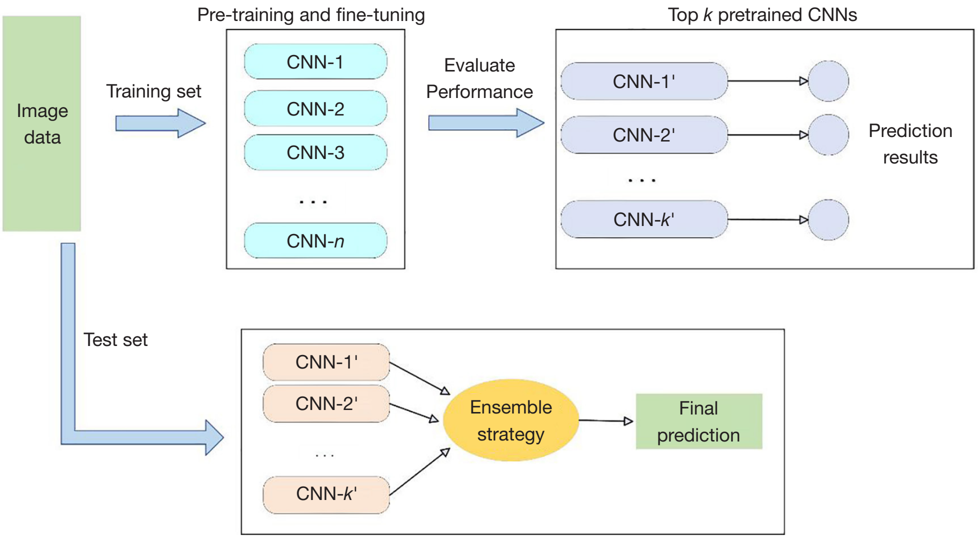

Figure 4 The architecture of the proposed ensemble model. CNN, convolutional neural network.

Table 2 Evaluation metrics

Formulas	No.
$\text{Accuracy} = \dfrac{TP + TN}{TP + TN + FP + FN}$	[1]
$\text{Precision} = \dfrac{TP}{TP + FP}$	[2]
$\text{Recall} = \dfrac{TP}{TP + FN}$	[3]
$\text{F1-score} = 2 \times \dfrac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	[4]

FN, false negative; FP, false positive; TN, true negative; TP, true positive.

multiple models as new features. The meta-learner is then used to obtain the final prediction. Naïve Bayes was selected as the meta-learner in this study because it is one of the most widely used classification schemes in stacking.

Model evaluation and visualization

Evaluation metrics

This study utilized accuracy, precision, recall, and F1-score as evaluation metrics for assessing the classification performance of the models. The terms TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. The calculation formulas are described in *Table 2*.

Prediction visualization

Gradient Weighted Class Activation Mapping (Grad-CAM) (40) is a technique used in DL to visualize and understand the decisions made by CNNs. It shows which regions of the input image are important for the network's predictions for specific classes. In this work, we used Grad-CAM to help us better understand the behavior of the model, fulfilling a key need for interpretability in DL models.

Results

Experiment setting

In this work, the experiments were conducted on a workstation with the following specifications: NVIDIA GEFORCE RTX 3090 with 24 GB of Video Random Access Memory (VRAM), Intel Core i9-11900 central processing unit (CPU) with a clock speed of 3.50 GHz, and 128 GB of Random Access Memory (RAM). The initial learning rate was set to 0.0001, and Adam was used as the optimizer. The batch size was 64, and the number of epochs was 30.

The dataset used in this experiment is the EBHI patch dataset that has been preprocessed in the Methods section. We performed the data splitting at the individual patches level to create training, validation, and testing sets,

Table 3 Performance of CNNs at 40×, 100×, 200×, and 400× on the validation set

CNNs	Accuracy (%)			
	40×	100×	200×	400×
ResNet101	93.31	92.88	94.88	94.14
ResNet152	95.28	94.95	95.43	95.27
Xception	95.32	97.66	95.79	97.04
DenseNet169	95.81	96.43	96.77	93.99
DenseNet201	94.50	96.01	96.57	95.13
EfficientNetB0	97.49	98.40	98.38†	96.87
EfficientNetB1	98.24†	98.38	98.05	97.78†
EfficientNetV2M	97.95	98.47†	98.23	97.59

†, best-achieved accuracy values. CNNs, convolutional neural networks.

Table 4 Ranking of the top three and five fine-tuned CNNs based on validation accuracy

Ranking	40×	100×	200×	400×
1	EfficientNetB1	EfficientNetV2M	EfficientNetB0	EfficientNetB1
2	EfficientNetV2M	EfficientNetB0	EfficientNetV2M	EfficientNetV2M
3	EfficientNetB0	EfficientNetB1	EfficientNetB1	Xception
4	DenseNet169	Xception	DenseNet169	EfficientNetB0
5	Xception	DenseNet169	DenseNet201	ResNet152

CNNs, convolutional neural networks.

where the image patches were randomly assigned to the different subsets. This may lead to potential data leakage bias such as image patches from the same patients were potentially included in both the training and testing sets. Nevertheless, we believe it will not affect the results much as it is a common practice to randomly split the data in machine learning. A stratified hold-out validation method was employed to ensure that each sub-dataset is mutually exclusive while maintaining a similar class distribution as much as possible. The specific partitioning ratios are as follows: 60% for the training set, 20% for the validation set, and 20% for the testing set.

The specific experimental steps are as follows: firstly, we used eight different pretrained deep CNN architectures (ResNet101, ResNet152, Xception, DenseNet169, DenseNet201, EfficientNetB0, EfficientNetB1, and EfficientNetV2M) to classify patches of colorectal cancer images at different magnifications (40×, 100×, 200×, 400×) using TL. Secondly, we chose the top three and top five DL models based on their performance in testing, and used various ensemble strategies to classify histopathological images of colorectal cancer. Finally, the ensemble models were evaluated to determine their reliability as a classification diagnostic system.

Experimental results

The accuracies of the different fine-tuned CNNs on the validation set at 40×, 100×, 200×, and 400× magnifications using the TL technique are shown in *Table 3*. EfficientNetB1 obtained the highest accuracy in two magnification datasets: 98.24% for 40× and 97.78% for 400×. EfficientNetV2M achieved the highest accuracy in the 100× (98.47%), and EfficientB0 achieved the highest accuracy in the 200× (98.38%) datasets. The best-achieved accuracy values are marked with "†" in *Table 3*.

The top three and five fine-tuned CNNs based on validation accuracy are listed in *Table 4*. The top three CNNs in terms of accuracy were variants of EfficientNet, Xception. For the top five CNNs, they were variants of

Table 5 Classification performance of fine-tuned and ensemble models on the 40× magnification factor testing set

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet101	93.33	93.63	92.77	93.20
ResNet152	95.28	93.59	97.37	95.42
Xception	95.32	95.81	94.68	95.24
DenseNet169	95.81	95.84	95.67	95.75
DenseNet201	94.51	94.03	94.94	94.48
EfficientNetB0	97.47	96.50	98.81	97.64
EfficientNetB1	98.23[†]	97.53	99.16	98.34
EfficientNetV2M	97.98	97.58	96.69	97.71
EL-MV3	98.64	97.59	98.72	98.64
EL-MV5	98.66	97.62	98.73	98.66
EL-UA3	98.77	99.16[†]	98.33	98.74
EL-UA5	98.78	98.87	98.60	98.73
EL-Sta3	98.93	98.94	98.95	98.94
EL-Sta5	99.11[†]	99.12	99.08[†]	99.10[†]

[†], best results for the metrics. "EL_MV", "EL_UA", and "EL_Sta" refer to Ensemble Model with majority voting, unweighted averaging, and stacking, respectively; and 3 or 5 after the ensemble model abbreviation indicates the top 3 or top 5 CNN models. CNN, convolutional neural network.

EfficientNet, DenseNet, Xception, and ResNet. These top-performing CNN models based on validation accuracy were selected as the base classifiers for the deep ensemble learning on the testing set.

In the ensemble learning experiments, we selected the top three and top five CNNs with the best classification performance as the base classifiers for ensemble learning models using different strategies to classify colorectal cancer histopathological images. The performance of the different fine-tuned base models and ensemble models on the testing set with 40× (*Table 5*), 100× (*Table 6*), 200× (*Table 7*), and 400× (*Table 8*) magnification factors are shown below. The best results for the metrics are marked with "[†]". The ensemble model names are abbreviated in the tables as "EL-MV", "EL-UA", and "EL-Sta" for majority voting, unweighted averaging, and stacking, respectively. Additionally, the annotation behind the abbreviated ensemble models indicates the number of individual CNNs involved in the ensemble learning. For instance, "EL-MV5" means the model ensembles the top five highest-accuracy CNNs using the majority voting strategy.

Table 5 shows the classification performance of different fine-tuned CNN models and ensemble models on the testing set with a 40× magnification factor. Among the eight fine-tuned CNN models, EfficientNetB1 achieved an accuracy of 98.23%, followed by EfficientNetV2M (97.98%) and EfficientNetB0 (97.47%). In the ensemble models, EL-Sta5 had the best performance in all indicators, especially in accuracy (99.11%) and F1-score (99.10%), which were at the highest levels. The EL-Sta3 model followed closely. Performance analysis for each metric: accuracy: EL-Sta5 (99.11%) and EL-Sta3 (98.93%) performed the best; precision: EL-UA3 (99.16%) was the highest, followed by EL-Sta5 (99.12%); recall: EL-Sta5 (99.08%) performed the best, followed by EL-Sta3 (98.95%); F1-score: EL-Sta5 (99.10%) and EL-Sta3 (98.94%) performed the best.

Table 6 shows the classification performance of different fine-tuned CNN models and ensemble models on the testing set with a 100× magnification factor. Among the eight fine-tuned CNN models, EfficientNetV2M achieved highest accuracy of 98.42%, followed by EfficientNetB0 (98.37%) and EfficientNetB1 (98.34%). In the ensemble models, EL-Sta5 had the best performance in all indicators, especially in accuracy (99.36%), recall (99.78%), which were at the highest levels. The EL-MV5 model followed with excellent indicators. Performance analysis for each metric: accuracy: EL-Sta5 (99.36%) and EL-Sta3 (99.14%) performed the best; precision: EL-MV5 (98.95%) was

Table 6 Classification performance of fine-tuned and ensemble models on the 100× magnification factor testing set

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet101	92.81	97.38	87.25	92.04
ResNet152	94.97	96.73	92.96	94.81
Xception	97.61	97.40	97.80	97.60
DenseNet169	96.34	95.89	96.65	96.26
DenseNet201	96.01	96.30	95.33	95.81
EfficientNetB0	98.37	98.01	98.84	98.42
EfficientNetB1	98.34	98.51	98.11	98.31
EfficientNetV2M	98.42†	98.16	98.83	98.50
EL-MV3	98.64	98.09	99.02	98.54
EL-MV5	99.10	98.95†	99.33	99.12
EL-UA3	98.71	98.72	99.24	98.98
EL-UA5	98.99	98.63	99.36	98.99
EL-Sta3	99.14	98.61	99.70	99.15†
EL-Sta5	99.36†	98.23	99.78†	99.00

†, best results for the metrics. "EL_MV", "EL_UA", and "EL_Sta" refer to Ensemble Model with majority voting, unweighted averaging, and stacking, respectively; and 3 or 5 after the ensemble model abbreviation indicates the top 3 or top 5 CNN models. CNN, convolutional neural network.

the highest, followed by EL-UA3 (98.72%); recall: EL-Sta5 (99.78%) performed the best, followed by EL-Sta3 (99.70%); F1-score: EL-Sta3 (99.15%) and EL-MV5 (99.12%) performed the best.

Table 7 shows the classification performance of different fine-tuned CNN models and ensemble models on the testing set with a 200× magnification factor. Among the eight fine-tuned CNN models, EfficientNetB1 achieved an accuracy of 98.24%, followed by EfficientNetV2M (98.22%). In the ensemble models, EL-Sta3 had the best accuracy (99.29%). The EL-UA5 model followed with excellent indicators. Performance analysis for each metric: accuracy: EL-Sta3 (99.29%) and EL-UA5 (99.18%) were the highest. Higher: EL-Sta5 (99.09%); precision: EL-Sta3 (99.26%) and EL-MV5 (99.07%) were the highest; recall: EL-Sta5 (99.88%) and EL-MV5 (99.30%) were the highest; F1-score: EL-Sta3 (99.29%) and EL-MV5 (99.18%) were the highest.

Table 8 shows the classification performance of different fine-tuned CNN models and ensemble models on the testing set with a 400× magnification factor. Among the eight fine-tuned CNN models, EfficientNetB1 achieved an accuracy of 97.72%, followed by EfficientNetV2M (97.58%) and Xception (96.87%). In the ensemble models, EL-Sta5 had the best performance in all indicators, especially in accuracy (98.96%), which were at the highest levels. Performance analysis for each metric: accuracy: EL-Sta5 (98.96%) and EL-Sta3 (98.72%) were the highest; precision: EL-Sta5 (98.94%) and EL-UA5 (98.89%) were the highest; recall: EL-UA3 (99.00%) and EL-Sta5 (98.95%) were the highest; F1-score: EL-Sta5 (98.94%) and EL-UA5 (98.70%) were the highest.

According to the results of these tables, the ensemble model performs the best on accuracy and is superior compared to the performance of the individual fine-tuned CNNs. The ensemble learning model that combines five base CNNs achieved the best classification performance in the experiments, outperforming its three counterparts. The reason for the improved performance lies in their diversity and ability to focus on different areas of interest within the images when more base models are included. In particular, the EL-Sta5 model performed the best on the three sub-datasets with magnifications (40×, 100×, 400×).

Discussion

Performance analysis of the base CNNs

Figure 5 shows the visualization heatmaps of the colorectal

Table 7 Classification performance of fine-tuned and ensemble models on the 200× magnification factor testing set

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet101	94.78	98.36	91.16	94.62
ResNet152	95.23	95.97	94.05	95.00
Xception	95.73	97.22	93.93	95.54
DenseNet169	96.75	96.96	96.45	96.70
DenseNet201	96.51	96.28	96.68	96.48
EfficientNetB0	97.96	98.47	97.21	97.83
EfficientNetB1	98.24†	97.73	98.33	98.03
EfficientNetV2M	98.22	97.74	98.71	98.22
EL-MV3	98.59	98.56	98.62	98.59
EL-MV5	98.89	99.07	99.30	99.18
EL-UA3	98.69	98.48	99.08	98.74
EL-UA5	99.18	98.26	98.72	98.72
EL-Sta3	99.29†	99.26†	99.28	99.29†
EL-Sta5	99.09	98.63	99.88†	99.16

†, best results for the metrics. “EL_MV”, “EL_UA”, and “EL_Sta” refer to Ensemble Model with majority voting, unweighted averaging, and stacking, respectively; and 3 or 5 after the ensemble model abbreviation indicates the top 3 or top 5 CNN models. CNN, convolutional neural network.

Table 8 Classification performance of fine-tuned and ensemble models on the 400× magnification factor testing set

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet101	94.04	92.69	95.00	93.83
ResNet152	95.17	92.90	97.70	95.24
Xception	96.87	97.43	96.60	97.01
DenseNet169	93.93	96.32	91.27	93.73
DenseNet201	95.03	93.50	96.89	95.16
EfficientNetB0	96.44	97.14	96.05	96.59
EfficientNetB1	97.72†	97.71	97.71	97.71
EfficientNetV2M	97.58	97.58	97.26	97.42
EL-MV3	98.19	97.69	98.72	98.22
EL-MV5	98.31	98.32	96.75	97.53
EL-UA3	98.14	98.43	99.00†	98.63
EL-UA5	98.69	98.89	98.51	98.70
EL-Sta3	98.72	98.72	97.59	98.15
EL-Sta5	98.96†	98.94†	98.95	98.94†

†, best results for the metrics. “EL_MV”, “EL_UA”, and “EL_Sta” refer to Ensemble Model with majority voting, unweighted averaging, and stacking, respectively; and 3 or 5 after the ensemble model abbreviation indicates the top 3 or top 5 CNN models. CNN, convolutional neural network.

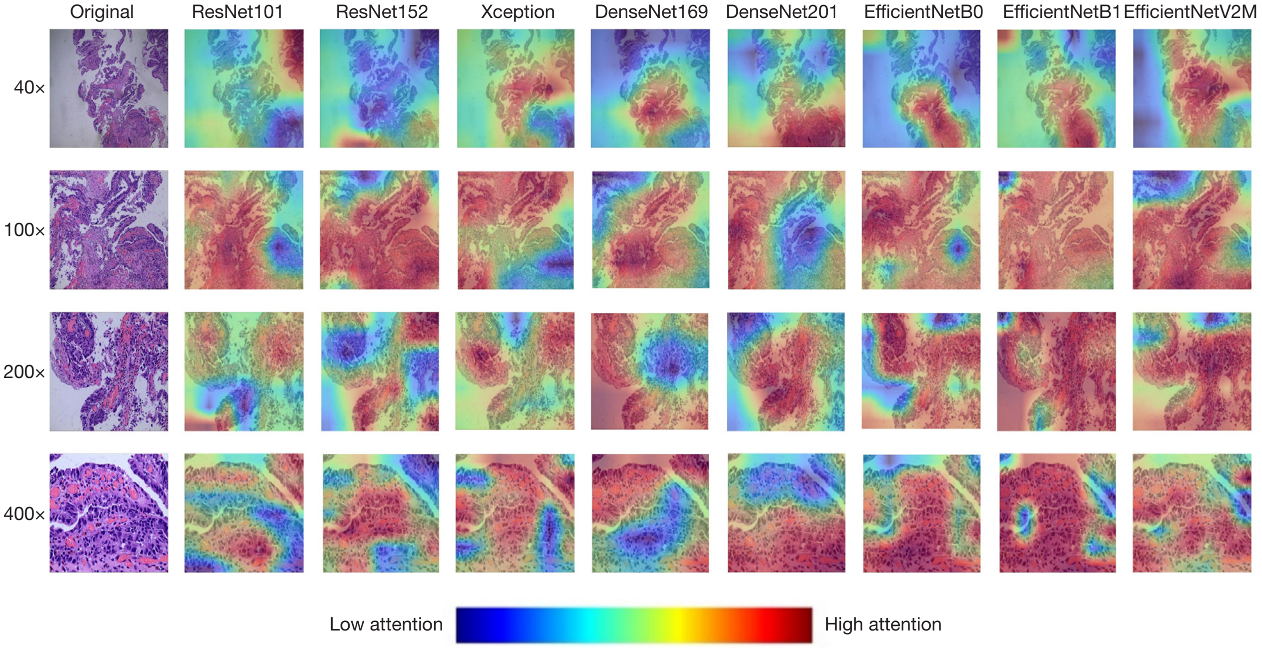

Figure 5 Samples and corresponding Grad-CAM heatmaps generated by eight CNN models under the four magnification factor images. Grad-CAM, Gradient Weighted Class Activation Mapping; CNN, convolutional neural network.

cancer histopathological sample patches on the eight base CNN models generated using Grad-CAM. Heatmaps are used to show the critical regions in the histopathology images where different models perform classification. The diversity of CNN structures was assessed by analyzing the variation in the response of the heatmaps. Fluctuations in heatmap intensity values were correlated with how much attention the models paid to the highlighted decision locations. Red color indicates more attention, and blue color indicates less attention. The regions of attention of the eight CNN models shown in the heatmaps can be well explained by the fact that EfficientNet, Xception, and DenseNet can give better results in terms of classification accuracy. By highlighting important regions, these architectures focus on wider activation regions, and image anomaly regions are covered more accurately.

Performance comparison with the state-of-the-art results

Table 9 provides a comparison of model performance between this study and recent state-of-the-art studies on the EBHI dataset. In the experiments of the compared papers, colorectal cancer images were divided into two categories: benign (normal, polyps, Low-grade IN) and malignant (high-grade IN, adenocarcinoma). The experiments were conducted only on images with a magnification of 200×. The training set, validation set, and test set were divided in a 4:4:2 ratio. Our experiment followed the experimental settings of the compared papers, and the models with the best performance (a based CNN and ensemble models) were selected for comparison of classification performance.

The experimental results in *Table 9* show that the classification performance of the proposed single CNN model, EfficientNetB1, and all ensemble models outperforms those in references (31) and (34). Specifically, the EfficientNetB1 model used in this study surpasses the performance metrics of the model in reference (34) by 1.97% to 4.77% across all measures. EfficientNetB1 achieves a classification accuracy of 98.57%, which is 3.2% higher than the 95.37% accuracy obtained by the single VGG16 model in reference (34).

When comparing single models and TL techniques on the EBHI dataset, our models consistently outperforms the one in reference (34). This may be attributed to two main factors: first, different CNN architectures have varying capabilities for feature extraction, leading to differences in classification performance. EfficientNetB1 proves to be better suited for the EBHI dataset than VGG16, demonstrating the importance of selecting the right model architecture. Second, this study applies data

Table 9 Comparison of classification performance between the proposed models and the comparison model with a dataset magnification of 200×

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
VGG16 (34)	95.37	96.40	94.30	95.30
CDSS (31)	91.1	88.74	94.36	91.46
EfficientNetB1	98.57	98.37	98.77	98.57
EL-MV3	98.96	98.40	99.31	98.85
EL-MV5	99.01	98.65	99.80	99.11
EL-UA3	98.97	97.79	99.77	98.78
EL-UA5	99.02	98.63	99.77	99.11
EL-Sta3	99.08	98.63	99.83	99.14
EL-Sta5	99.26	99.07	99.91	99.44

"EL_MV", "EL_UA", and "EL_Sta" refer to Ensemble Model with majority voting, unweighted averaging, and stacking, respectively; and 3 or 5 after the ensemble model abbreviation indicates the top 3 or top 5 CNN models. CNN, convolutional neural network.

preprocessing techniques, including patch cropping and data augmentation, to significantly increase the size of the training set. Patch cropping also removes irrelevant information from images, and data balancing ensures an even distribution across all categories. These preprocessing steps were not used in references (31) and (34), which could be a key reason for the superior performance of our model.

In the comparison of ensemble model classification performance, all the models proposed in this paper outperform those in reference (31). Specifically, our EL-Sta5 model achieves a top classification accuracy of 99.26%, which is 8.16% higher than the 91.1% reported in reference (31). The strong performance of our ensemble models may be due to the complementary nature of the five classifiers, enabling more effective extraction of low-level features like contours, colors, and textures, as well as high-level semantic features. In contrast, the ensemble model in reference (31) consists of only two base classifiers (ConvNeXt-base and ConvNeXt-Tiny) combined using a simple averaging strategy. As ConvNeXt is known for its simplified and efficient architecture, the feature extraction capacity of these base classifiers is limited and may only excel in specific applications. Additionally, reference (31) did not explore other ensemble strategies, which could explain why its overall classification performance on the EBHI dataset is lower than that of the models proposed in this paper.

Extended experiment

Our proposed ensembles models achieved the best classification performance among current the state-of-the-art studies on the EBHI dataset. To demonstrate the effectiveness and robustness of the proposed ensemble models and to prove that there is no sample limitation, we further conducted extended experiments on different colon cancer histopathology datasets. Two extended datasets were used: LC25000 and NCT-CRC-HE-100K. The LC25000 dataset is a histopathology image dataset of lung and colon cancer. It is divided into five categories with 5,000 images in each category, totaling 25,000 images. The NCT-CRC-HE-100K dataset is a H&E-stained histological image dataset of human colorectal cancer and normal tissues, containing 100,000 non-overlapping image patches in nine tissue classes.

In the extended experiments, we used the same preprocessing steps, models, and experimental setup as described in the Results section. The data splits for both datasets were 60% for training, 20% for validation, and 20% for testing. We selected three ensemble models proposed in this paper: EL-MV5, EL-UA5, and EL-Sta5. Based on the results from the previous section, we observed that ensemble learning models comprising five base CNN models achieved the best classification performance across multiple subdatasets. Therefore, we chose ensemble models consisting of five base classifiers. These base classifiers were the top five CNN networks from the EBHI dataset: EfficientNetB0, EfficientNetB1, EfficientNetV2M, Xception, and DenseNet169. We retrained them using the two extended datasets separately. Then the base classifiers were combined with three different ensemble strategies to

Table 10 Accuracy of different models on two datasets

Models	Accuracy (%)	
	LC25000	NCT-CRC-HE-100K
EL-MV5	99.67	99.68
EL-UA5	99.74	99.73
EL-Sta5	99.78	99.80
Mehmood *et al.* (20)	98.40	–
Omar *et al.* (41)	99.44	–
Ghosh *et al.* (32)	–	99.13
Mohammed *et al.* (26)	–	99.48

"EL_MV", "EL_UA", and "EL_Sta" refer to Ensemble Model with majority voting, unweighted averaging, and stacking, respectively; and 3 or 5 after the ensemble model abbreviation indicates the top 5 CNN models. CNN, convolutional neural network.

form the three ensemble models.

For each model, the CNN networks were fine-tuned, and the output SoftMax layers were adjusted according to the number of classes in the respective datasets. The LC25000 dataset contains five classes, so the output layer was set to five nodes, while the NCT-CRC-HE-100K dataset has nine classes, so the output layer was set to nine nodes. The extended experimental process involved training and validating the three fine-tuned models on the two datasets, followed by classification testing. The classification accuracy results for both datasets are shown in *Table 10*.

As shown in *Table 10*, our proposed ensemble models demonstrate good generalization ability, achieving the highest overall accuracy of 99.78% and 99.80% on the two datasets using EL-Sta5, respectively. The proposed ensemble model EL-Sta5 has higher classification accuracies than the state-of-the-art methods on both datasets. Additionally, more than 99.6% overall accuracy was also obtained using EL-MV5 and EL-UA5. The proposed ensemble models show excellent performance in classifying different colorectal cancer histopathology datasets. These results demonstrate that our proposed ensemble models have better classification performance and generalization in dealing with different histopathological image datasets and multi-class classification tasks of colon cancer.

Limitations of our proposed study

Although our proposed ensemble models managed to achieve state-of-the-art results on the EBHI dataset as well as the two extended datasets, this work has several limitations. Compared to single models, ensemble learning combines multiple base models, leading to increased training costs, higher consumption of computational resources and time, and greater model complexity, which in turn raises the difficulty of debugging. To address the challenges related to computational resources and time, leveraging multiple graphic processing units (GPUs) for parallel processing can accelerate training and reduce time consumption. When tackling model complexity, it is crucial to selectively integrate a small number of efficient base models and apply effective combination strategies, such as stacking. The improvement in ensemble model performance is not driven by the quantity of base classifiers, but rather by their quality and the effectiveness of the combination method used. In addition, the initial weights and hyperparameters are transferred from the ImageNet datasets that have very different image features and characteristics from the histopathological images, the model parameters may not be optimal for pathology image classification. To address the challenge of improving knowledge learning outcomes, optimization can be achieved by thoroughly investigating the impact of specific dataset characteristics on pre-trained model performance. By optimizing the model's initial weights, hyperparameter settings and the use of dynamic augmentation, the accuracy of cancer detection can be further enhanced.

Conclusions

This study developed deep ensemble learning models for multi-class colorectal cancer classification using TL with pre-trained networks such as ResNet, DenseNet, Xception, and EfficientNet. We fine-tuned and pre-trained multiple CNN models using colorectal cancer histopathological image patches magnified at 40×, 100×, 200×, and 400×, and employed three ensemble strategies to build the ensemble models. Experimental results show that EfficientNet variant performed exceptionally well as a single model, achieving the highest classification accuracy. Ensemble learning models, built from the top five classifiers based on classification performance, achieved the state-of-the-art results. The EL-Sta5 ensemble model performed best on the 100× magnification test set, with an accuracy of 99.36%. The study highlights that ensemble learning significantly improves performance in colorectal cancer histopathological image classification compared to individual TL models. Additionally, we enhanced the dataset by performing patch cropping, data augmentation, and balancing operations, which improved the model's generalization and reduced the risk of overfitting. Future research will focus on reducing the computational costs of ensemble models and improving their computational efficiency. Tailored TL strategies should be designed to address the specific characteristics of histopathological images, further enhancing the model's knowledge learning capability. In addition, more advanced models like Vision Transformer (e.g., ViT 32) and larger architectures would be studied to enhance model diversity, performance and relevance in the future.

Acknowledgments

The authors would like to extend their heartfelt gratitude to the Guangxi Key Laboratory of Big Data and Guangxi Key Laboratory of Seaward Economic Intelligent System Analysis and Decision-making for their generous support, which provided the essential facilities and funding for our project. Additionally, we would like to acknowledge the authors of the EBHI, LC25000 and NCT-CRC-HE-100K datasets for making their datasets publicly available. The models and some parameters can be accessed at the following link: https://github.com/MissLingxizi/CEM_moldel.git.

Footnote

Reporting Checklist: The authors have completed the TRIPOD+AI reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-1641/rc

Funding: This study was supported by Guangxi First-Class Discipline Statistics Construction Project Fund, and Guangxi Higher Education Institutions Young and Middle-aged Teachers' Basic Research Capacity Enhancement Project (No. 2024KY0669).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-1641/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study utilized publicly available online datasets and did not involve any human or animal subjects. Therefore, ethical approval was not required for this study.

References

1. Mengash HA, Alamgeer M, Maashi M, Othman M, Hamza MA, Ibrahim SS, Zamani AS, Yaseen I. Leveraging Marine Predators Algorithm with Deep Learning for Lung and Colon Cancer Diagnosis. Cancers (Basel) 2023.
2. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, Bray F. Cancer statistics for the year 2020: An overview. Int J Cancer 2021. [Epub ahead of print]. doi: 10.1002/ijc.33588.
3. Chhikara BS, Parang K. Chemical Biology LETTERS Global Cancer Statistics 2022: the trends projection analysis. Chem Biol Lett 2023;10:451.
4. Yin Z, Yao C, Zhang L, Qi S. Application of artificial intelligence in diagnosis and treatment of colorectal cancer: A novel Prospect. Front Med (Lausanne) 2023;10:1128084.
5. Li J, Liu J, Yue H, Cheng J, Kuang H, Bai H, Wang Y, Wang J. DARC: Deep adaptive regularized clustering for histopathological image classification. Med Image Anal 2022;80:102521.
6. Li X, Li C, Rahaman MM, Sun H, Li X, Wu J, Yao Y, Grzegorzek M. A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. Artif Intell Rev 2022;55:4809-78.
7. Senan EM, Alsaade FW, Al-Mashhadani MIA, Aldhyani THH, Al-Adhaileh MH. Classification of histopathological images for early detection of breast cancer using deep learning. Journal of Applied Science and Engineering 2021;24:323-9.
8. Ben Hamida A, Devanne M, Weber J, Truntzer C, Derangère V, Ghiringhelli F, Forestier G, Wemmert C.

Deep learning for colon cancer histopathological images analysis. Comput Biol Med 2021;136:104730.

9. Li C, Liu J, Tang J. Simultaneous segmentation and classification of colon cancer polyp images using a dual branch multi-task learning network. Math Biosci Eng 2024;21:2024-49.

10. Xu H, Cha YJ, Clemenceau JR, Choi J, Lee SH, Kang J, Hwang TH. Spatial analysis of tumor-infiltrating lymphocytes in histological sections using deep learning techniques predicts survival in colorectal carcinoma. J Pathol Clin Res 2022;8:327-39.

11. Lou J, Xu J, Zhang Y, Sun Y, Fang A, Liu J, Mur LAJ, Ji B. PPsNet: An improved deep learning model for microsatellite instability high prediction in colorectal cancer from whole slide images. Comput Methods Programs Biomed 2022;225:107095.

12. Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. Sci Rep 2020;10:1504.

13. Koo JC, Ke Q, Hum YC, Goh CH, Lai KW, Yap WS, Tee YK. Non-annotated renal histopathological image analysis with deep ensemble learning. Quant Imaging Med Surg 2023;13:5902-20.

14. Al-Mamun Provath M, Deb K, Jo KH. Classification of Lung and Colon Cancer Using Deep Learning Method. In: Na I, Irie G. editors. Frontiers of Computer Vision. IW-FCV 2023. Communications in Computer and Information Science, vol 1857. Singapore: Springer; 2023:56-70.

15. Neto PC, Montezuma D, Oliveira SP, Oliveira D, Fraga J, Monteiro A, Monteiro J, Ribeiro L, Gonçalves S, Reinhard S, Zlobec I, Pinto IM, Cardoso JS. An interpretable machine learning system for colorectal cancer diagnosis from pathology slides. NPJ Precis Oncol 2024;8:56.

16. Kumar A, Vishwakarma A, Bajaj V. CRCCN-Net: Automated framework for classification of colorectal tissue using histopathological images. Biomedical Signal Processing and Control 2023;79:104172.

17. Tummala S, Kadry S, Nadeem A, Rauf HT, Gul N. An Explainable Classification Method Based on Complex Scaling in Histopathology Images for Lung and Colon Cancer. Diagnostics (Basel) 2023;13:1594.

18. Kalatzis D, Spyratou E, Karnachoriti M, Kouri MA, Orfanoudakis S, Koufopoulos N, Pouliakis A, Danias N, Seimenis I, Kontos AG, Efstathopoulos EP. Advanced Raman Spectroscopy Based on Transfer Learning by Using a Convolutional Neural Network for Personalized Colorectal Cancer Diagnosis. Optics 2023;4:310-20.

19. Luo R, Bocklitz T. A systematic study of transfer learning for colorectal cancer detection. Informatics in Medicine Unlocked 2023;40:101292.

20. Mehmood S, Ghazal TM, Khan MA, Zubair M, Naseem MT, Faiz T. Malignancy Detection in Lung and Colon Histopathology Images Using Transfer Learning with Class Selective Image Processing. IEEE Access 2022;10:25657-68.

21. Yong MP, Hum YC, Lai KW, Lee YL, Goh CH, Yap WS. Histopathological Cancer Detection Using Intra-Domain Transfer Learning and Ensemble Learning. IEEE Access 2023;12:1434-57.

22. Yong MP, Hum YC, Lai KW, Lee YL, Goh CH, Yap WS, Tee YK. Histopathological Gastric Cancer Detection on GasHisSDB Dataset Using Deep Ensemble Learning. Diagnostics (Basel) 2023;13:1793.

23. Ohata EF, Chagas JVS das, Bezerra GM, Hassan MM, de Albuquerque VHC, Filho PPR. A novel transfer learning approach for the classification of histological images of colorectal cancer. J Supercomput 2021;77:9494-519.

24. Sharkas M, Attallah O. Color-CADx: a deep learning approach for colorectal cancer classification through triple convolutional neural networks and discrete cosine transform. Sci Rep 2024;14:6914.

25. Talukder MdA, Islam MdM, Uddin MA, Akhter A, Hasan KF, Moni MA. Machine Learning-based Lung and Colon Cancer Detection using Deep Feature Extraction and Ensemble Learning. Expert Syst Appl 2022;205:117695.

26. Mohammed M, Mwambi H, Mboya IB, Elbashir MK, Omolo B. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Sci Rep 2021;11:15626.

27. Li C, Xue D, Kong F, Hu Z, Chen H, Yao Y, Sun H, Zhang L, Zhang J, Jiang T, Yuan J, Xu N. Cervical Histopathology Image Classification Using Ensembled Transfer Learning. In: Pietka E, Badura P, Kawa J, Wieclawek W. editors. Information Technology in Biomedicine. ITIB 2019. Advances in Intelligent Systems and Computing, vol 1011. Springer, Cham; 2019:26-37.

28. Zhu Z, Wang S, Zhang Y. ROENet: A ResNet-Based Output Ensemble for Malaria Parasite Classification. Electronics (Basel) 2022;11:2040.

29. Muller D, Soto-Rey I, Kramer F. An Analysis on Ensemble Learning Optimized Medical Image Classification with Deep Convolutional Neural Networks. IEEE Access 2022;10:66467-80.

30. Khazaee Fadafen M, Rezaee K. Ensemble-based multi-

tissue classification approach of colorectal cancer histology images using a novel hybrid deep learning framework. Sci Rep 2023;13:8823.

31. Yengec-Tasdemir SB, Aydin Z, Akay E, Dogan S, Yilmaz B. Improved classification of colorectal polyps on histopathological images with ensemble learning and stain normalization. Comput Methods Programs Biomed 2023;232:107441.
32. Ghosh S, Bandyopadhyay A, Sahay S, Ghosh R, Kundu I, Santosh KC. Colorectal Histology Tumor Detection Using Ensemble Deep Neural Network. Engineering Applications of Artificial Intelligence 2021;100:104202.
33. Iqbal S, Qureshi AN, Alhussein M, Aurangzeb K, Kadry S. A Novel Heteromorphous Convolutional Neural Network for Automated Assessment of Tumors in Colon and Lung Histopathology Images. Biomimetics (Basel) 2023;8:370.
34. Hu W, Li C, Rahaman MM, Chen H, Liu W, Yao Y, Yao Y, Sun H, Grzegorzek M, Li X. EBHI: A new Enteroscope Biopsy Histopathological H&E Image Dataset for image classification evaluation. Physica Medica 2023;107:102534.
35. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition [Internet]. Available online: http://image-net.org/challenges/LSVRC/2015/
36. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 21-26 July 2017; Honolulu, HI, USA. IEEE; 2017:1251-8.
37. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks [Internet]. Available online: https://github.com/liuzhuang13/DenseNet
38. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: International Conference on Machine Learning. PMLR; 2019:6015-114.
39. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Trans Knowl Data Eng. 2010;22:1345-59.
40. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2017 IEEE International Conference on Computer Vision (ICCV); 22-29 October 2017; Venice, Italy. IEEE; 2017:618-26.
41. Omar LT, Hussein JM, Omer LF, Qadir AM, Ghareb MI. Lung and Colon Cancer Detection Using Weighted Average Ensemble Transfer Learning. 2023 11th International Symposium on Digital Forensics and Security (ISDFS); 11-12 May 2023; Chattanooga, TN, USA. IEEE; 2023:1-7.

Cite this article as: Ke Q, Yap WS, Tee YK, Hum YC, Zheng H, Gan YJ. Advanced deep learning for multi-class colorectal cancer histopathology: integrating transfer learning and ensemble methods. Quant Imaging Med Surg 2025;15(3):2329-2346. doi: 10.21037/qims-24-1641