STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, March 2026, pp 1995–2012.
Published online in International Academic Press (www.IAPress.org)

Optimized Deep Ensemble Framework for Colorectal Polyp Detection and Clinical Deployment Design

Fayza Elshorbagy*, Ehab Elsalamouny, Marwa F. Mohamed

Department of Computer Science, Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt

Abstract Early and accurate detection of colorectal polyps is critical for reducing colorectal cancer risk and improving patient outcomes. This paper introduces an ensemble deep transfer learning framework with Bayesian hyperparameter optimization for robust colorectal polyp classification. The method combines three state-of-the-art backbones—ResNet50, EfficientNetB0, and InceptionV3—whose outputs are fused via probability averaging to improve reliability. Stratified 10-fold cross-validation provides unbiased performance estimates, while Bayesian optimization fine-tunes model parameters for high accuracy and efficiency. Experiments on three benchmark datasets demonstrate excellent results, achieving 99.56% accuracy on CP-CHILD-A, 99.40% on CP-CHILD-B, and 92.80% on Kvasir V2. To illustrate clinical usability, we also designed user interface prototypes as a computer-aided diagnostic (CAD) system, showing how the framework could be integrated into real-world screening workflows. These results highlight the potential of the proposed approach for real-time, clinically deployable colorectal polyp detection.

Keywords Colorectal polyp detection, Ensemble learning, Bayesian optimization, Transfer learning, Stratified cross-validation, CAD system

DOI: 10.19139/soic-2310-5070-3006

1. Introduction

Colorectal cancer (CRC) remains one of the leading causes of cancer-related morbidity and mortality worldwide, with its incidence steadily increasing in many regions [1]. Early detection of precancerous polyps during colonoscopy is critical to prevent disease progression and improve patient outcomes. However, manual inspection of colonoscopy images is time consuming, operator dependent and subject to variability in precision, especially in cases of small, flat, or sessile polyps [2]. These challenges highlight the need for computer-aided solutions that can help clinicians detect and classify colorectal polyps more accurately and consistently.

In recent years, deep learning — especially convolutional neural networks (CNNs) — has greatly improved the way we classify medical images and detect lesions [3, 4]. CNN-based methods have shown very promising results in identifying and segmenting polyps [5, 6]. However, relying on a single model can have some drawbacks. Such models may overfit to the training data, struggle to perform well on new or different datasets, and lose accuracy when faced with unseen cases. Another challenge is choosing the best hyperparameters for these models. This process is often done manually, requires a lot of computing power, and can make it difficult to reproduce results or scale the approach.

Transfer learning has emerged as a powerful paradigm for medical image analysis, enabling the reuse of feature representations learned from large-scale natural image datasets such as ImageNet [23]. By initializing models with pretrained weights, transfer learning mitigates the challenge of limited annotated medical datasets,

*Correspondence to: Fayza Elshorbagy (Email: UGS.140151@ci.suez.edu.eg). Department of Computer Science, Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt (41522).

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2026 International Academic Press

accelerates convergence, and often yields superior generalization performance compared to training from scratch. Numerous studies have demonstrated that fine-tuning pretrained CNN backbones leads to significant improvements in classification, detection, and segmentation tasks across a wide range of medical imaging modalities [24, 25].

Ensemble learning is a powerful way to overcome the weaknesses of using a single model by combining the strengths of several models. In medical image analysis, this approach helps reduce overfitting, lowers prediction variability, and improves performance on different types of data. By averaging or voting on the outputs of multiple models, ensembles produce more stable and trustworthy results — an important benefit for clinical applications like colorectal polyp detection. Recent research shows that ensembles of CNNs often achieve better accuracy and sensitivity than any single model, especially when working with difficult or imbalanced datasets [26].

To address these limitations, this work proposes an ensemble deep transfer learning framework that integrates three state-of-the-art CNN backbones: ResNet50 [27], EfficientNetB0 [28], and InceptionV3 [29] and optimizes them using Bayesian hyperparameter tuning to achieve robust and reproducible classification performance. Probability-based ensemble fusion is employed to improve prediction reliability, and stratified 10-fold cross-validation is used to obtain unbiased performance estimates. Additionally, we designed user interface prototypes for a computer-aided diagnostic (CAD) system, demonstrating how the proposed framework can be integrated into clinical workflows for real-time polyp screening.

The contributions of this work can be summarized as follows:

1. We present a robust ensemble transfer learning framework that integrates multiple CNN architectures to enhance colorectal polyp classification performance.
2. We employ Bayesian optimization to systematically tune hyperparameters, ensuring reproducibility and improving overall model efficiency.
3. We conduct a comprehensive ablation study to demonstrate the effectiveness of the ensemble strategy compared to individual backbone models, confirming its superior generalization capability.
4. We validate the proposed framework on three benchmark datasets—CP-CHILD-A, CP-CHILD-B, and Kvasir V2—achieving accuracies of 99.56%, 99.40%, and 92.80%, respectively.
5. We propose a conceptual CAD system interface, emphasizing the framework's practical potential for real-world clinical deployment.

2. Related Work

In this section, we review existing research related to colorectal polyp detection and classification, focusing on three key areas : deep learning approaches for polyp classification, ensemble and hybrid frameworks, and optimization strategies for medical images.

Deep learning, especially CNN-based architectures, has transformed the field of colorectal image analysis. Studies such as Raseena et al. [7] have benchmarked multiple CNN backbones including VGG19, ResNet50, and MobileNetV3 on four major datasets (PolypsSet, CP-CHILD-A, CP-CHILD-B, and Kvasir V2), demonstrating that transfer learning with fine-tuning can yield high classification accuracy. Other works proposed transformer-based solutions, e.g., DeepCPD [8], which leverages multi-head self-attention for robust polyp categorization. Similarly, John Lewis et al. [9] combined CNN and transformer modules in PSNet, achieving state-of-the-art performance for polyp segmentation with clearer boundary detection.

Mukhtorov et al. [10] developed an explainable deep learning approach for endoscopic image classification using ResNet152 with Grad-CAM visualizations. Their method achieved 98.28% training and 93.46% validation accuracy, highlighting its effectiveness for medical image analysis. Zhang et al. [11] developed SSD-GPNet, a CNN based on the Single Shot MultiBox Detector (SSD) architecture, for real-time gastric polyp detection. These studies collectively show that deep learning models can effectively capture discriminative features for polyp detection but also face challenges like overfitting, poor generalization to new datasets.

To mitigate generalization gaps, ensemble learning has been widely explored. Auzine et al. [12] combined InceptionV3, InceptionResNetV2, and VGG16 into an ensemble trained on the Kvasir dataset, achieving superior accuracy through model diversity. Hybrid frameworks have also emerged, integrating CNNs with spatial attention mechanisms [13] to leverage both local and global feature dependencies. Although these approaches improve accuracy, they are often computationally expensive. To address this, we combine model outputs through probability averaging for stable predictions and use Bayesian optimization to efficiently explore the hyperparameter space. This allows us to find near-optimal configurations with fewer training runs, saving time and computational resources while improving reproducibility.

Optimization methods are widely used to improve model performance in medical image classification. Traditional grid and random search methods can be slow and often overfit. Researchers have explored metaheuristic approaches such as Genetic Algorithms [14], Grey Wolf Optimization [15], and Harris Hawk Optimization [16] to tune models or select features, reporting better accuracy and lower complexity. Others, like Elsayed et al. [17] and Mohamed et al. [18], used swarm intelligence and hybrid optimizers for tasks such as COVID-19 and breast cancer image analysis. In this work, we use Bayesian optimization, which builds a probabilistic model to explore the search space efficiently, leading to consistent tuning with faster convergence.

Despite impressive results from CNNs, transformers, and ensembles, most studies still face three main challenges: overfitting, lack of reproducibility, and limited focus on real-time usability. Our work addresses these by combining transfer learning, probability-based ensemble fusion, and Bayesian tuning into a practical and clinically oriented solution.

3. Methodology

This section describes the proposed framework for the classification of colorectal polyps. Our approach integrates three key components: (i) dataset preparation and preprocessing, (ii) CNN backbone training with Bayesian-optimized hyperparameters, and (iii) ensemble fusion for final predictions. An overview of the methodology is illustrated in Figure 1 and the proposed framework is summarized in Algorithm 1.

3.1. Dataset and Preprocessing

Images from 1,600 pediatric colonoscopies (ages 0–18 years) were used to construct the **CP-CHILD-A** and **CP-CHILD-B** datasets. The CP-CHILD-A dataset consists of 8,000 colonoscopy images, including 1,000 polyp and 7,000 non-polyp samples. The CP-CHILD-B dataset contains 1,500 images, with 400 polyp and 1,100 non-polyp samples. Both datasets are publicly available and can be accessed at [19]. Table 1 summarizes the datasets used in this study. The Kvasir V2 dataset originally comprised eight classes of high-quality endoscopic images of the gastrointestinal (GI) tract, annotated and verified by medical professionals [38]. Collected using high-definition endoscopes under the Vestre Viken Health Trust, which serves a population of approximately 470,000 people, the dataset was refined in this study to include only two relevant classes—polyps and nonpolyps—directly associated with colorectal conditions. The resulting balanced subset of 2,000 images (1,000 per class) is publicly available at [41].

Table 1. Summary of CP-CHILD and Kvasir V2 datasets used in this study.

Dataset	Total Images	Polyp Images	Non-Polyp Images
CP-CHILD-A	8,000	1,000	7,000
CP-CHILD-B	1,500	400	1,100
Kvasir V2	1,000	1,000	2,000

All images were resized to $224 \times 224 \times 3$ pixels to match the input requirements of the CNN backbones. To enhance generalization and minimize overfitting, we applied *on-the-fly* data augmentation [20] during training.

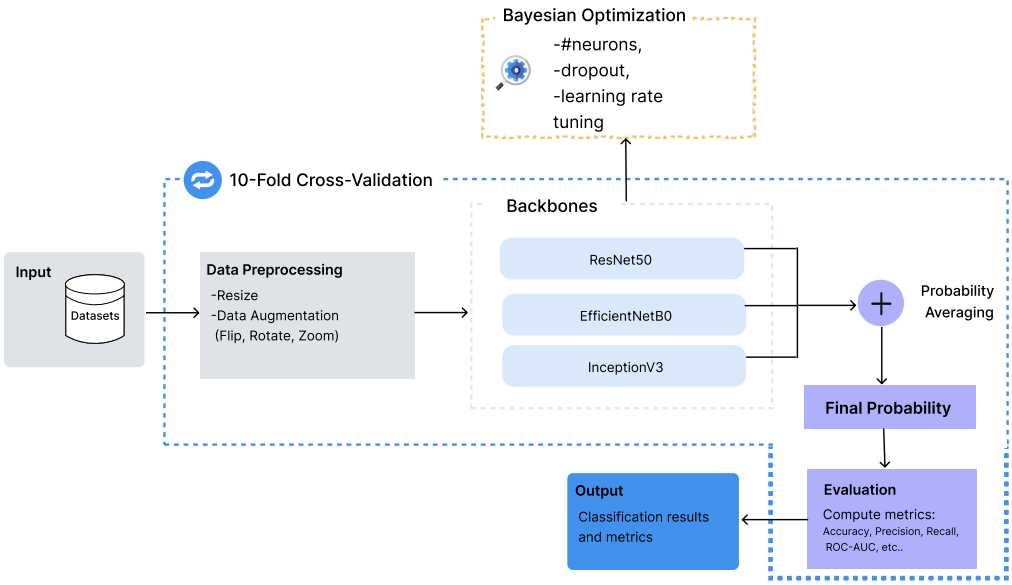


Figure 1. The workflow of the proposed framework.

Instead of permanently enlarging the dataset, transformations were performed dynamically at runtime so that each training epoch encountered slightly different image variations. This strategy exposed the network to a richer set of visual patterns, helping it become more robust to natural variations in clinical settings.

The augmentation process included random horizontal flips to simulate changes in colonoscope orientation, small rotations of up to $\pm 8^{\circ}$ to account for camera tilts, and zoom-in operations of up to 8% to mimic variations in colonoscope proximity and polyp size. In some cases, minor intensity adjustments were applied to reproduce differences in lighting conditions. By combining these transformations, the effective training set size was virtually expanded, making the model less sensitive to spatial orientation, scale, and illumination changes. Figure 2 illustrates examples of the augmentation process applied to a single polyp image, showing how variability is introduced without permanently increasing dataset size.

All data augmentations were applied dynamically and exclusively within each training fold after data splitting. For each iteration of the stratified 10-fold cross-validation, augmentation operations were performed *on-the-fly* only on the training subset of that fold. The validation subset remained completely separate and was never augmented, ensuring that no augmented variant of any image could appear in both the training and validation sets. This strict separation effectively prevented data leakage and preserved the integrity of the cross-validation procedure.

3.2. Backbone Models

Our framework integrates three well-established convolutional neural network backbones pretrained on ImageNet: ResNet50 , EfficientNetB0 , and InceptionV3. These architectures were selected because they complement each other in terms of representation power, efficiency, and ability to capture multi-scale features, making them highly suitable for colorectal polyp classification.

ResNet50 introduces residual connections that allow very deep networks to train effectively by mitigating the vanishing gradient problem. This enables the extraction of rich hierarchical features, which is crucial for capturing subtle visual differences in medical images [27].

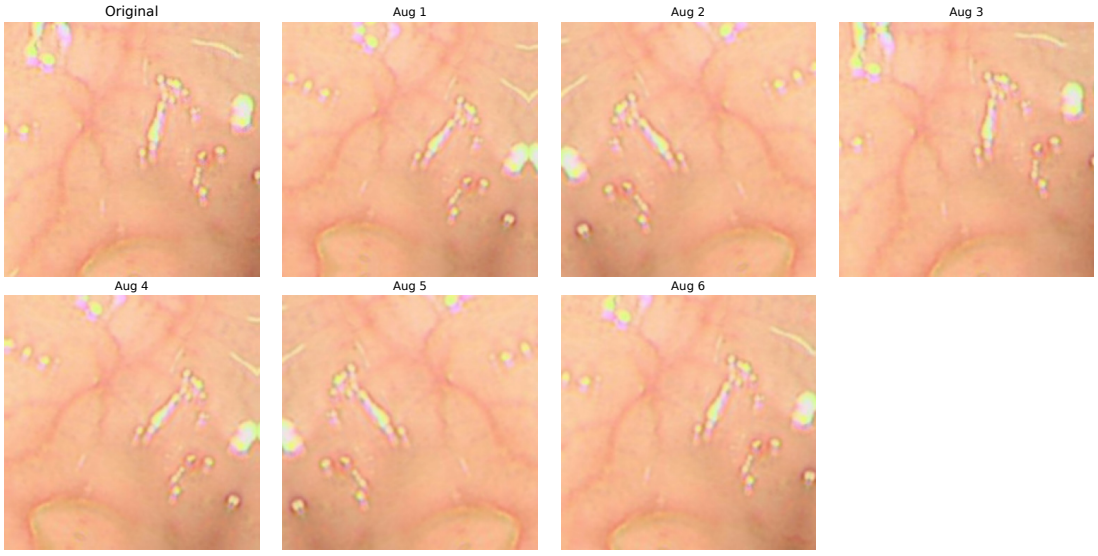


Figure 2. Example of on-the-fly data augmentation applied to a polyp image.

EfficientNetB0 employs a compound scaling strategy that uniformly balances network depth, width, and input resolution [28]. As a result, it provides strong accuracy with significantly fewer parameters and lower inference time compared to many conventional architectures, making it well-suited for deployment in real-time or resource-constrained clinical settings.

InceptionV3 uses factorized convolutions and inception modules that process input features at multiple scales simultaneously [29]. This design is particularly advantageous for polyp detection, where lesions may vary greatly in size, shape, and texture.

By combining these three backbones in an ensemble, our framework benefits from their complementary strengths: ResNet50's hierarchical feature extraction, EfficientNetB0's computational efficiency, and InceptionV3's multi-scale representation. This fusion, followed by Bayesian hyperparameter optimization, ensures reliable and well-balanced predictions while keeping the overall model lightweight.

3.3. Bayesian Hyperparameter Optimization

Choosing appropriate hyperparameters is a crucial step in developing deep learning models, as factors such as the number of neurons (u), dropout rate (p), and learning rate (η) have a direct impact on both model accuracy and generalization. Rather than relying on trial-and-error methods or exhaustive grid search, which are computationally expensive and time-consuming, we adopted Bayesian Optimization to systematically determine the best hyperparameter configuration [30, 31].

Bayesian Optimization builds a probabilistic surrogate model to approximate how hyperparameters affect model performance. After each training iteration, the surrogate model is updated with the new results, and an acquisition function is used to decide which hyperparameter values to evaluate next. This process strikes a balance between exploration—trying novel regions of the hyperparameter space that might lead to improvement—and exploitation—focusing on areas already known to yield strong performance.

This strategy significantly reduces unnecessary evaluations by avoiding regions that are unlikely to improve accuracy, leading to faster convergence toward optimal solutions. In our framework, Bayesian Optimization was used to search for the optimal number of units (u) in the dense layer (64–512), dropout rates (p) (0.1–0.5), and learning rates (η) (10^{-3} or 10^{-4}). The resulting hyperparameter configurations consistently improved classification

performance across folds while reducing overfitting, thereby making the approach both computationally efficient and reproducible compared to manual tuning or random search.

3.4. Ensemble Strategy

To improve classification robustness and reduce the risk of overfitting to a single model, we adopted a probability-based ensemble approach. For each backbone — ResNet50, EfficientNetB0, and InceptionV3 — we trained an independent model using the optimal hyperparameters obtained from Bayesian optimization. Each model outputs a probability score P for the positive (polyp) class. The final prediction is obtained by averaging the probabilities of all three models:

$$P_{\text{ensemble}} = \frac{P_{\text{ResNet}} + P_{\text{EffNet}} + P_{\text{Inception}}}{3}$$

where P_{ensemble} is the aggregated probability for a given input image. A classification threshold of 0.5 is then applied to determine the final label:

$$\hat{y} = \begin{cases} 1 & \text{if } P_{\text{ensemble}} \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

This probability-level fusion allows each model to contribute proportionally to the final decision, mitigating the impact of misclassifications from any single backbone. Compared to majority voting on class labels, probability averaging preserves confidence information and often yields smoother and more reliable predictions, which is critical for medical decision support systems.

3.5. Evaluation

To ensure unbiased performance assessment, we employed stratified 10-fold cross-validation, which preserves the class distribution across folds. In each fold, class weights were used to address the natural class imbalance between polyp and non-polyp samples. Training was further stabilized by early stopping and adaptive learning rate scheduling: if validation performance plateaued, the learning rate was automatically reduced, enabling more precise convergence.

The framework was evaluated using a comprehensive set of metrics: accuracy, precision, recall, F1-score, Cohen's Kappa, and the area under the receiver operating characteristic curve (ROC–AUC). Performance was reported as the mean $\pm$ standard deviation across folds. Additionally, confusion matrices and ROC curves were generated to provide a detailed view of classification behavior and model discriminative ability.

The key classification and agreement metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{6}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. p_o represents the observed agreement between predicted and ground truth labels, and p_e is the agreement expected by chance. Finally, the ROC–AUC metric measures the area under the ROC curve, which plots the true positive rate against the false positive rate at multiple thresholds. Higher AUC values indicate better class separability.

Algorithm 1: Ensemble Deep Transfer Learning with Bayesian Optimization and 10-Fold CV

Input: Dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, folds $K = 10$
Output: Aggregated metrics: Accuracy, Precision, Recall, F1, Confusion matrix, ROC–AUC
Preprocess: Resize all images to $224 \times 224 \times 3$. No augmentation is applied before splitting to ensure strict data separation.
Define backbones: ResNet50, EfficientNetB0, and InceptionV3 (ImageNet weights, remove top). For each backbone add: GAP $\rightarrow$ Dense(u) $\rightarrow$ Dropout(p) $\rightarrow$ Sigmoid.
Bayesian tuning: Use Bayesian optimization to search hyperparameters $\Theta = \{u, p, \eta\}$ (units, dropout, learning rate) for each backbone; objective: validation accuracy.
for $k = 1$ **to** K **do**
  Split D into train $D_{\text{train}}^{(k)}$ and test $D_{\text{test}}^{(k)}$ (stratified).
  Compute class weights on $D_{\text{train}}^{(k)}$.
  Apply on-the-fly data augmentation (*flip, rotation* $\pm 8°$, *zoom* $\leq 8\%$) **only** to $D_{\text{train}}^{(k)}$ during training. The validation fold $D_{\text{test}}^{(k)}$ remains pristine and unaugmented to prevent data leakage.
  For each backbone $b \in \{\text{ResNet}, \text{EffNet}, \text{Inc}\}$:
    - Initialize model with best hyperparameters Θ_b from Bayesian tuning.
    - Train on $D_{\text{train}}^{(k)}$ with early stopping.
    - Predict probabilities $P_b^{(k)}$ on $D_{\text{test}}^{(k)}$.

  Compute ensemble probability: $P_{\text{ens}}^{(k)} = \frac{1}{3} \sum_b P_b^{(k)}$.
  Threshold: $\hat{y}^{(k)} = \mathbf{1}\big[P_{\text{ens}}^{(k)} \geq 0.5\big]$.
  Store predictions and compute fold metrics (Acc, Prec, Rec, F1).
Aggregate results across folds: report mean $\pm$ std of metrics, confusion matrix, ROC curve, and AUC.

Algorithm 1 summarizes the entire pipeline: data preprocessing and augmentation, Bayesian hyperparameter tuning for each backbone, independent training, probability averaging for ensemble predictions, and final performance evaluation using stratified 10-fold cross-validation.

4. Results and Discussion

This section presents the experimental results obtained from the proposed ensemble framework and discusses their significance. We first describe the experimental setup, including implementation details and hyperparameter tuning results. Then, we report quantitative findings for the CP-CHILD-A, CP-CHILD-B, and Kvasir V2 datasets, supported by confusion matrices and ROC curves to illustrate classification performance. To further assess the contribution of each component within the ensemble, we present an ablation study analyzing the effect of individual classifiers and fusion strategies. We also benchmark our approach against recent state-of-the-art methods on the same datasets to confirm its competitiveness. Finally, we demonstrate a conceptual clinical deployment design, highlighting how the framework could be integrated into real-world diagnostic workflows.

4.1. Experimental Setup

To validate the proposed ensemble framework, all experiments were implemented in Python using the TensorFlow deep learning library and executed on Google Colab with GPU acceleration enabled. The development environment consisted of a PC running Windows 11, equipped with an Intel Core i7-10510U CPU (1.80 GHz) and 16 GB of RAM.

Hyperparameter tuning was performed independently for each backbone and dataset using Bayesian optimization with 5 trials per model. The search space covered the number of dense-layer units ($u \in [64, 512]$), dropout probability ($p \in [0.1, 0.5]$), and learning rate ($\eta \in \{1\mathrm{e}{-3}, 1\mathrm{e}{-4}\}$). Each model was fine-tuned for transfer learning using ImageNet-initialized weights and trained under stratified 10-fold cross-validation to ensure generalization.

Table 2 summarizes the optimal hyperparameters found for each backbone on all three datasets (CP-CHILD-A, CP-CHILD-B, and Kvasir V2). Notably, for the Kvasir V2 dataset, the optimal configuration favored a higher dropout rate in ResNet50 and moderate dense-layer sizes in EfficientNetB0 and InceptionV3, which aligns with the dataset's greater visual variability compared to CP-CHILD, indicating that stronger regularization improved generalization.

The reported hyperparameters correspond to the final models used in ensemble and ablation experiments, ensuring a fair comparison across architectures and datasets.

Table 2. Best hyperparameters found by Bayesian optimization for each backbone across datasets. "Units" refers to the number of neurons in the added dense layer; "Drop" is the dropout rate; "LR" is the learning rate.

Dataset	Backbone	Units (u)	Drop (p)	LR (η)
CP-CHILD-A	ResNet50	320	0.40	1e−4
	EfficientNetB0	64	0.50	1e−3
	InceptionV3	128	0.10	1e−3
CP-CHILD-B	ResNet50	384	0.10	1e−4
	EfficientNetB0	512	0.20	1e−3
	InceptionV3	256	0.10	1e−4
Kvasir v2	ResNet50	320	0.50	1e−4
	EfficientNetB0	256	0.10	1e−4
	InceptionV3	128	0.40	1e−3

Across datasets, the tuning process revealed consistent trends: EfficientNetB0 often required a lower dropout and moderate dense-layer size to balance its lightweight architecture, while ResNet50 benefited from higher dropout rates to prevent overfitting, especially on smaller datasets like CP-CHILD-B. InceptionV3 favored intermediate configurations with slightly higher learning rates, consistent with its faster convergence behavior. These tuned parameters were subsequently adopted in the ablation study and ensemble evaluation to ensure optimized and comparable performance across all datasets.

4.2. Quantitative Results on CP-CHILD-A

Table 3 reports the performance of the proposed ensemble framework on CP-CHILD-A across 10 folds of stratified cross-validation. The model achieves consistently high results, with a mean accuracy of $99.56\% \pm 0.25$, precision of $97.99\% \pm 1.52$, recall of $98.58\% \pm 0.89$, and F1-score of $98.27\% \pm 0.78$. These results indicate excellent stability, as reflected by the low standard deviations across folds. The ROC-AUC value of 99.89% confirms the strong discriminative power of the model.

Table 3. Performance metrics across 10-fold cross-validation for the proposed ensemble framework on CP-CHILD-A.

Fold	Accuracy	F1-score	Precision	Recall	Kappa	ROC-AUC
1	99.62	98.31	98.86	97.75	98.09	99.98
2	99.75	98.88	98.88	98.88	98.74	99.96
3	99.75	98.88	98.88	98.88	98.74	99.94
4	99.88	99.50	100.00	99.01	99.43	99.97
5	99.25	97.32	97.32	97.32	96.89	99.88
6	99.62	98.45	96.94	100.00	98.23	99.96
7	99.50	97.94	98.96	96.94	97.65	99.53
8	99.62	98.34	97.80	98.89	98.13	99.76
9	99.62	98.49	98.00	98.99	98.28	99.95
10	99.00	96.61	94.21	99.13	96.02	99.93
Mean	**99.56**	**98.27**	**97.99**	**98.58**	**98.02**	**99.89**
Std	**0.25**	**0.78**	**1.52**	**0.89**	**0.92**	**0.13**

Beyond overall accuracy, greater emphasis is placed on the **Recall (Sensitivity)** and **F1-score**, which are particularly meaningful for imbalanced medical datasets. The high recall of 98.58% confirms that the proposed ensemble correctly identifies nearly all true polyp cases, while the F1-score of 98.27% reflects a balanced trade-off between sensitivity and precision. Such a balance is critical in medical imaging, where missing a polyp (false negative) can have more serious consequences than a false alarm.

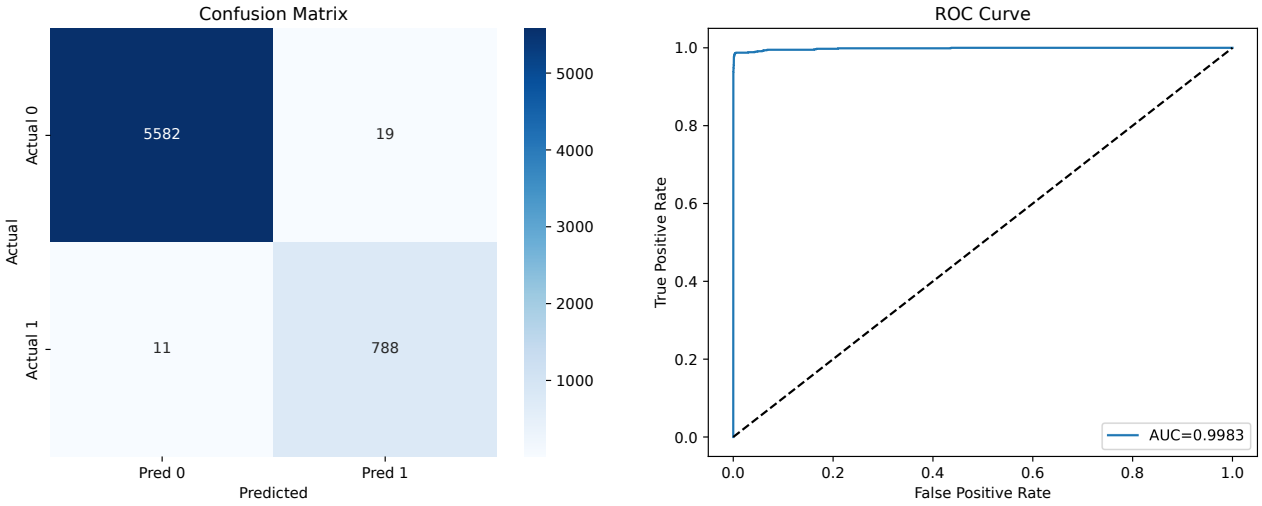
Figure 3 illustrates the model's classification behavior on CP-CHILD-A. The confusion matrix shows that the model correctly classified most samples, with only **11 false negatives (FN)** and **19 false positives (FP)** across the entire test set. This very low number of false negatives demonstrates that the ensemble rarely overlooks true polyp regions, which is essential for safe clinical deployment. Furthermore, the calculated **Specificity**—the proportion of correctly classified non-polyp samples—reached **99.66%**, highlighting the model's strong ability to correctly identify normal mucosa and avoid unnecessary alerts.

The ROC curve remains very close to the ideal top-left shape, confirming high separability between classes across all thresholds. Overall, the ensemble achieves an excellent balance between high sensitivity (minimizing missed polyps) and high specificity (reducing false alarms), demonstrating its robustness and suitability for practical colonoscopy screening.

4.3. Quantitative Results on CP-CHILD-B

Table 4 summarizes the performance of the proposed ensemble framework on CP-CHILD-B under 10-fold stratified cross-validation. The model maintained consistently high performance across folds, achieving a mean accuracy of $99.40\% \pm 0.70$, precision of $98.03\% \pm 2.62$, recall of $99.78\% \pm 0.65$, and F1-score of $98.88\% \pm 1.31$. These results confirm the robustness of the ensemble framework, with very low variability between folds. The mean Cohen's Kappa of 98.47% indicates excellent agreement beyond chance, while the ROC-AUC of 99.93% demonstrates the framework's near-perfect discrimination between polyp and non-polyp regions.

Given the class imbalance in CP-CHILD-B, **Recall (Sensitivity)** and **F1-score** provide a more clinically meaningful assessment of model performance than accuracy alone. The ensemble achieved an exceptionally high recall of 99.78%, indicating that almost every true polyp instance was correctly detected. The corresponding F1-score of 98.88% shows that this sensitivity was achieved without sacrificing precision, ensuring balanced detection. Such performance is crucial for clinical settings, where minimizing false negatives directly translates to improved diagnostic safety.



(a) Confusion matrix

(b) ROC curve

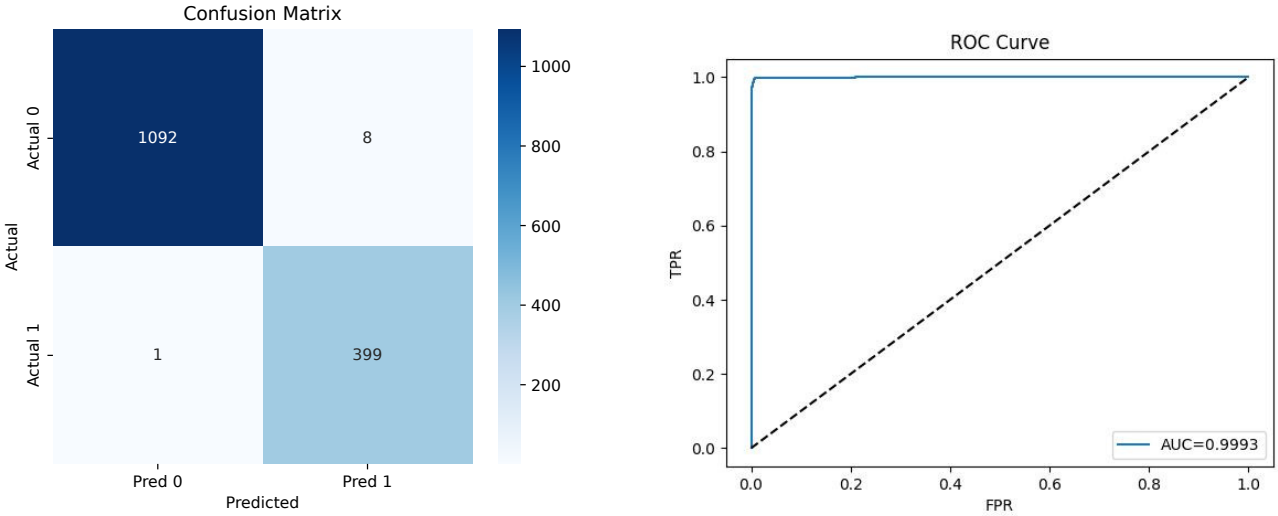
Figure 3. Confusion matrix and ROC curve for the proposed ensemble framework on CP-CHILD-A.

Table 4. Performance metrics across 10-fold cross-validation for the proposed ensemble framework on CP-CHILD-B.

Fold	Accuracy	F1-score	Precision	Recall	Kappa	ROC-AUC
1	100.00	100.00	100.00	100.00	100.00	100.00
2	98.00	96.55	93.33	100.00	95.15	100.00
3	99.33	98.82	97.67	100.00	98.36	100.00
4	98.67	97.06	94.29	100.00	96.20	99.92
5	100.00	100.00	100.00	100.00	100.00	100.00
6	100.00	100.00	100.00	100.00	100.00	100.00
7	100.00	100.00	100.00	100.00	100.00	100.00
8	100.00	100.00	100.00	100.00	100.00	100.00
9	98.67	97.44	95.00	100.00	96.54	99.93
10	99.33	98.90	100.00	97.83	98.42	99.44
Mean	**99.40**	**98.88**	**98.03**	**99.78**	**98.47**	**99.93**
Std	**0.70**	**1.31**	**2.62**	**0.65**	**1.78**	**0.17**

Figure 4 visualizes these findings. The confusion matrix reveals that the model produced only **1 false negatives (FN)** and **8 false positives (FP)** in total, underscoring its reliability in distinguishing true polyp regions from normal non-polyp regions. The computed **Specificity** reached **99.27%**, showing that the model not only detects nearly all true polyps but also correctly rejects the vast majority of non-polyp areas. This high specificity ensures reduced unnecessary alerts, which is particularly valuable in real-world colonoscopy assistance systems.

The ROC curve remains tightly aligned with the top-left boundary, confirming a near-ideal balance between sensitivity and specificity across varying thresholds. Overall, the ensemble framework demonstrated exceptional consistency and clinical relevance on CP-CHILD-B, maintaining excellent detection sensitivity while keeping false positives extremely low — a desirable property for computer-aided polyp screening.



(a) Confusion matrix

(b) ROC curve

Figure 4. Confusion matrix and ROC curve for the proposed ensemble framework on CP-CHILD-B.

4.4. Quantitative Results on Kvasir V2

The Kvasir v2 dataset used in this study is **balanced**, containing an equal number of polyp and non-polyp images (1000 each). This balanced setup provides a fair ground for evaluating the model's ability to generalize across both classes without the influence of class imbalance.

Table 5 presents the quantitative performance of the proposed ensemble framework across 10 folds of stratified cross-validation. Compared with the CP-CHILD datasets, the results on Kvasir v2 are relatively lower, which is expected given the higher variability and more diverse imaging conditions within this dataset. Nevertheless, the model maintains strong and stable performance, achieving a mean accuracy of $92.80\% \pm 1.14$, precision of $90.47\% \pm 2.77$, recall of $96.09\% \pm 1.14$, and F1-score of $93.17\% \pm 1.54$. The ROC-AUC of $97.51\% \pm 0.40$ demonstrates that the ensemble retains excellent discriminative ability, even under more challenging conditions.

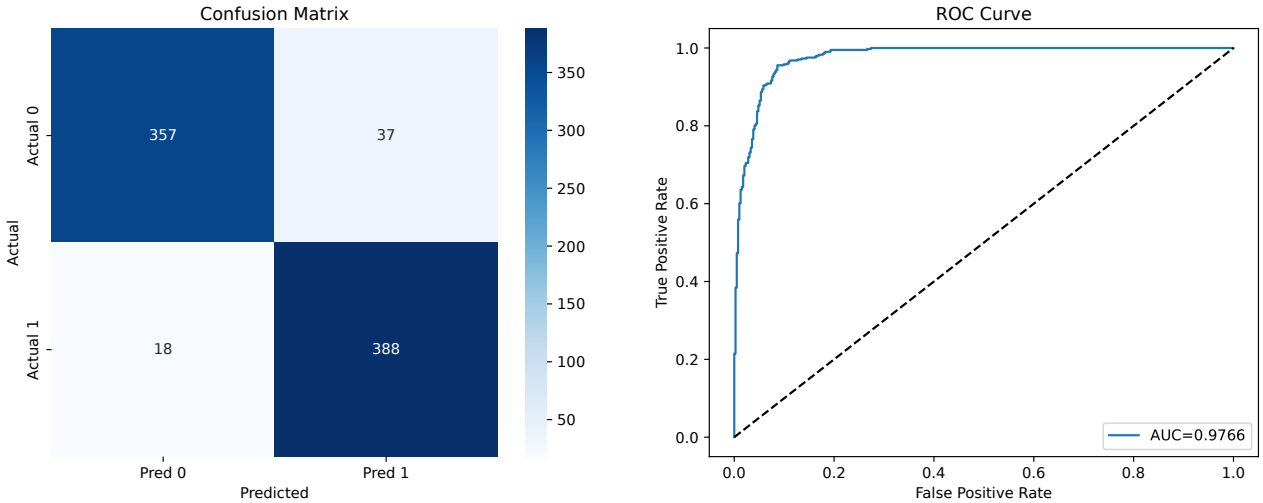
Table 5. Performance metrics across 10-fold cross-validation for the proposed ensemble framework on Kvasir v2.

Fold	Accuracy	F1-score	Precision	Recall	Kappa	ROC-AUC
1	93.50	94.37	92.37	96.46	86.69	97.36
2	93.00	94.07	90.24	98.23	85.57	97.39
3	93.00	94.07	90.24	98.23	85.57	97.46
4	92.50	93.09	90.99	95.28	84.90	97.93
5	91.00	90.62	86.14	95.60	82.02	97.76
6	92.50	92.39	90.10	94.79	85.01	96.74
7	94.50	95.15	94.74	95.58	88.80	97.78
8	94.50	94.84	94.39	95.28	88.95	98.24
9	91.00	90.62	86.14	95.60	82.02	97.27
10	92.50	92.46	89.32	95.83	85.02	97.22
Mean	**92.80**	**93.17**	**90.47**	**96.09**	**85.45**	**97.51**
Std	**1.14**	**1.54**	**2.77**	**1.14**	**2.22**	**0.40**

While the accuracy is slightly lower than that observed on the CP-CHILD datasets, the model still demonstrates **strong generalization ability** across different imaging domains. The high recall of 96.09% suggests that the model is capable of detecting the vast majority of true polyp cases, while the F1-score of 93.17% indicates a good balance between sensitivity and precision. This is particularly meaningful since real-world endoscopic images often exhibit variations in lighting, texture, and appearance, which can challenge even robust models.

Figure 5 illustrates the model's classification behavior on the Kvasir v2 dataset. The confusion matrix shows that most samples were correctly classified, with only **18 false negatives (FN)** and **37 false positives (FP)** observed across the test set. Given the balanced data distribution, these results confirm that the ensemble is effective in both detecting true polyps and avoiding excessive false alarms. The corresponding **Specificity** of approximately **90.59%** further highlights its reliability in identifying non-polyp areas.

The ROC curve, with an AUC of **0.9766**, remains close to the ideal top-left shape, indicating a clear separation between classes. Overall, although the performance on Kvasir v2 is slightly lower than on institution-specific datasets, the ensemble framework maintains **high sensitivity, stability, and robustness**, demonstrating promising potential for deployment across diverse endoscopic imaging settings.



(a) Confusion matrix

(b) ROC curve

Figure 5. Confusion matrix and ROC curve for the proposed ensemble framework on the Kvasir v2 dataset.

4.5. Ablation Study: Ensemble vs Individual Backbones

To evaluate whether the ensemble's added complexity is justified, we compared the final ensemble against each individually optimized backbone (ResNet50, EfficientNetB0, and InceptionV3). All models were trained under identical experimental conditions, including the same image preprocessing, data augmentation, optimizer, learning rate schedule, batch size, and 10-fold cross-validation protocol. This ensures that any observed differences in performance are attributable solely to the model architecture and the ensemble fusion strategy rather than to training configuration.

Table 6 presents the comparative performance between the individual backbones and the proposed ensemble framework across all datasets. Across the three datasets, the ensemble consistently achieves the highest mean values for accuracy, F1-score, precision, recall, and ROC-AUC, demonstrating the benefit of integrating complementary feature representations from multiple architectures. While individual models such as ResNet50 and EfficientNetB0 already deliver strong results, particularly on the CP-CHILD datasets, the ensemble provides an additional

Table 6. Ablation study comparing individual backbones with the proposed ensemble framework across all datasets. Reported values are mean ± std over 10 folds.

Dataset	Model	Acc(%)	Prec(%)	Rec(%)	F1(%)	AUC(%)
CP-CHILD-A	ResNet50	99.24±0.20	96.08±2.26	97.76±1.38	96.89±1.03	99.83±0.18
	EfficientNetB0	98.81±0.34	93.83±2.69	96.86±1.84	95.29±1.43	99.83±0.13
	InceptionV3	98.09±0.92	91.61±4.99	94.08±2.98	92.70±2.47	99.13±0.31
	Proposed	**99.56±0.25**	**97.99±1.52**	**98.58±0.89**	**98.27±0.78**	**99.89±0.13**
CP-CHILD-B	ResNet50	99.07±0.53	96.77±2.06	99.78±0.65	98.24±0.94	99.91±0.22
	EfficientNetB0	98.40±0.61	95.61±1.36	98.43±2.07	96.98±1.19	99.92±0.11
	InceptionV3	95.87±1.63	89.06±3.87	96.31±2.60	92.49±2.61	99.30±0.61
	Proposed	**99.40±0.70**	**98.03±2.62**	**99.78±0.65**	**98.88±1.31**	**99.93±0.17**
Kvasir V2	ResNet50	89.70±2.10	87.30±4.36	93.01±2.11	89.98±2.13	95.77±1.20
	EfficientNetB0	90.45±2.21	87.39±2.84	94.66±2.72	90.84±1.97	96.14±1.10
	InceptionV3	86.80±2.80	84.45±3.96	90.26±2.36	87.22±2.76	94.49±2.42
	Proposed	**92.80±1.14**	**90.47±2.77**	**96.09±1.14**	**93.17±1.54**	**97.51±0.40**

performance margin of approximately 0.3–1.5% in most metrics. This improvement, though modest in absolute terms, is consistent across all folds and datasets, indicating that the ensemble fusion enhances robustness rather than merely fitting a specific dataset.

For the CP-CHILD-A and CP-CHILD-B datasets, which contain more homogeneous imaging conditions, the ensemble slightly outperforms ResNet50 and EfficientNetB0, achieving peak accuracies of 99.56% and 99.40%, respectively. These results suggest that while the individual networks already capture discriminative texture and boundary details effectively, the ensemble leverages their complementary strengths to achieve more reliable predictions and lower variance across folds.

On the more diverse and challenging Kvasir v2 dataset, the performance gap becomes more evident. The ensemble achieves an accuracy of 92.80% and an F1-score of 93.17%, surpassing the best single backbone (EfficientNetB0) by around 2.3% and 2.3%, respectively. This confirms that the ensemble generalizes better to variations in illumination, color tone, and surface texture—conditions under which single backbones may exhibit inconsistent feature activation or misclassification of subtle polyp patterns. The consistently higher ROC-AUC of 97.51% further validates its improved discriminative capability.

It is worth noting that EfficientNetB0, despite being a lightweight model, performs competitively across all datasets, achieving near-ensemble performance in several metrics. This observation is particularly relevant for real-time or resource-limited environments, where the marginal gain of the ensemble may be traded off for faster inference and reduced computational cost. Nevertheless, for applications demanding the highest possible reliability, especially in clinical decision-support systems, the ensemble's superior balance between sensitivity and specificity justifies its additional complexity.

Overall, this ablation study clearly demonstrates that combining multiple optimized backbones leads to a measurable and consistent improvement in predictive stability and generalization. The results validate the ensemble's design rationale and highlight its practical advantage in handling dataset variability while maintaining strong performance.

4.6. Comparison with State-of-the-Art Methods

Table 7 presents a comprehensive comparison between the proposed ensemble framework and several state-of-the-art approaches on the CP-CHILD-A, CP-CHILD-B, and Kvasir v2 datasets. Our method consistently demonstrates competitive or superior performance across most evaluation metrics while maintaining reliable generalization through 10-fold cross-validation.

For CP-CHILD-A, the proposed ensemble achieves an accuracy of 99.56%, outperforming ResNet50 [7] by over 2% and slightly surpassing the ResNet152GAP model from Wei Wang et al. [22]. These results are obtained under a rigorous stratified 10-fold cross-validation setting, providing a more reliable and unbiased estimate of performance compared to prior studies that rely on single-train-test splits.

Similarly, on CP-CHILD-B, our model achieves 99.40% accuracy and the highest recall (99.78%), reflecting excellent sensitivity with minimal missed polyp cases. Although the DeepCPD model [8] reports a slightly higher accuracy, its lack of cross-validation makes direct comparison less conclusive. Overall, the ensemble achieves strong and stable results across both in-house datasets, supporting its robustness and reproducibility.

On the more challenging and diverse Kvasir v2 dataset, the ensemble maintains robust generalization, achieving an accuracy of 92.80% and F1-score of 93.17%. While the absolute accuracy is slightly lower than on the CP-CHILD datasets, this reflects the increased variability in imaging conditions, color tone, and surface textures typical of publicly sourced data. Notably, the recall of 96.09% highlights that the model continues to detect most true polyp cases even under more heterogeneous conditions. Compared to other methods such as the Spatial-Attention ConvMixer [38], Explainable Deep Learning framework [39], and Attention-Guided CNN [40], the proposed ensemble achieves comparable or higher sensitivity, demonstrating its balanced trade-off between recall and overall accuracy. These findings indicate that the ensemble design enhances feature diversity and robustness across different data domains while maintaining a clinically acceptable performance level.

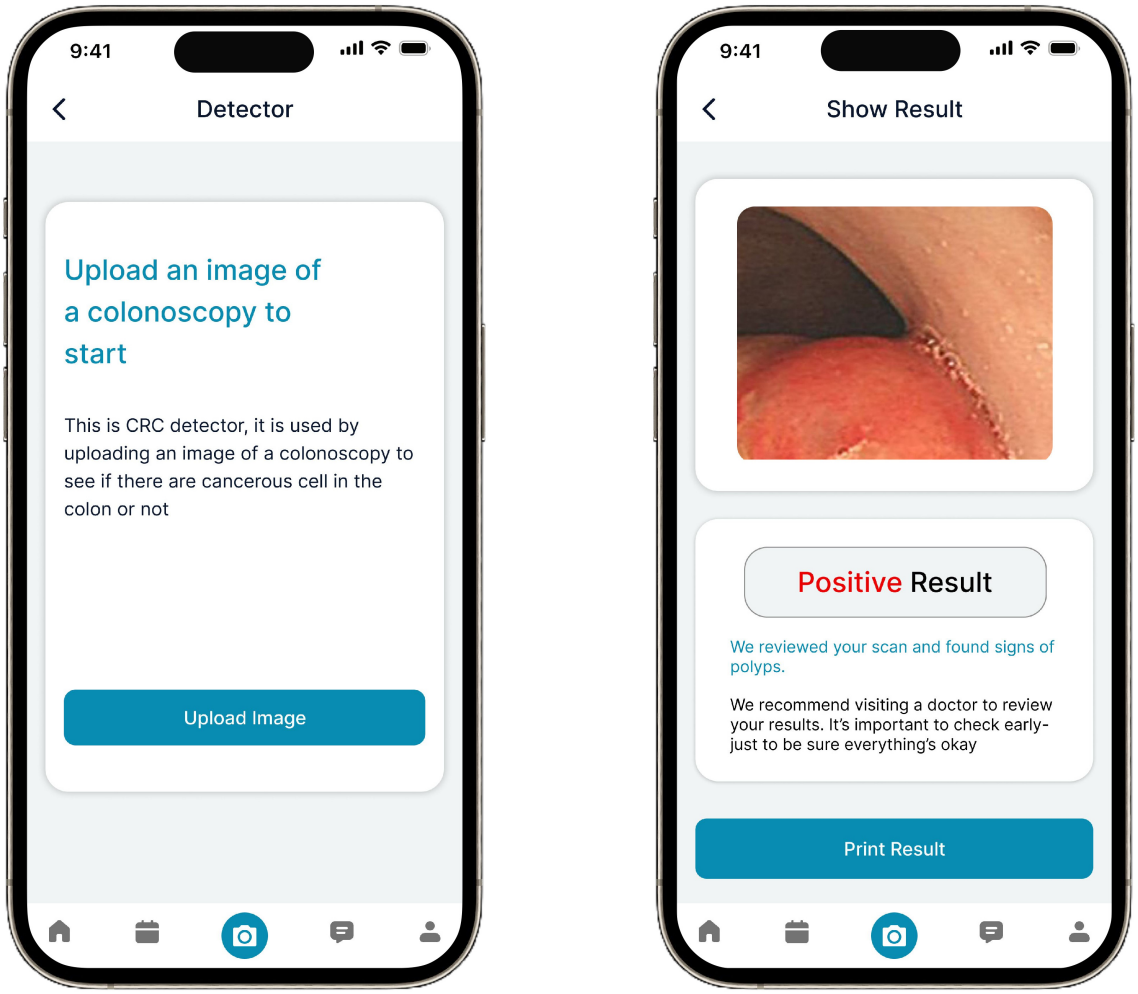
Table 7. Comparison of the proposed framework with prior work on CP-CHILD and Kvasir v2 datasets.

Dataset	Method	Architecture	Acc(%)	Rec(%)	Pre(%)	CV Applied?
CP-CHILD-A	**Proposed**	ResNet50+EffNetB0+IncV3 (Ensemble)	**99.56**	**98.58**	**97.99**	**Yes (10-fold)**
	Raseena et al. [21]	VGG19	99.10	98.00	97.51	No
	Raseena et al. [7]	ResNet50	97.20	95.00	91.34	No
	Wei Wang et al. [22]	ResNet152GAP	99.29	97.55	-	No
CP-CHILD-B	**Proposed**	ResNet50+EffNetB0+IncV3 (Ensemble)	**99.40**	**99.78**	98.03	**Yes (10-fold)**
	Raseena et al. [8]	ViT	99.75	99.00	100.0	No
	Raseena et al. [7]	ResNet50	98.75	99.00	96.11	No
	Wei Wang et al. [22]	ResNet152GAP	99.35	97.70	-	No
Kvasir v2	**Proposed**	ResNet50+EffNetB0+IncV3 (Ensemble)	**92.80**	**96.09**	**90.47**	**Yes (10-fold)**
	Demirbaş et al. [38]	Spatial-Attention ConvMixer	93.37	93.37	93.66	No
	Mukhtorov et al. [39]	ResNet152 + Grad–CAM	93.46	-	-	No
	Lonseko et al. [40]	Attention-Guided CNN	93.19	92.70	92.80	**Yes (5-fold)**

4.7. Clinical Deployment Design

To highlight how our framework could be used in real clinical practice, we created conceptual user interface (UI) mockups using Figma. These mockups demonstrate how the system might fit into the typical workflow of a clinician. Figure 6 presents two example screens. The first (a) is an image upload screen, where the clinician or technician can easily select colonoscopy *frames or still images* for analysis, either manually or by importing them directly from the endoscopy capture system. The second (b) is a results screen that displays the predicted risk probability together with a simple color-coded interpretation guide, allowing clinicians to quickly assess whether a polyp is likely present. Our goal with this design is to make results easy to interpret at a glance and to support rapid decision-making immediately after image acquisition. While these UIs are currently conceptual mockups, they are designed for future integration using RESTful APIs to communicate with the trained model hosted on a local inference device.

Beyond the user interface, Figure 7 illustrates a high-level architecture of the proposed clinical deployment pipeline. During colonoscopy, still frames are captured by the *endoscope* and exported to a nearby *edge inference device*. This device, equipped with a GPU or embedded accelerator, runs the optimized ensemble model and hosts the integrated UI. The clinician can upload or review images directly on this device, which performs classification and visualizes the predicted results on the connected medical monitor. The classified images and prediction scores can then be stored securely within the hospital network for reporting and follow-up analysis.



(a) (b)

Figure 6. UI prototype for image-based colorectal polyp classification using the proposed ensemble framework integrated on an edge inference device.

Deploying AI systems in clinical environments faces key challenges. Integration with existing endoscopy devices requires compatibility with medical imaging standards such as DICOM [32] for data exchange. All patient images must be processed locally to comply with data privacy laws, including HIPAA [33] in the United States and GDPR [34] in Europe. In addition, computer-aided detection (CADe) tools must undergo regulatory approval—such as FDA 510(k) clearance [35] or CE Mark certification under European MDCG guidance [36, 37]—to ensure clinical safety, reliability, and interoperability across medical environments.

As a key step toward clinical translation, future work will focus on prospective validation of the ensemble framework using still colonoscopy images acquired from real-world screening sessions. Collaborations with clinical partners will enable evaluation of classification accuracy, inference latency, and user experience across diverse imaging devices and patient populations. Feedback from gastroenterologists will guide interface refinement

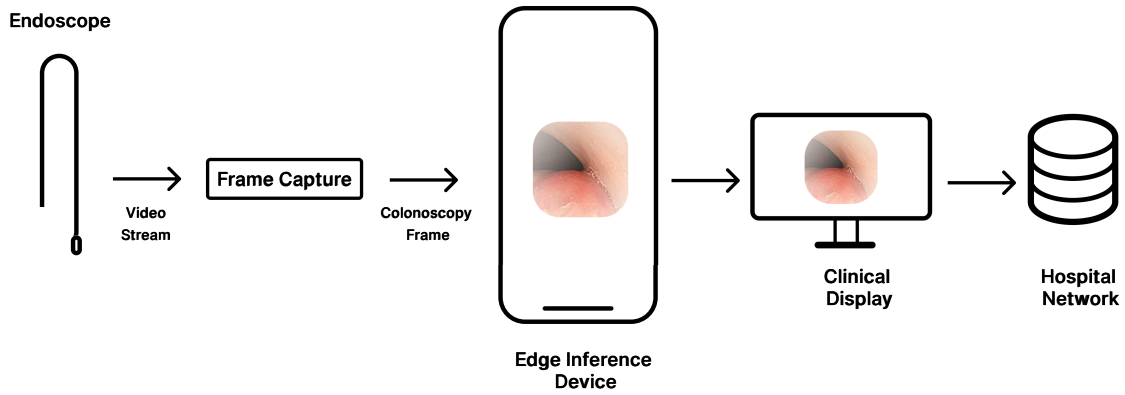


Figure 7. High-level architecture of the proposed clinical deployment pipeline.

and workflow optimization. Ultimately, multi-center studies will be pursued to demonstrate generalizability and to provide the clinical evidence required for regulatory approval and large-scale adoption in endoscopic practice.

5. Conclusion

This study presented an optimized deep ensemble framework for robust colorectal polyp detection, integrating three state-of-the-art CNN backbones—ResNet50, EfficientNetB0, and InceptionV3—optimized using Bayesian hyperparameter tuning. By combining model predictions through probability averaging and employing stratified 10-fold cross-validation, the proposed framework achieved high reliability and reduced the risk of overfitting.

Comprehensive experiments on three benchmark datasets—CP-CHILD-A, CP-CHILD-B, and Kvasir V2—demonstrated the framework's strong generalization ability across diverse imaging domains. The ensemble achieved accuracies of 99.56%, 99.40%, and 92.80%, respectively, with consistently high precision, recall (sensitivity), F1-scores, Cohen's Kappa, and ROC–AUC values. These results confirm the method's robustness and clinical relevance, especially in maintaining high sensitivity for accurate polyp detection.

An extensive ablation study further validated the ensemble's effectiveness by comparing it against individual backbone models. The results showed that the ensemble consistently outperformed single networks across all datasets, highlighting the benefit of combining complementary feature representations from multiple architectures to enhance predictive stability and generalization.

In addition, we benchmarked our approach against recent state-of-the-art methods on the same datasets. The proposed ensemble achieved competitive or superior performance, particularly under rigorous stratified cross-validation, offering a more reliable and unbiased estimate than single-split evaluations used in prior studies.

Beyond algorithmic performance, a conceptual clinical deployment design was introduced, illustrating how the framework can be integrated into real-world computer-aided diagnostic (CAD) workflows using mobile or edge-based devices equipped with GPUs or AI accelerators. This design bridges the gap between research and clinical application, supporting rapid and consistent decision-making during colonoscopy screenings.

Future work may explore partial fine-tuning of backbone networks to capture more task-specific features, integration with real-time colonoscopy video streams, and the development of lightweight models suitable for edge deployment. In addition, future work will incorporate a computational complexity analysis, including benchmarking inference time, parameter count, and FLOPs, to better quantify the trade-off between accuracy and efficiency. Exploring the integration of explainable AI techniques such as Grad-CAM or SHAP could also provide valuable insights into model decision-making, thereby increasing transparency and trust among medical practitioners. Overall, the proposed framework represents a promising step toward reliable, scalable, and clinically deployable colorectal polyp detection systems.

REFERENCES

1. H. Sung, J. Ferlay, R. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries*, CA: A Cancer Journal for Clinicians, vol. 71, Feb. 2021, doi: 10.3322/caac.21660.
2. J. Tan, P. Pickhardt, Y. Gao, Z. Liang, W. Cao, M. Pomeroy, Y. Huo, L. Li, M. Barish, and A. Abbasi, *3D-GLCM CNN: A 3-dimensional gray-level co-occurrence matrix based CNN model for polyp classification via CT colonography*, IEEE Transactions on Medical Imaging, vol. PP, pp. 1–1, Dec. 2019, doi: 10.1109/TMI.2019.2963177.
3. C. Gelu-Simeon, A. A. Bonnet, A. Bakhouche, and P. Han, *Deep Learning Model Applied to Real-Time Delineation of Colorectal Polyps*, BMC Medical Informatics and Decision Making, vol. 25, no. 1, pp. 1–14, 2025, doi: 10.1186/s12911-025-03047-y.
4. D. Babuc, T. Ivaşcu, M. Ardelean, and D. Onchiş, *Bionnica: A Customizable Deep Neural Network Architecture for Colorectal Polyps' Premalignancy Risk Evaluation with Masked Autoencoders*, Multimedia Tools and Applications, July 2025, doi: 10.1007/s11042-025-21058-9.
5. E. Aniq, F.-A. El Ghanaoui, and M. Chakraoui, *Vision Transformers for Breast Cancer Mammographic Image Classification*, Statistics Optimization & Information Computing, June 2025, doi: 10.19139/soic-2310-5070-2539.
6. M. Ahamed, M. Islam, M. Nahiduzzaman, M. Karim, M. Ayari, and A. Khandakar, *Automated Detection of Colorectal Polyp Utilizing Deep Learning Methods With Explainable AI*, IEEE Access, vol. PP, pp. 1–1, Jan. 2024, doi: 10.1109/ACCESS.2024.3402818.
7. T. Raseena, J. Kumar, and S. Balasundaram, *A Residual Learning Approach Towards the Diagnosis of Colorectal Disease Effectively*, in Lecture Notes in Computer Science, pp. 160–172, Jan. 2024, doi: 10.1007/978-3-031-53085-2_14.
8. T. P. Raseena, J. Kumar, and S. Balasundaram, *DeepCPD: deep learning with vision transformer for colorectal polyp detection*, Multimedia Tools and Applications, pp. 1–24, Feb. 2024, doi: 10.1007/s11042-024-18607-z.
9. J. Lewis, Y. Cha, and J. Kim, *Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images*, Scientific Reports, vol. 13, Jan. 2023, doi: 10.1038/s41598-023-28530-2.
10. D. Mukhtorov, R. Madinakhon, S. Muksimova, and Y.-I. Cho, *Endoscopic image classification based on explainable deep learning*, Sensors, vol. 23, p. 3176, Mar. 2023, doi: 10.3390/s23063176.
11. X. Zhang, F. Chen, T. Yu, J. An, Z. Huang, J. Liu, W. Hu, L. Wang, H. Duan, and J. Si, *Real-time gastric polyp detection using convolutional neural networks*, PLoS One, vol. 14, p. e0214133, Mar. 2019, doi: 10.1371/journal.pone.0214133.
12. M. Auzine, M. H.-M. Khan, S. Baichoo, N. Gooda Sahib-Kaudeer, P. Bissoonauth-Daiboo, X. Gao, and Z. Heetun, *Development of an ensemble CNN model with explainable AI for the classification of gastrointestinal cancer*, PLOS ONE, vol. 19, Jun. 2024, doi: 10.1371/journal.pone.0305628.
13. Z. Lonseko, P. Adjei, W. Du, C. Luo, D. Hu, L. Zhu, T. Gan, and N. Rao, *Gastrointestinal Disease Classification in Endoscopic Images Using Attention-Guided Convolutional Neural Networks*, Applied Sciences, vol. 11, p. 12, Nov. 2021, doi: 10.3390/app112311136.
14. P. Saxena, S. Huque, S. Vhatkar, K. Ramana, C. Anand, and D. Durai, *Genetic Algorithm Optimization of Feature Selection for Medical Image Classification*, International Journal of Soft Computing, vol. 14, pp. 3354–3360, May 2024, doi: 10.21917/ijsc.2024.0471.
15. M. Deif, H. Attar, A. Amer, H. Issa, M. Khosravi, and A. Solyman, *A New Feature Selection Method Based on Hybrid Approach for Colorectal Cancer Histology Classification*, Wireless Communications and Mobile Computing, vol. 2022, May 2022, doi: 10.1155/2022/7614264.
16. W. Nagy and H. Alsalamah, *Efficient Harris Hawk Optimization (HHO)-Based Framework for Accurate Skin Cancer Prediction*, Mathematics, vol. 11, p. 3601, Aug. 2023, doi: 10.3390/math11163601.
17. M. E. A. Elaziz, A. Dahou, N. Alsaleh, A. Elsheikh, A. Saba, and M. Ahmadein, *Boosting COVID-19 Image Classification Using MobileNetV3 and Aquila Optimizer Algorithm*, Entropy, vol. 23, Oct. 2021, doi: 10.3390/e23111383.
18. A. Mohamed, A. Saba, M. Hassan, H. M. Youssef, A. Dahou, A. Elsheikh, A. El-Bary, M. E. A. Elaziz, and R. Ibrahim, *Boosted Nutcracker Optimizer and Chaos Game Optimization with Cross Vision Transformer for Medical Image Classification*, Egyptian Informatics Journal, vol. 26, p. 100457, Jun. 2024, doi: 10.1016/j.eij.2024.100457.
19. CP Child, *Dataset available at* https://figshare.com/articles/dataset/CP-CHILD_zip/12554042?file=23383508, accessed Apr. 21, 2025.
20. K. Malialis, D. Papatheodoulou, S. Filippou, C. Panayiotou, and M. Polycarpou, *Data augmentation on-the-fly and active learning in data stream classification*, arXiv preprint arXiv:2210.06873, Oct. 2022, doi: 10.48550/arXiv.2210.06873.
21. T. Raseena, Jitendra Kumar, and S. Balasundaram, *Exploring the Effectiveness of Deep Learning Architectures for Colorectal Polyp Detection: Performance Analysis and Insights*, SN Computer Science, vol. 5, Apr. 2024, doi: 10.1007/s42979-024-02825-1.

22. W. Wang, J. Tian, C. Zhang, Y. Luo, X. Wang, and J. Li, *An improved deep learning approach and its applications on colonic polyp images detection*, BMC Medical Imaging, vol. 20, Jul. 2020, doi: 10.1186/s12880-020-00482-3.
23. ImageNet Winning CNN Architectures (ILSVRC), *Kaggle Discussion*, Apr. 2025. [Online]. Available: https://www.kaggle.com/discussions/getting-started/149448. [Accessed: Apr. 15, 2025].
24. H. Shin, H. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. Summers, *Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning*, IEEE Transactions on Medical Imaging, vol. 35, Feb. 2016, doi: 10.1109/TMI.2016.2528162.
25. M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, *Transfusion: Understanding Transfer Learning with Applications to Medical Imaging*, arXiv preprint arXiv:1902.07208, Feb. 2019, doi: 10.48550/arXiv.1902.07208.
26. T. G. Dietterich, *Ensemble methods in machine learning*, in Multiple Classifier Systems: First International Workshop, MCS 2000, Lecture Notes in Computer Science, pp. 1–15, Jan. 2000. ISBN: 3-540-67704-6.
27. K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, Jun. 2016, doi: 10.1109/CVPR.2016.90.
28. M. Tan and Q. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, arXiv preprint arXiv:1905.11946, May 2019, doi: 10.48550/arXiv.1905.11946.
29. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, *Rethinking the Inception Architecture for Computer Vision*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, doi: 10.1109/CVPR.2016.308.
30. J. Snoek, H. Larochelle, and R. P. Adams, *Practical Bayesian Optimization of Machine Learning Algorithms*, Advances in Neural Information Processing Systems, vol. 25, 2012.
31. R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon, *Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020*, NeurIPS 2020 Competition and Demonstration Track, PMLR, pp. 3–26, 2021.
32. National Electrical Manufacturers Association (NEMA), *Digital Imaging and Communications in Medicine (DICOM) Standard*, Rosslyn, VA, USA, 2020. [Online]. Available: https://www.dicomstandard.org/
33. U.S. Department of Health and Human Services, *Health Insurance Portability and Accountability Act (HIPAA)*, Washington, DC, USA, 1996. [Online]. Available: https://www.hhs.gov/hipaa/index.html
34. European Parliament and Council of the European Union, *General Data Protection Regulation (GDPR)*, Official Journal of the European Union, Regulation (EU) 2016/679, Apr. 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj
35. U.S. Food and Drug Administration (FDA), *Artificial Intelligence and Machine Learning in Software as a Medical Device: Action Plan*, Silver Spring, MD, USA, 2022. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device
36. European Commission, Medical Device Coordination Group (MDCG), *FAQ on Interplay between the Medical Devices Regulation (MDR), the In Vitro Diagnostic Regulation (IVDR), and the AI Act — MDCG 2025-6*, Brussels, Belgium, June 2025. [Online]. Available: https://health.ec.europa.eu/latest-updates/mdcg-2025-6-faq-interplay-between-medical-devices-regulation-vitro-diagnostic-medical-devices-2025-06-19_en
37. European Commission, Medical Device Coordination Group (MDCG), *Guidance on Standardization for Medical Devices — MDCG 2021-5 Rev.1*, Brussels, Belgium, July 2024. [Online]. Available: https://health.ec.europa.eu/latest-updates/update-mdcg-2021-5-rev1-guidance-standardisation-medical-devices-july-2024-2024-07-02_en
38. A. Demirbaş, H. Üzen, and H. Fırat, *Spatial-attention ConvMixer architecture for classification and detection of gastrointestinal diseases using the Kvasir dataset*, Health Information Science and Systems, vol. 12, Apr. 2024, doi: 10.1007/s13755-024-00290-x.
39. D. Mukhtorov, R. Madinakhon, S. Muksimova, and Y.-I. Cho, *Endoscopic Image Classification Based on Explainable Deep Learning*, Sensors, vol. 23, p. 3176, Mar. 2023, doi: 10.3390/s23063176.
40. Z. Lonseko, P. Adjei, W. Du, C. Luo, D. Hu, L. Zhu, T. Gan, and N. Rao, *Gastrointestinal Disease Classification in Endoscopic Images Using Attention-Guided Convolutional Neural Networks*, Applied Sciences, vol. 11, p. 11136, Nov. 2021, doi: 10.3390/app112311136.
41. K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, *KVASIR: A Multi-Class Image Dataset for Computer-Aided Gastrointestinal Disease Detection*, in *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*, Taipei, Taiwan, 2017, pp. 164–169, doi: 10.1145/3083187.3083212.