



PolyDSS: computer-aided decision support system for multiclass polyp segmentation and classification using deep learning

Abdelrahman I. Saad¹ · Fahima A. Maghraby¹ · Osama M. Badawy²

Received: 15 July 2023 / Accepted: 7 December 2023 / Published online: 24 December 2023
© The Author(s) 2023

Abstract

Colorectal cancer (CRC) is a malignant condition that affects the colon or rectum, and it is distinguished by abnormal cell growth in these areas. Colon polyps, which are abnormalities, can turn into cancer. To stop the spread of cancer, early polyp detection is essential. The timely removal of polyps without submitting a sample for histology is made possible by computer-assisted polyp classification. In addition to Locally Shared Features (LSF) and ensemble learning majority voting, this paper introduces a computer-aided decision support system named PolyDSS to assist endoscopists in segmenting and classifying various polyp classes using deep learning models like ResUNet and ResUNet++ and transfer learning models like EfficientNet. The PICCOLO dataset is used to train and test the PolyDSS model. To address the issue of class imbalance, data augmentation techniques were used on the dataset. To investigate the impact of each technique on the model, extensive experiments were conducted. While the classification module achieved the highest accuracy of 0.9425 by utilizing the strength of ensemble learning using majority voting, the proposed segmenting module achieved the highest Dice Similarity Coefficient (DSC) of 0.9244 using ResUNet++ and LSF. In conjunction with the Paris classification system, the PolyDSS model, with its significant results, can assist clinicians in identifying polyps early and choosing the best approach to treatment.

Keywords Colorectal cancer · Paris classification · Transfer learning · Ensemble learning · Deep learning · Locally shared features

1 Introduction

Colorectal cancer is a prevalent type of neoplasm worldwide and is adapted to either the rectal region or colon. This type of tumor causes critical illness, contributing significantly to fatality rates concerning carcinomas

globally. The second-most likely cancer that leads to death in the US is colorectal cancer. By 2023, approximately 153,020 people are expected to have CRC, and of those, 52,550 will pass away from this disease, with 19,550 cases and 3750 deaths occurring among people under the age of 50 [1, 2]. Therefore, understanding this disease through research gains importance due to its common occurrence tendency, prevention through early diagnosis potential and need for efficacious curative modalities available. Polyps are abnormal formations that form within the wall of one's colon or rectum. Often beginning as small polyps with no cancer indications (known as adenomatous polyps) [3, 4], these developments can become life-threatening if ignored and progress to full-blown colorectal cancer. The size, quantity, and type of affected growths all have an impact on polyp-to-cancer transformation. Certain types of adenomas, in particular, have a higher proclivity to progress to malignancy than other types. Regular screenings, such as colonoscopies, are routine practices that ensure timely

✉ Abdelrahman I. Saad
abdelrahman.saad@aast.edu

Fahima A. Maghraby
fahima@aast.edu

Osama M. Badawy
obadawy@aast.edu

¹ College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport, Cairo, Egypt

² College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt

identification and removal of these problematic adenomatous growths, lowering the chances of developing colorectal cancer. Medical decision support systems have a significant impact on the treatment of colorectal cancer by providing timely, evidence-based recommendations and information to healthcare professionals [5, 6]. These systems use cutting-edge computer models and algorithms to analyze patient data. These systems' capacity to enhance clinical decisions, lessen diagnostic blunders, and enhance healthcare outcomes is a key characteristic. Medical decision support systems have the potential to completely transform colon cancer care by enabling patients to receive the most appropriate and effective treatments based on their own specific demands. These systems have the capacity to analyze complex data and provide real-time line guidance. A significant role is played by computer-aided diagnosis (CADx), a medical imaging and diagnosis technology, particularly in colonoscopy [7]. Using machine learning and deep learning methods, CADx helps medical professionals analyze anomalies or possible diseases in images such as colonoscopy images [8]. Through the investigation of numerous datasets and complex patterns, CADx systems can provide valuable insights while also improving diagnosis accuracy and efficiency [9]. CADx can aid in the detection of polyps, lesions, or other abnormalities within the colon during a colonoscopy, assisting clinicians in identifying potential precursors to colorectal cancer. Transfer learning models have evolved to be extremely efficient tools in all areas of machine learning, and their applicability in the classification of polyps is no exception. Identification and characterization of abnormal gastrointestinal growths are critical tasks for polyp classification, which is an important imaging task. Models developed using massive datasets like ImageNet can be fine-tuned on smaller, domain-specific datasets containing polyp images using transfer learning, which contributes to overcoming the limitations of large medical datasets because the cost of acquiring data and annotation is high. Transfer learning is used when there is insufficient training data for a model. On many different samples, deep neural networks are trained for new tasks using transfer learning, and the weights are inherited. According to [10], the Paris classification is crucial for the precise and consistent classification of polyps. It gives doctors access to superb information that they can use to decide on the best course of treatment. The Paris classification helps surgeons identify potential risks and predict the likelihood of cancer by categorizing polyps based on their external appearance and features, such as size, shape, or surface features. In order to facilitate collaboration among healthcare professionals and allow for consistency in reporting, the classification system offers a common language for doctors and scientists. A standardized classification allows physicians

to more precisely identify the best management strategies, such as surveillance intervals, endoscopic resection procedures, or surgical intervention referrals. Our goal is to create a computer-aided decision support system (CADSS) that will classify polyps into six categories based on the Paris classification: 0-Ip, 0-Ips, 0-Is, 0-IIa, 0-IIa/c, and 0-IIb. Correct classification of these categories, assessment of the main decisions about removing or resecting polyps, and identification of removal techniques if a dangerous polyp is present will help colonoscopists in their treatment. Two modules are proposed in this study. The first module is a segmentation module that improves image features by combining Residual U-Net (ResUNet) and Residual U-Net Plus Plus (ResUNet++) with Locally Shared Features (LSF). The model segments the images, and the resulting masks are used to draw an outline around the polyp-containing area. The first module's output is then used to classify polyps using a transfer learning model called EfficientNet. Data augmentation techniques such as flipping, random rotation, scaling, contrast modification, and zooming are employed to compensate for the imbalance of existing classes. Independently, five variations of EfficientNet models are trained, and a majority voting method is used to combine predictions from various models to make predictions more accurately. Due to its unique combination of U-Net and ResNet characteristics, ResUNet++ is a perfect choice for medical image segmentation. Its capacity to capture fine-grained information using skip connections, as well as its multi-scale context integration via densely coupled skip pathways, makes it extremely proficient at reliably segmenting objects of varied sizes and shapes in complex medical images. Furthermore, its proven record of delivering state-of-the-art performance in segmentation tasks, as well as its fine-tuning competences for coping with specific medical datasets, demonstrates its applicability for the challenging field of medical imaging. EfficientNet provides valuable benefits for image classification in the medical area, stressing both efficiency and speed, which are critical for quick and accurate diagnoses. Its scalability is ensured by the compound scaling method, which ensures that model size corresponds to computational resources and dataset complexity. Furthermore, the model makes use of transfer learning by expanding on its highly effective feature extraction skills from ImageNet pretraining, allowing for rapid adaptation to the requirements of medical image classification. Furthermore, the robust generalization of EfficientNet to numerous medical image different shapes, anatomical structures, and diseases makes it a good solution for the diverse and complicated nature of medical image analysis tasks. The main contributions of this paper are:

1. A novel two-module medical decision support system for segmenting and classifying polyps.
2. Addressing the problem of image shortages using image augmentation and transfer learning models.
3. Using Paris classification as a reference, we classify polyps into six classes: 0-Ip, 0-Ips, 0-Is, 0-IIa, 0-IIa/c, and 0-IIb.
4. Providing a concise summary of the nature of the polyps based on endoscopic appearance, class, number of images in each class, description, view, risk assessment, and proper resection technique.
5. Five EfficientNet model variations were trained, tested, and evaluated in order to find the most efficient model for classifying polyps.
6. Employing the ensemble learning technique with majority voting to enhance the accuracy and the model's robustness.

1.1 Paper organization

The structure of the paper is organized as follows: Section 2 introduces medical image processing approaches and the Paris classification system. Section 3 introduces the literature survey of prior studies. Section 4 presents the dataset, applied methods and proposed model in detail. Section 5 discusses implementation details, model evaluation methodology, evaluation metrics, experiments and discussion, comparison with existing methods, and finally a detailed analysis of the ablation studies. The graphical user interface (GUI) and its evaluation are viewed in Sect. 6. Section 7 presents the limitations and expected future work for this study. All ethical considerations are discussed in Sect. 8. Finally, in Sect. 9, the study conclusion is presented.

2 Background

This section provides an overview of the key components of our research. Fundamentals of image segmentation and classification concepts are presented. The section also delves into the Paris Classification System, which acts as the framework for colorectal polyp management and treatment. These key components are critical for understanding the novel approaches and findings provided in this research.

2.1 Medical image processing

Image segmentation and classification are vital methods in medical imaging, especially when addressing colorectal polyps. Image segmentation is the accurate demarcation

and identification of regions or objects within an image, which allows the polyps to be separated from the tissue around them. This procedure assists in the localization and quantification of polyps, which is necessary for appropriate diagnosis and treatment planning. The U-Net architecture is a well-known option for medical image segmentation. It is a convolutional neural network (CNN) [11] created for biomedical image processing. It comprises a contracting path for capturing context and a broadening symmetric path for exact localization. Image classification, on the other hand, seeks to classify these segmented regions, deciding whether the discovered polyps are benign, cancerous, or belong to a certain type or class. Deep learning models such as AlexNet, which learns complex image features using a deep convolutional layer [12], and Visual Geometry Group Net (VGG), which consists of small-sized convolutional filters that are easy to adapt and implement for different image classification tasks, while DenseNet aids in feature reuse and improves gradient flow for precise feature extraction. The previously described models can aid in classification by assessing the shape, texture, and other characteristics of the segmented polyps, allowing for early detection and accurate risk estimation. Together, these methods serve a critical role in the advancement of colorectal polyp identification and therapy, eventually leading to improved patient experiences and healthcare effectiveness.

2.2 Paris classification

In the field of gastrointestinal disorders, the Paris classification is a crucial approach, especially for the detection and management of colorectal polyps. By establishing a standard method of categorizing polyps based on traits like morphology, size, and other traits, it enables doctors to choose effective treatments. There are two methods for removing polyps: endoscopic mucosal resection (EMR) and endoscopic submucosal dissection (ESD) [13, 14] as

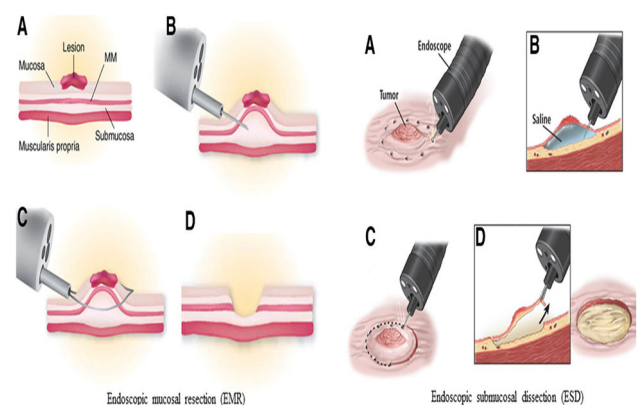


Fig. 1 Operation procedures for EMR compared to ESD [18]

shown in Fig. 1. EMR includes preparation in which the patient is typically sedated or anesthetized. The endoscope is inserted into the target area of the digestive tract. Following that, an injection of a solution, usually a saline solution mixed with medication, into the submucosal layer beneath the lesion is performed. Finally, a wire loop, or snare, is inserted through the endoscope and placed around the lesion. The loop is then tightened, allowing the lesion to be cut and removed from the underlying tissue. In ESD, the patient is prepared and sedated or anesthetized, as with EMR. The endoscope is inserted into the digestive tract and placed near the lesion of interest. The lesion's border is marked with a needle or other tools to create a demarcation line to guide the subsequent dissection, and solution is injected below the lesion to create a fluid padding and lift the lesion, allowing for a clearer working space for dissection. Finally, the lesion is carefully separated from the underlying tissue by carefully dissecting the submucosal layer. After making an incision along the marked border, the submucosal layer is exposed. By evaluating the Paris classification, physicians can estimate the risks of malignancy associated with polyps and decide which approach is appropriate. The classification system is intended to improve treatment outcomes by assisting healthcare professionals in selecting the best method for polyp removal, promoting effective management of colon tumors, and reducing the need for invasive procedures. We created a table that contains both PICCOLO dataset details based on Paris classification as well as the potential risk and resection technique for each polyp class, as shown in Table 1 [15–17].

3 Literature survey

This section delves into a thorough analysis of previous studies, an essential journey through the process of research that has formed the foundations for the current

understanding of the problem at hand. This investigation not only gives important background information, but it also acts as a stepping stone for examining the advancements and gaps in the current body of knowledge.

In 2021, Hsu et al. [19] extracted features using a CNN model that included convolution, batch normalization, implementing ReLU functions and performing max pooling operations. They came up with an accurate, mobile-friendly computer-aided diagnostic system. The authors used colonoscopy images provided by the CVC Clinic [20] and Chang Gung Medical Hospital's Department of Gastroenterology and Hepatology. The CVC Clinic dataset includes 612 continuous images derived from 29 white-light (WL) images. Using the YOLOv2 and YOLOv3 models, the authors' study achieved an overall accuracy of 94.9% and 94.4% for polyp detection, respectively. In terms of recognizing polyps, the findings showed that narrow-band imaging (NBI) images were more precise than WL images, with an accuracy of 82.8% for NBI and 72.2% for WL.

In 2022, Lo et al. [21] used various approaches to contrast the performance variations between deep learning with deep convolutional neural networks (DCNN) features and machine learning with texture features, the performance variations between different DCNN architectures, and the performance variations between DCNN models trained from scratch and transfer learning, respectively. They also used gray-level co-occurrence matrix (GLCM) and gabor textures to extract features, which they then combined in various classifiers to create polyp classification models. In addition, to manage the many features, they used principal component analysis as the feature selection method. The authors analyzed colonoscopy images from 1991 patients who had colonoscopies between January 1, 2018 and July 27, 2018. Among these patients, biopsy results showed that 206 adenocarcinomas, 732 adenomas, and 1053 hyperplastic polyps were present. The authors discovered that DCNN can achieve significant performance

Table 1 PICCOLO dataset description based on Paris classification in addition to the risk and resection technique

Paris classification polyp classes					
Endoscopic appearance	Class	Description	No. of images	Risk	Resection technique
Protruded lesions	0-Ip	Protruded, pedunculated	384	Low-moderate	EMR
	0-Ips	Sub pedunculated	300	Low-moderate	EMR
	0-Is	Protruded, sessile	640	Low-moderate	EMR or ESD
Flat elevated lesion	0-IIa	Slightly elevated (flat)	1364	High	EMR or ESD
	0-IIa/c	Flat elevation with central depression (flat)	174	High	ESD
Flat lesion	0-IIb	Flat mucosal change (completely flat)	106	High	ESD
N/a	N/a	N/a	465	–	–

and that, in the case of good-quality image data, training and building a model from nothing is an appealing approach. They also discovered that AlexNet trained from scratch had the highest accuracy of 96.4%, while GLCM texture features in the B channel had 75.6% accuracy. One of the study's limitations is that training from scratch may take longer in clinical use than transfer learning, although there is a significant accuracy difference. For future work, they suggested that the CAD system can assist gastroenterologists in determining the types of polyps as part of colonoscopy and that additional CAD systems for gastrointestinal inflammatory conditions like crohn's syndrome and colitis with ulceration may be evolved.

In 2023, Krenzer et al. [22] developed several novel methods for automated polyp classification, including transfer learning with various classification strategies (1-nn, centroid, support vector machine (SVM)), supervised pretraining, and transferring styles as a step in augmentation. Convolutional Neural Networks (CNNs), one type of deep learning technique, were also used to identify and categorize polyps. They also introduced a two-step process for classifying polyps based on the Paris classification scheme, which involves locating and cropping the polyp on the image, followed by classifying the polyp using a transformer on the cropped area. They also created a strong and an effective system for classifying polyps using the NICE criteria that can classify polyps in a precise and consistent manner. The study describes two datasets that were used in the analysis. The SUN database [23], which is a colonoscopy videos from an openly available database served as the initial dataset. The dataset contains 100 colonoscopy videos, each lasting about 10 minutes. The dataset contains 48 polyps that have been classified using the Paris classification system. The EndoData dataset, created by the document's authors at the University Clinic of Würzburg [24], is the second dataset used. Colonoscopy videos were annotated using a framework for faster endoscopic annotation in the dataset. The dataset contains 100 cases with a total of 23,154 polyp images. The polyps are classified according to the NICE classification system and are divided into four types: Is, Ip, Isp, and Ila. With an accuracy of 89.35%, the Paris classification system outperformed all previous papers in the literature in terms of state-of-the-art performance on clinical data. The NICE classification system proved the practicality of the few-shot learning approach in unavailable data environments in the endoscopic domain with a competing accuracy of 81.34%. The lack of readily available data is one of the study's major obstacles, particularly for uncommon conditions and incidents, as well as the cost and time required to obtain expert annotations. Another challenge is selecting a transfer learning dataset that aligns with the target domain's similarity notions. The authors also pointed out that the

accuracy of the models is restricted by the shortage of labeled datasets for the polyp area of study. Furthermore, the image distribution on the test datasets was unbalanced, which may reduce the significance of the test results. The authors proposed that random image flipping appears to be important in polyp characterization and should be investigated further in future research.

In 2023, Yue et al. [25] proposed a CAD method for automatic classification of endoscopic images by feeding them to the deep neural network (DNN) in order to obtain representative features and processing them with a global average pooling procedure and a novel cost-sensitive loss function called the class-imbalanced (CI) loss for endoscopic image classification, which can adapt to focus more on hard samples and address the challenges posed by imbalanced and hard samples in endoscopic image classification. In their study, the authors used two datasets. The first dataset was a gastrointestinal dataset collected from Baerum Hospital in Norway during gastroscopy and colonoscopy procedures. It contains 10,662 JPEG-formatted labeled images containing anatomical landmarks, pathological findings, and normal esophageal and colonic findings. This dataset's data have an imbalance ratio of 191; the data distribution is very skewed. The second dataset was the Hyper-Kvasir dataset [26], which is a large multiclass public dataset with 23 categories, including various stages of esophagitis, ulcerative colitis. This dataset poses a difficult classification problem and is highly imbalanced. As it relates to the binary class classification task, the authors developed a polyp dataset that consists of 22,935 images with rich parts and contents that were gathered from Shenzhen University General Hospital between 2019 and 2021. The study's findings revealed that when implemented in pyramid vision transformer version 2 (PVTv2-B1), their proposed class imbalance (CI) loss function outperformed other competing loss functions when evaluating the binary class classification task using six metrics for evaluation and five evaluation metrics in the multiclass classification task, with the highest accuracy of 94.75% in binary classification and 90.82% in multiclass classification. While using the Inception-v3 model, they achieved the highest accuracy of 91.55% in multiclass classification. They propose that future work could enhance the performance of the model by combining the dynamically weighted balanced (DWB) loss with an exponential function in the CI loss term.

In 2023, Shen et al. [27] used deep learning (DL) models (CNN and EfficientNet-b0) to detect and classify polyps in colonoscopy images. They used a large hospital-based dataset derived from several hospitals to improve the DL models' accuracy, and they compared results of the model classification to those of a group of pathologists. The models were put into an interactive GUI (graphical user

interface) and given the name EndoAim TM system. For their study, the authors used a large hospital-based dataset from three hospitals. Each hospital contributed 50 colonoscopy videos to the dataset, for a total of 150 colonoscopy videos. They also collected 385 images of narrow-band imaging (NBI) and single polyp images. The NBI images contained 193 images of non-adenoma tissue and 192 images of adenoma tissue, in addition to a dataset for a preliminary study that contains 10,000 images of colonoscopy polyps [28]. For hospitals A, B, and C, the model performance for polyp detection averaged 0.9516, with lesion-based sensitivity values of 0.9817, 0.9389, and 0.9360, respectively. With respect to polyp classification, the system achieved a mean average precision (mAP) of 0.89 and a sensitivity of 0.92. The authors promised to validate models on other patients from different medical organizations in a future study.

In 2023, Lewis et al. [29] proposed polyp segmentation network (PSNet), which is a dual encoder–decoder architecture used in medical image segmentation. The PS encoder, a novel CNN-based encoder, and a transformer-based encoder make up the dual encoder. For their study, the authors used five publicly available datasets: Kvasir-SEG [30], CVC-ClinicDB [20], CVC-ColonDB [31], ETIS [32], and EndoScene [33]. Using training, validation, and test datasets, they assessed the performance of their model in comparison with other recent models. According to the authors, their model produced average mean dice (mDice) and mean intersection over union (mIoU) scores of 0.863 and 0.797 for each of the five datasets. They addressed important issues like model overfitting and accurately capturing polyp characteristics like size and texture in their insightful research.

In 2023, Zhu et al. [34] presented a new feature extraction module called the global-local context module (GLCM) and a multi-modality cross-attention (MMCA) module for integrating background data, polyp regions, and boundary areas. These modules have been incorporated into CRCNet, a foundational network for polyp segmentation. Pyramid networks (FPN) are a feature of the architecture of CRCNet, which uses an altered variant of the standard U-Net as its foundation. It has an evenly distributed encoder-decoder architecture and a total of five layers. The authors of the study used two datasets: Kvasir-SEG [30] and CVC-ClinicDB [20]. There are 1,000 different colonoscopy images in the Kvasir-SEG collection, each with a unique size, angle, and texture. All images are expertly labeled by experienced professionals to precisely meet clinical requirements. CVC-ClinicDB contains 612 colonoscopy images. A number of metrics, including dice coefficient, mIoU, recall, precision, accuracy, and F2-score, were used to evaluate CRCNet's performance. In terms of segmentation accuracy, the results showed that

CRCNet performed better than other state-of-the-art methods, with dice scores on the Kvasir-SEG and CVC-ClinicDB datasets of 91.59% and 95.02%, respectively. They mentioned that one of the limitations of their study was the extremely small amount of data that were used, which had an impact on the model's performance and necessitated optimization. In order to calculate the cost and size of the model, they will also try to integrate it with system clinics. In the future, the author plans to use saturation modification, edge-aware blind deblurring, and object elimination techniques to enhance the level of accuracy of edge segmentation and polyp localization. Model quantization and distillation techniques can be used to shrink existing models and incorporate high-resolution hardware devices such as medical endoscopes.

In conclusion, the proposed study represents a substantial advance in the field of coloscopy by tackling a variety of principal limitations mentioned in prior research. In the beginning, we addressed the issue of limited datasets by applying data augmentation methods and utilizing transfer learning models. This not only enriches the accessible data but also improves the generalizability of the proposed approach. Furthermore, PolyDSS accurately detects small polyps, including those with flat characteristics, which were frequently missed in previous studies. The proposed segmentation and classification model surpasses prior attempts, making it a more robust tool for coloscopy specialists. One of the study's most notable accomplishments is the development of a user-friendly graphical interface that merges decision support system capabilities with computer-aided diagnosis. This integration empowers endoscopists by providing them with important insights and recommendations, thereby enabling more accurate and rapid decision-making during procedures. Furthermore, the PolyDSS model excelled at classifying polyps using the standard Paris classification method, allowing for more informed decisions about resection techniques. This not only assists in the prevention of colorectal cancer, but it also aids in early detection.

In summary, the proposed study fills significant gaps in the coloscopy field highlighted in Table 2 by improving data availability, segmentation and classification accuracy, and practical decision support. By addressing the limitations of earlier studies, we present a comprehensive method to greatly enhance patient outcomes.

4 Materials and methods

This section presents a comprehensive view of the dataset used in this study, as well as a broad spectrum of techniques and methods used to support the PolyDSS model.

Table 2 Summary of literature survey

Author (s)	Dataset	Preprocessing	Methodology	Evaluation tools	Advantage	Disadvantage
Hsu et al. [19]	–CVC Clinic –Linkou –Chang Gung Medical Hospital	–RGB images to gray images –Feature extraction using CNN	–Polyp detection using YOLOv2, YOLOv3 –Polyp classification	–Accuracy –Precision –Recall –F1-measure –F2-measure	–Improved patient outcome –Avoided unnecessary polypectomy –Reduced misdiagnosis rate of endoscopists	–Polyps in images less than 1,600 pixels are missed –Hard to distinguish Neoplastic and Hyperplastic polyps than 1,800 pixels
Lo et al. [21]	Taipei Medical University –Joint Institutional dataset	–GLCM to extract features –PCA to reduce feature dimension	–Feature extraction –Built polyp classification model by combining features	–Accuracy –Sensitivity –Specificity	DCNN can accomplish significant performance and training from scratch	Dataset is limited
Krenzer et al. [22]	–SUN database –Endo Data dataset	–Detected and cropped polyp –Classified the polyp by transformer –Used FSL for data scarcity	Two pretrained CNN models and a pretrained transformer were used for Paris classification	F1-score	81.34% illustrating the FSL approach's efficacy in data-scarce scenarios	–Shortage of labeled datasets –Test datasets were unbalanced –High cost and time for expert annotations
Yue et al. [25]	–Hyper-Kvasir dataset –Baerum Hospital –Shenzhen University Hospital	DNN to extract features from images	A DNN-based classification method consists of a feature extractor and a classifier	–ACC –BACC –AUC –MCC –Kappa –G-mean –F1	Class imbalance (CI) loss reweighting scheme to pay more attention to the minority classes	Some hard samples can lower the classification accuracy
Shen et al. [27]	Dataset derived from several hospitals	Blurry and Low-contrast images removed	An AI-based system for the detection and classification of polyps using colonoscopy images	–Accuracy –Sensitivity –Specificity –AUC –F1-score –mAP –ROC	Models are deployed into an Olympus CV290 machine, which is ready to use in clinical practice	Small validation size, affected model's accuracy and robustness
Lewis et al. [29]	–Kvasir –SEG –CVC –ClinicDB –CVC –ColonDB –ETIS –EndoScene	–Images resized to 512×512 –Encoder and decoder as a merge module –Used LFE module and CCM	Dual encoder-decoder architecture	–mDice –mIoU	Polyps of a wider variety of shapes, colors, and sizes have a significant effect on the accuracy of the model	Models need to be simplified
Zhu et al. [34]	–Kvasir –SEG –CVC –ClinicDB	–Used GLCM to capture global and local information –Applied MMCA module to integrate background data, and boundary areas	Encoder-decoder framework for polyp segmentation	–Dice coefficient –mIoU –Recall –Precision –Accuracy –F2-score	In comparison with traditional methods, the proposed model is very competitive in terms of accuracy and computing efficiency	–Extremely small amount of data that was used –The size of the model and cost of testing must be reconsidered

4.1 Dataset

The PICCOLO dataset [35] was used in this study. The dataset was obtained from Hospital Universitario and contains a total of 3433 precisely annotated images, which are divided into two primary categories: narrow-band images (1302) and white-light images (2131), which were taken from 76 lesions and 40 patients, respectively. Additionally, the dataset is associated with metadata that includes the number of polyps that were present throughout the operation, their sizes, two polyp classification systems (Paris and Nice), preliminary and final diagnosis, and histological stratification. The dataset is divided into six classes: 0-Ip, 0-Ips, 0-Is, 0-IIa, 0-IIa/c, and 0-IIb. Table 3 presents the number of images in each class, while Fig. 2 shows the polyp class distribution, and finally, Fig. 3 shows a representative image for each class. Along with the clinical metadata, the dataset contains train, test, and validation folders, as well as masks and polyps folders within each folder. We encountered the issue that the images are not classified into classes or categorized based on any classification system. This issue was handled by applying a simple preprocessing algorithm that matched the video code column, the Paris classification column, and the image name in the existing folders. This enabled us to classify each image into its own class based on the Paris classification system.

4.2 Methods

In this section, the study introduces the methods used to improve the performance of image segmentation and classification tasks. Deep learning models, notably ResU-Net and ResUNet++, are used, and they are combined with locally shared features. Furthermore, multiple EfficientNet model versions (B0, B1, B2, B3, B4, and B5) are used to provide a strong foundation for the classification tasks. Furthermore, ensemble learning is used, with a majority voting procedure, to collectively improve the predictive skills of the EfficientNet model versions.

Table 3 The number of images in each class in the train, validation, and test sets [35]

Class	Train Set	Validation Set	Test Set
0-Ip	274	81	29
0-Ips	245	41	14
0-Is	433	176	31
0-IIa	1052	263	49
0-IIa/c	27	122	25
0-IIb	-	48	58

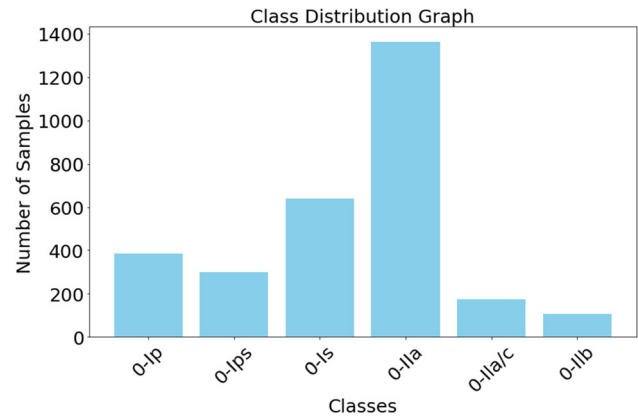


Fig. 2 Polyp class distribution (total number of images in train, validation, and test sets)

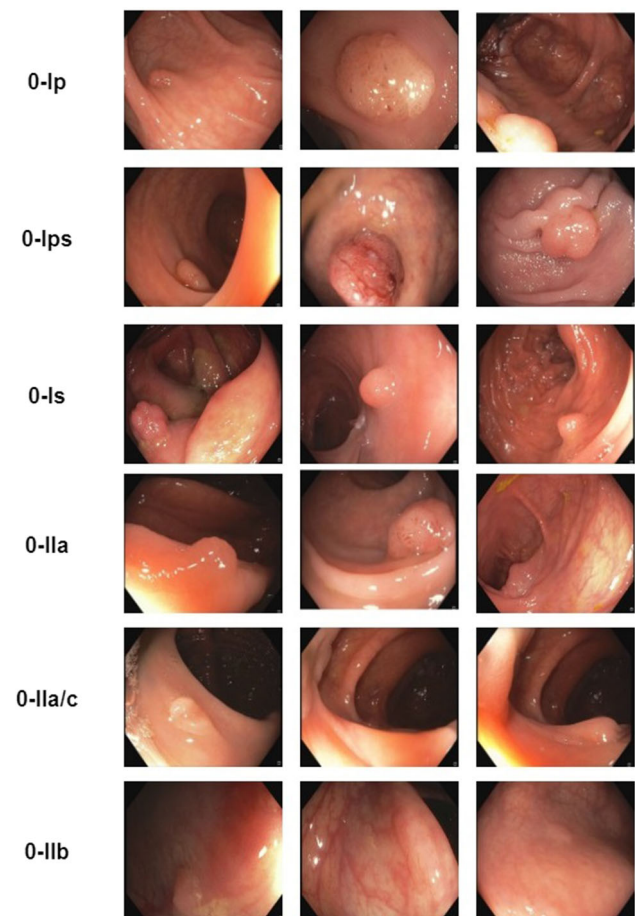


Fig. 3 PICCOLO Dataset polyp image samples from each class

4.2.1 Data augmentation techniques

In this study, we used various data augmentation techniques to overcome class imbalances and improve the performance of deep learning models [36]. These augmentation techniques were chosen based on a combination

of actual findings and acknowledged best practices in the fields of computer vision and image processing. The goal was to include diversity and variability in the training data while keeping the enhanced data representative of the underlying classes. Here is a complete overview of the data augmentation techniques that used, as well as how they were chosen and implemented:

1. *Flipping* Images were flipped or horizontally mirrored. When working with polyp images, flipping can be advantageous. Because polyps can develop in many orientations inside the colon, rotating the images horizontally can assist provide more training samples to guarantee the model learns to recognize polyps regardless of their orientation.
2. *Random rotation* To imitate images from various angles, random rotation was used. Polyps can be found inside the colon at a variety of angles and orientations. Random rotation augmentation is required for training the model to be insensitive to polyp orientation fluctuations. This is especially crucial when it comes to increasing the model's ability to recognize polyps in endoscopic images where the camera angle varies.
3. *Scale augmentation* The process of scaling augmentation entailed resizing images to varied dimensions, whether larger or smaller than the original size. Polyps can differ in shape and size based on the scale of the image. Scaling augmentation is useful for training the model to recognize polyps of varying sizes, making it more resistant to polyps that appear differently in different endoscopic examinations.
4. *Contrast modification* Images were subjected to contrast modification, which altered the intensity of pixel values. Contrast adjustment enables the model to accommodate variations in lighting and precisely detect and classify polyps in images with varying levels of illumination.
5. *Zooming* In order to segment and classify polyps, the model must focus on the region of interest within the image. By focusing on polyp regions within a larger endoscopic image, zooming augmentation is useful for training the model to recognize and classify polyps.

Previous work in the field and its relevance to the specific problem being addressed in this study influenced the selection of these augmentation strategies. These strategies were used to increase dataset diversity while maintaining class features such as polyp forms and attributes. They were used during the training data preprocessing to generate augmented versions of the original images, thereby extending the training dataset. These augmentation techniques were used with proper parameter settings to guarantee a fair representation of the classes and to minimize overfitting. In addition, we used rigorous testing and cross-

validation to identify the most successful combination of augmentations.

In conclusion, we chose data augmentation techniques based on their capability to mitigate class imbalance, increase model generalization, and make the model more resistant to fluctuations in input data. These techniques were used and verified to improve the overall performance of the deep learning model.

4.2.2 Deep learning models

Two deep learning models were used in segmenting polyps: ResUNet and ResUNet++. The ResUNet architecture combines the U-Net architecture with residual learning, which relies on the idea of residual blocks [37]. Residual blocks introduce skip connections, which facilitate the ability of the model to get past certain layers, allowing information to flow directly from the initial layers to later layers. ResUNet++ is a ResUNet broadening that improves the skip connections to collect deeper contextual information [38]. It presents several routes or branches in the network's encoder and decoder sections, enabling the model to learn features at various scales [39]. ResUNet and ResUNet++ have both been used effectively for polyp segmentation in medical image analysis [40–43]. Their ability to utilize skip connections and residual learning has enabled them to handle complex and diverse image datasets effectively. So, in the segmentation module, we used the two models to segment the polyps, which will be used as input for the classification module.

4.2.3 Locally shared feature (LSF)

The concept of benefiting from information from adjacent regions in an image to boost segmentation performance is referred to as locally shared features (LSF) [44]. Locally shared features play a vital role in polyp segmentation, where accurately defining polyp boundaries are critical. The model is able to better represent the shape, texture, and spatial relationships between pixels by taking into account the contextual information surrounding a polyp region, resulting in more precise segmentation results. Integrating LSF with ResUNet and ResUNet++ improves segmentation abilities by taking into account spatial relationships and contextual information near polyp regions. The LSF equation can be donated as follows:

$$L_{i,j} = \phi(X_i, W_{lsc}) \quad (1)$$

$L_{i,j}$ donates the feature representation at a specific position in the resulting segmentation map, ϕ expresses the function used to apply LSF, and finally, W_{lsc} represents the locally shared convolutional weights along the positions in the

output map, while X_i represents the image or feature map. When compared to fully connected layers, this enables the model to capture spatial dependencies and context information while minimizing the number of parameters. We developed the LSF as a convolutional block that uses a convolutional operation through exactly the same number of weights at various spatial locations in order to integrate it with ResUNet and ResUNet++. The feature maps that originate from these layers are then combined with the equivalent feature maps from the ResUNet or ResUNet++ architecture. This fusion is accomplished by shifting and concatenating the feature maps from the locally shared convolutional layers with the already-existing feature maps. The locally shared features are then integrated with the remaining parts of the network by passing the fused feature maps through additional convolutional layers.

convergence, and better computational resource utilization. These pretrained models are capable of identifying relevant features from polyp images even when relatively little data is available for training by utilizing knowledge learned from large-scale datasets and complex tasks. In this study, we used five EfficientNet variants: B0, B1, B2, B3, and B4. For significant performance and efficiency, it is crucial that there be a sufficient number of layers and parameters in the EfficientNet models, especially those from B0 to B4. The number of layers in EfficientNet models allows for the capture of complex patterns and representations from the input data, whereas the number of parameters in EfficientNet models influences the model's capacity to learn and represent information. Table 4 compares the number of layers and parameters of various EfficientNet models.

Algorithm 1 WAEI

```

Input: Training-data, validation-data, validation-labels, model-configs, and test-data
Output: Final-predictions
Models = [ ];
Predictions = [ ];
for model-config in model-configs do
    | Model = train-model(training-data, model-config);
    | Models.append(model);
end
for model in models do
    | Model-predictions = make-predictions(model, validation-data);
    | Predictions.append(model-predictions);
end
Weights = assign-weights(predictions, validation-labels);
Final-predictions = weighted-average(predictions, weights);
Test-data-predictions = make-predictions-using-ensemble-learning(final-predictions, test-data);

```

4.2.4 Transfer learning models

For improving performance and effectiveness in the field of polyp classification, transfer learning is essential. A collection of convolutional neural network (CNN) models called EfficientNet was created in order to achieve outstanding performance while minimizing computational costs and model size [45]. The architecture of EfficientNet is based on a compound scaling technique that steadily increases the depth, width, and resolution of the network [46–48]. Known for their effectiveness and accuracy, EfficientNet variants can be used as pretrained models in transfer learning for polyp classification tasks. This considerably reduces the requirement for large data collections and training time. Transfer learning with EfficientNet variations helps the model reap the benefits of the generalization power of the pretrained models, resulting in improved polyp classification accuracy, more rapid

4.2.5 Ensemble learning and majority voting

Ensemble learning is an effective machine learning technique that brings together various models to make better

Table 4 The number of layers and parameters of different EfficientNet model architectures

Model	Number of layers	Number of parameters
B0	132	5.3 million
B1	186	7.9 million
B2	186	9.2 million
B3	210	12.3 million
B4	258	19.5 million
B5	312	30.6 million
B6	360	43.3 million
B7	438	66.7 million

predictions [49–51]. Majority voting is one of the most common approaches in ensemble learning. Ensemble learning can be used to improve the performance of the classifiers [52–55] in regards to polyp classification using EfficientNet variations. We generated an ensemble of classifiers by training multiple EfficientNet models with different initializations or hyperparameters. During inference, each classifier predicts a polyp image, and the class with the most votes is chosen as the final prediction. This method reduces any biases or restrictions of individual models, resulting in enhanced accuracy and robustness for polyp classification tasks. To compute the majority vote, \hat{y} is the class label to be predicted, and $C_m(\mathbf{x})$ is the number of classifiers.

$$\hat{y} = \text{mode}\{C_1(\mathbf{x}), C_2(\mathbf{x}), \dots, C_m(\mathbf{x})\} \tag{2}$$

Assume we have six class labels to predict in our case study: 0-Ip, 0-Ips, 0-Is, 0-IIa, 0-IIa/c, and 0-IIb labeled as class 0, class 1, class 2, class 3, class 4, and class 5, and we have five classifiers B0, B1, B2, B3, and B4. If the classifier results are as follows:

- B0 predicts 0-Ip (class 0)
- B1 predicts 0-Is (class 2)
- B2 predicts 0-IIa (class 3)
- B3 predicts 0-IIa (class 3)
- B4 predicts 0-IIa (class 3)

$$\hat{y} = \text{mode}\{0, 2, 3, 3, 3\} = 3 \tag{3}$$

Using Eq. (2), we would classify the final output class sample as “class 3” using a majority vote, which is 0-IIa, thereby making it a flat elevated lesion. Assume the following classifier results:

- B0 predicts 0-Ip (class 0)
- B1 predicts 0-Is (class 2)
- B2 predicts 0-IIa (class 2)
- B3 predicts 0-IIa (class 3)
- B4 predicts 0-IIa (class 3)

We can observe a tie here, and we cannot decide the final predicted class. In this case, the predictions of various individual models or learners in an ensemble are combined using the weighted average ensemble learning (WAEL) technique demonstrated in Algorithm 1, which is utilized in deep learning and statistics. This method uses a weighted average of all the various model predictions to determine the final prediction which can be calculated as follows:

$$\begin{aligned} FinalP = & (W1 \times P1 + W2 \times P2 + \dots \\ & + WN \times PN) / (W1 + W2 + \dots + WN) \end{aligned} \tag{4}$$

where *FinalP* is the final prediction for a given data point since it represents the ensemble’s prediction for that data point after merging the predictions from several individual models. As for the weights given to each individual model in the ensemble,

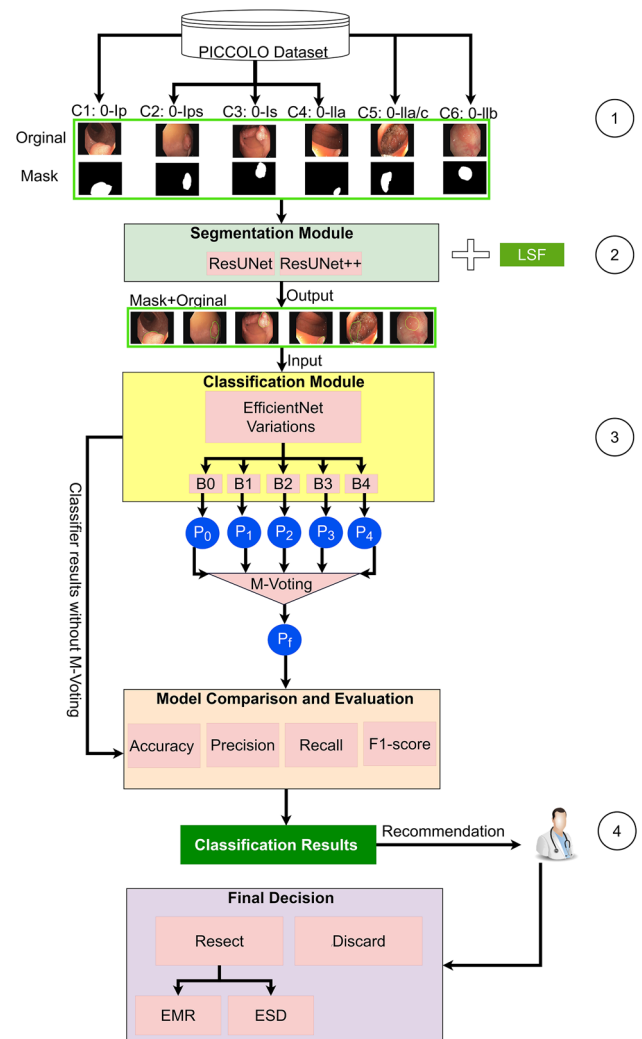


Fig. 4 An overview of the PolyDSS model flow for polyp segmentation and classification

they are represented by $W1, W2,$ and $WN,$ while the predictions that each individual model in the ensemble makes for the same data point are represented by $PN.$ Each model provides its own prediction based on its learned patterns and characteristics. The purpose of assigning weights is to give greater significance or influence to specific models in the ensemble based on their performance or reliability. Higher weights are often given to models that are more dependable or have higher confidence in their predictions, whereas lower weights or even exclusion from the ensemble may be given to less reliable models.

4.3 Proposed model for polyp segmentation and classification

In this section, the PolyDSS model for polyp segmentation and classification consists of four phases and is shown in Figure 4. A detailed explanation of each phase is presented below.

4.3.1 Phase 1: dataset preprocessing

Several essential tasks were carried out during the dataset preprocessing phase to set up the data for later analysis and deep learning model training. The initial dataset included a clinical information file as well as three core folders: train, test, and validation, with subfolders masks and polyps for masks and associated polyp images in each. These images were initially disorganized and unclassified. To address this, images were classified using a simple algorithm into their proper class using the clinical metadata's frame number and class information based on the Paris classification system. Image format and size were standardized to maintain consistency and compatibility, with all images shrunk to 512×512 pixels and converted to the PNG format. Furthermore, data augmentation techniques were used to increase the diversity of the dataset and improve model robustness. Flipping, random rotation, scale augmentation, contrast change, and zooming were among the techniques used. Experiments were run using three different levels of augmentation, yielding datasets with 500, 1364, and 2000 images per experiment. The goal of this modification in dataset design was to make it easier to train more resilient and generalizable deep learning models for tasks like polyp detection and analysis.

4.3.2 Phase 2: segmentation module

The segmentation phase began with the use of two primary inputs: the preprocessed original medical image and its related mask image. These images were used to lay the groundwork for subsequent segmentation tasks. To achieve reliable polyp segmentation within medical images, deep learning models, notably the ResUNet and ResUNet++ architectures coupled with the LSF approach, were used as segmentation models trained on the dataset. This technique produced predicted mask pictures that outlined the regions of interest, especially the polyps. A contour-finding algorithm was then used to improve the visual interpretability of the results. This algorithm compares the original image to the anticipated mask image to determine the polyp boundaries. Following that, a green outline was placed on the original image, precisely outlining the area of interest, which in this case was the location and form of the polyps. This stage was critical in distinguishing and visualizing polyp regions within medical imaging, allowing for subsequent analysis and classification tasks.

4.3.3 Phase 3: classification module

During the classification phase, the process began with the utilization of the original medical image, which featured a distinct green contour outlining the area of interest, which

corresponded to the polyps. This highlighted image was used as the input for the classification module. To properly classify the highlighted images and detect various polyp types, a set of five EfficientNet versions (B0, B1, B2, B3, and B4) were trained on the given dataset. An ensemble learning approach, specifically the majority voting technique, was used to boost the overall performance and dependability of the classification models. The predictions provided by the multiple EfficientNet models were used in this technique, and the final classification decision was based on a majority vote. This ensemble learning method not only enhanced classification accuracy but also reduced the risk of overfitting, resulting in a precise and reliable classification of the discovered polyps.

4.3.4 Phase 4: model comparison and evaluation

During the model comparison and evaluation phase, the classification models were thoroughly evaluated using well-established evaluation metrics such as accuracy, precision, recall, and F1-score. This meticulous assessment sought to assess the models' ability to correctly classify the indicated medical images. In addition to individual model evaluations, a comparative study was performed to compare the performance of the ensemble learning model with the majority voting technique to individual models. The evaluation's findings were then presented to healthcare professionals as recommendations for their input cases. These recommendations included key information such as predicted polyp class, class risk, and resection recommendations, which provided valuable assistance for healthcare professionals in their clinical decision-making processes. Finally, based on the system's recommendations, healthcare professionals made final decisions, leveraging the potential of AI-driven classification and recommendations to improve the precision and quality of their diagnostic and treatment planning efforts.

5 Experimental results

In this section, we will present experimental implementation details, evaluation metrics used to evaluate the models, results comparisons and discussions, and finally ablation studies.

5.1 Implementation details

The PolyDSS model was trained on a system equipped with an 11th Gen Intel (R) Core (TM) i7-11800H @ 2.30 GHz processor, 16 GB of Random Access Memory (RAM), a NVIDIA GeForce RTX 3060 Graphical Processing Unit (GPU), and 1 Terabyte (TB) of Solid State

Drive (SSD) storage. The Anaconda 2.0 environment, Windows 10 Pro as an operating system, and the Python 3.7 programming language were used for all of the experiments. The data were split into 80% training, 10% validation, and 10% testing, and the PyTorch framework is utilized to implement each model.

5.2 Proposed model evaluation methodology

In our study, we used multiple evaluation metrics, including the Dice Similarity Coefficient (DSC), accuracy, precision, recall, and the F1-score. These metrics were selected due to their ability to provide a complete assessment of the effectiveness of our polyp segmentation and classification model. The selection of these metrics was intentional, as they each serve various roles in assessing the model's effectiveness. DSC measures the spatial overlap between predicted and ground truth polyp areas, which is necessary for determining segmentation accuracy. DSC is especially appropriate for polyp segmentation since it compensates for both erroneous positives and false negatives. Polyp areas can vary in size and form, and DSC is sensitive to these changes. Accuracy, an important metric for classification, assesses overall correctness, whereas precision focuses on avoiding false positives, a critical aspect in polyp classification to avoid unneeded clinical interventions. Recall, also known as sensitivity, assesses the ability of the model to gather all polyp instances and guarantee that they are not missed. Finally, particularly dealing with class imbalance, the F1-Score provides a balanced evaluation as an ideal blend of precision and recall. These measures, when combined, provide a strong foundation for assessing the model's performance in polyp segmentation and classification, addressing the task's unique problems and requirements.

5.2.1 Evaluation metrics

Dice Similarity Coefficient (DSC), Precision, and Recall were used to measure the performance of different model variations in the segmentation module, while Accuracy, Precision, Recall, and F1-Score were used in the classification module. DSC is commonly used to evaluate the output of image segmentation operations because it measures the similarity between two sets of data. The following Eq. (4) defines DSC:

$$\text{DSC}(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (5)$$

DSC is twice the overlapped area between A and B divided by the number of pixels in each image. The accuracy metric is used to assess the classifier's performance in accurately predicting the class labels of a given dataset. It

is calculated by dividing the total number of instances in the dataset by the number of correctly classified instances. Precision is the percentage of positive samples that match the image ground truth, while recall assesses the model's accuracy in detecting the number of captured positive samples. The F1-score combines a classifier's precision and recall into a single measurement by calculating their harmonic mean. Equations (5)–(9) define dice Similarity coefficient, accuracy, precision, recall, and f1-score, respectively.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (6)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

In this case study, TP stands for true positive (the number of patients who were correctly identified as having polyps), TN stands for true negative (the number of patients who were correctly identified as healthy and free of polyps), FP stands for false positive (the number of patients who were incorrectly identified as having polyps), and FN stands for false negative (the number of patients who were incorrectly identified as healthy and free of polyps).

5.2.2 Results and discussion

In this section, we will demonstrate the experiments and the results. The B4 model demonstrates the highest accuracy of 0.7000, as shown in Table 5, after expanding the dataset to 500 images per class. Additionally, it indicates good precision, recall, and F1-score values of 0.7952, 0.7021, and 0.7458, respectively. With an accuracy of 0.6400 and a relatively high F1-score of 0.6687, the B3 model follows closely. Compared to the more advanced models (B2, B3, and B4), the B0 and B1 have lower

Table 5 Performance of EfficientNet versions on the PICCOLO dataset's various classes after adding 500 additional images per class

Model	Accuracy	Precision	Recall	F1-Score
B0	0.5333	0.6328	0.4655	0.5364
B1	0.5800	0.8598	0.4532	0.5935
B2	0.6267	0.7820	0.5561	0.6500
B3	0.6400	0.7267	0.6193	0.6687
B4	0.7000	0.7952	0.7021	0.7458
Ensemble learning	0.7500	0.8164	0.8204	0.8184

Table 6 Performance of EfficientNet versions on the PICCOLO dataset's various classes after adding 1364 additional images per class

Model	Accuracy	Precision	Recall	F1-Score
B0	0.7347	0.8092	0.7676	0.7879
B1	0.7579	0.8596	0.7495	0.8008
B2	0.8227	0.9206	0.8392	0.8780
B3	0.8570	0.9480	0.8728	0.9088
B4	0.9034	0.9594	0.9285	0.9437
Ensemble learning	0.9425	0.9878	0.9503	0.9687

Table 7 Performance of EfficientNet versions on the PICCOLO dataset's various classes after adding 2000 additional images per class

Model	Accuracy	Precision	Recall	F1-Score
B0	0.7250	0.8961	0.6831	0.7752
B1	0.7642	0.9201	0.7352	0.8173
B2	0.8117	0.9456	0.7987	0.8660
B3	0.8325	0.9599	0.8245	0.8870
B4	0.8458	0.9645	0.8404	0.8982
Ensemble learning	0.8667	0.9757	0.8647	0.9168

accuracy and F1-scores. This is because larger models can learn more patterns and representations from the data, improving classification performance.

Referring to Table 6, all EfficientNet models perform better with the addition of 1364 images per class. The B4 model has the highest accuracy of 0.9034, high precision of 0.9594, high recall of 0.9285, and a high F1-score of 0.9437. The B3 model performs exactly well, with accuracy of 0.8570, high precision of 0.9480, and recall of 0.8728 values. In comparison with the B2, B3, and B4 models, the B0 and B1 models are less accurate. This is due to the core concept of EfficientNet, which is based on the idea of compound scaling, which entails scaling the model's width, depth, and resolution all at once. The number of layers and the quality of feature maps increase as the model

Table 8 Performance of EfficientNet versions on the PICCOLO dataset's various classes before augmentation

Model	Accuracy	Precision	Recall	F1-Score
B0	0.6094	0.6164	0.6405	0.6282
B1	0.7441	0.7814	0.7989	0.7901
B2	0.7845	0.8653	0.8146	0.8392
B3	0.7912	0.8529	0.8447	0.8488
B4	0.8114	0.8498	0.8829	0.8660
Ensemble learning	0.8316	0.8784	0.8945	0.8864

versions advance from B0 to B4. As a result, the models are better able to distinguish between various classes because they can capture more tiny details and progressively extract features at various levels.

Table 7 shows that the B4 model maintains the highest accuracy of 0.8458 and also indicates high values for precision of 0.9645, recall of 0.8404, and F1-score of 0.8982 after expanding the dataset to 2000 images per class. With an accuracy of 0.8325, a precision of 0.9599, and an F1-score of 0.8870, the B3 model follows closely.

The performance of the EfficientNet models improves practically as the dataset size increases when compared to the initial dataset in Table 8. In comparison with the lower versions (B0 and B1) of EfficientNet models, the higher versions (B2, B3, and B4) more frequently achieve better accuracy and F1-scores. The findings indicate that larger datasets help more complex models perform better on classification tasks. Figure 5 displays the segmentation module's qualitative analysis, showing the segmentation output using ResUNet + LSF and ResUNet++ + LSF compared to ground truth and the classification output using ensemble learning majority voting of EfficientNet variations. Due to the use of LSF, which makes it simpler to combine high-resolution information with global context, the PolyDSS can distinguish small polyps from background tissue as shown in polyp class 0-llb that shares the same color, which is more challenging. This allows the network to gather the local detailed features and contextual information required for precise segmentation.

5.2.3 Comparison with existing methods

Table 9 provides a comparative analysis of multiple methods in the field of colonoscopy that evaluates the performance of various models on diverse datasets. Lo et al. [21] investigated several models, including AlexNet, Inception-V3, ResNet-101, and DenseNet-201, with accuracy scores ranging from 78.2% to 87.7%. Nevertheless, they did not mention precision or F1-score. Krenzer et al. [22] applied their model to the SUN and EndoData datasets, achieving accuracy scores of 89.35% and 87.42%, respectively, with precision and recall values in the 80% range. On the HYPER-KVASIR dataset, Yue et al. [25] used MobileNet-v2, Inception-v3, and PVTv2-B1, achieving accuracy scores above 90%, but precision and recall values were not provided. Huang et al. [56] used their model to get an accuracy score of 87.1% while maintaining balanced precision and recall values on the Chang Bing Show Chwan Memorial Hospital dataset. Notably, the proposed method, an EfficientNet ensemble learning model on the PICCOLO dataset, outperformed earlier studies with a remarkable accuracy of 94.25%. Also, the PolyDSS model has a good precision of 98.78% and recall of 95.03%,

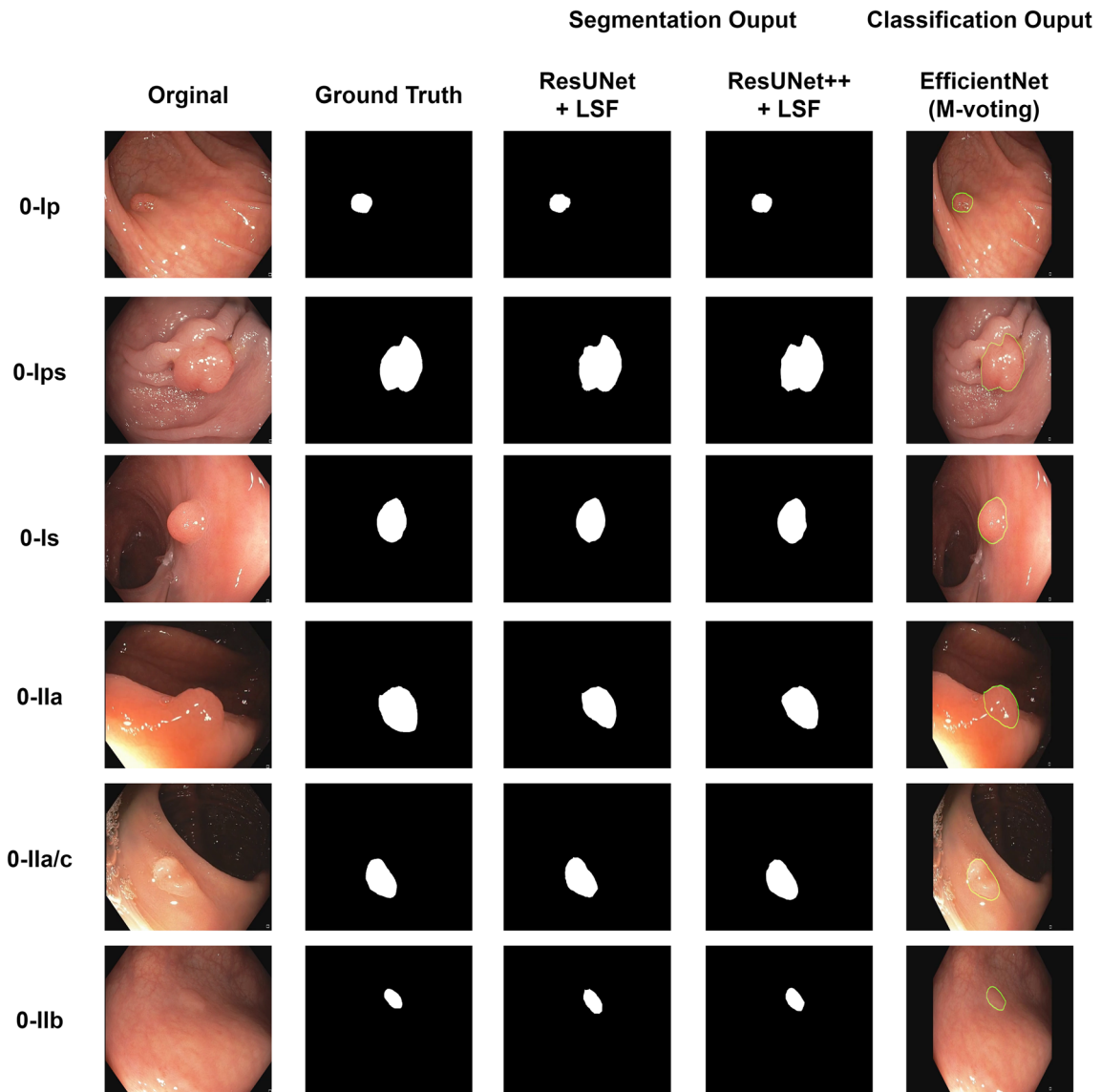


Fig. 5 Analysis of the segmentation and classification module’s results

Table 9 Comparing the results of the PolyDSS model with the existing models, significant values are in bold

Author (s)	Model	Dataset	Accuracy	Precision	Recall	F1-score
Lo et al. [21]	Alex	Created by authors	81.3	–	90.4	–
	Inception-V3		78.2	–	67.0	–
	ResNet-101		85.3	–	81.9	–
	DenseNet-201		87.7	–	83.0	–
Krenzer et al. [22]	[22]	SUN	89.35	84.76	79.10	81.28
		EndoData	87.42	80.09	78.83	79.45
Yue et al. [25]	MobileNet-v2	HYPER-KVASIR	90.44	–	–	65.12
	Inception-v3		91.55	–	–	62.84
	PVTv2-B1		90.82	–	–	62.54
Huang et al. [56]	[56]	Chang Bing Show Chwan Memorial Hospital	0.871	0.859	0.896	0.875
Saad et al.	PolyDSS	PICCLO	0.9425	0.9878	0.9503	0.9687

yielding an impressive F1-score of 96.87%. This demonstrates the efficacy of our approach to polyp segmentation and classification, providing a potentially valuable contribution to the area of interest by addressing some of the limitations noted in prior studies, such as data scarcity and low accuracy.

5.2.4 Ablation studies

In the section that follows, we are going to discuss how different parts of our model affect the outcome of our final output. Data augmentation has become a critical method for resolving issues with small datasets in deep learning applications. The performance of the segmentation and classification can be enhanced when segmenting or classifying a specific disease with only a limited amount of data. This is done by enhancing the dataset. In this ablation study, we primarily looked into how data augmentation affected the segmentation and classification of six classes related to colorectal cancer. Three distinct levels of data augmentation were used in each experiment. In the first experiment, 500 images from each class were added using a variety of augmentation methods, including flipping, random rotation, scaling, contrast modification, and zooming. In the second experiment, all classes were expanded to 1364 images, creating a larger dataset for training and evaluation. The third experiment, in the same manner, produced 2000 images for each class in an effort to further enhance the data. The main criteria for this augmentation approach are class balance, model generalization, and data quality. The rationale behind the choices to augment the number of images in the dataset is that initial class distribution is very unbalanced, with 106 to 1364 images per class, depending on the class. A minimum amount of data augmentation would be to increase each class to 500 images. The dataset size is increased without being overinflated, which could be computationally expensive and result in overfitting if not handled appropriately. When compared to more severe augmentation algorithms, the augmentation of 500 images is computationally manageable and does not require a lot of computational power. The number of augmented images per class in the following experiment equals the largest number of original images 1364. By guaranteeing that each class contains an equal number of original and augmented images, this decision helps maintain the distribution of the original data. A moderate augmentation technique can involve enhancing to match the original images with the most original image counts. It does not add synthetic data that significantly outweigh the original data, which can be desired in medical imaging scenarios where synthetic data adds ambiguity. Additionally, because the distribution of the augmented data matches that of the original data, we

will have greater confidence in its quality and relevance. Increasing the number of images in each class to 2000 is a bold data augmentation tactic. This decision may be advantageous because it greatly expands the dataset size and exposes the model to a wider range of data variances. Aggressive augmentation may improve generalization since it exposes the model to a wider variety of samples during training. In addition to working with complex medical images that could have different lighting conditions, orientations, or other factors, which is the case in polyps, this may be extremely helpful. The class imbalance problem can be addressed by augmenting 2000 images, which reduces the likelihood that the model will favor overrepresented classes in its predictions. Finally, our goal was to carry out extensive experiments to find the optimal augmentation approach for accurate segmentation and classification as well as investigate how different levels of data augmentation affected the system's performance. The segmentation module's performance was evaluated using DSC, precision, and recall, while the classification module's performance was evaluated using accuracy, precision, recall, and the F1-score.

5.2.5 Ablation study 1: how data augmentation affects the segmentation module

We used the augmented dataset to train, validate, and test our model for each experiment, and then, we reported the results. Without data augmentation, the ResUNet + LSF model generated preliminary findings with a Dice Similarity Coefficient (DSC) of 0.7890, as shown in Table 10. When the dataset was increased to 500 images, the model significantly improved, with the DSC increasing by about 7.35%, from 0.7890 to 0.8625, as shown in Table 11. The segmentation accuracy has increased by 7.35%, which is a significant improvement. The ResUNet++ + LSF model's DSC increased by roughly 9.7%, from 0.8071 to 0.9041, as shown in Table 11, indicating an even more impressive improvement that signifies a notable 9.7% increase in segmentation accuracy. In the second experiment, where the dataset was increased to 1364 images, both models displayed improvements. The ResUNet + LSF model's DSC was 0.8994, which is an improvement of about 3.69%

Table 10 Performance of the segmentation module for the PICCOLO dataset using ResUNet + LSF and ResUNet++ + LSF before augmentation

Model	DSC	Precision	Recall
ResUNet + LSF	0.7890	0.7580	0.7964
ResUNet++ + LSF	0.8071	0.7836	0.7694

Table 11 Performance of the segmentation module for the PICCOLO dataset using ResUNet + LSF and ResUNet++ + LSF after 500 augmented images per class

Model	DSC	Precision	Recall
ResUNet + LSF	0.8625	0.8417	0.8156
ResUNet++ + LSF	0.9041	0.8912	0.8778

Table 12 Performance of the segmentation module for the PICCOLO dataset using ResUNet + LSF and ResUNet++ + LSF after 1364 augmented images per class

Model	DSC	Precision	Recall
ResUNet + LSF	0.8994	0.8797	0.8615
ResUNet++ + LSF	0.9244	0.8839	0.8827

Table 13 Performance of the segmentation module for the PICCOLO dataset using ResUNet + LSF and ResUNet++ + LSF after 2000 augmented images per class

Model	DSC	Precision	Recall
ResUNet + LSF	0.8345	0.8163	0.7653
ResUNet++ + LSF	0.8574	0.8289	0.8048

over the outcomes of the earlier experiment, as shown in Table 12. Similar improvements were seen in the ResUNet++ + LSF model, with the DSC rising from 0.9041 to

0.9244, indicating a 2.03% increase in segmentation accuracy, as shown in Table 12. The models performed differently in the third experiment, where the dataset was expanded to 2000 images. The DSC for the ResUNet++ + LSF model decreased to 0.8574, while the DSC for the ResUNet + LSF model was 0.8345, which, respectively, represents a decrease of 6.49% and 6.7% as shown in Table 13 compared to the previous experiment; we believe this is the result of over-augmentation.

5.2.6 Ablation study 2: how data augmentation affects the classification module

The models’ accuracy, precision, recall, and F1-score varied in the initial results without data augmentation for the classification module. The performance metrics underwent significant changes after the dataset was expanded to 500 images. While some models’ accuracy declined, others’ precision, recall, and F1-score increased in some circumstances. For instance, B1 showed a significant improvement in classification precision, increasing by about 7.82% from 0.7814 to 0.8596, as shown in Fig. 6. Further advancements were seen across a number of metrics in the second experiment, where the dataset was expanded to 1364 images. Compared to the prior experiment, the models showed improved accuracy, precision, recall, and F1-score. Notably, B4’s accuracy increased significantly by about 9.2%, reaching 0.9034, as shown in Table 6. Additionally, B3’s precision increased by roughly 9.51% to 0.9480, demonstrating a significant improvement

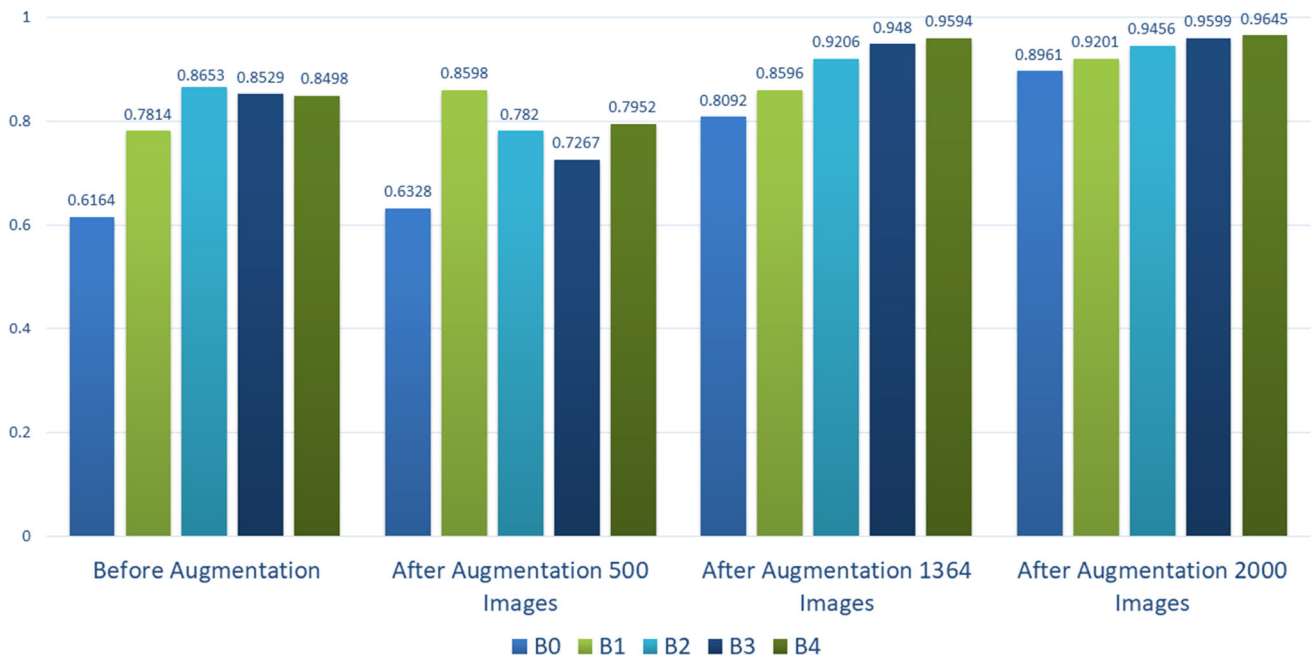


Fig. 6 EfficientNet models precision metric evaluation

in classification precision, as shown in Fig. 6. The models maintained a rise in performance in the third experiment, which increased the dataset to 2000 images. In comparison with the earlier experiments, the accuracy, precision, recall, and F1-score improved. From 0.9034 in the second experiment to 0.8458 in the third, B4’s accuracy did not increase, as shown in Tables 6 and 7. B2’s precision increased significantly as well, rising by about 2.5% to 0.9456, as shown in Fig. 6, demonstrating an improvement in classification precision. The overall effects of data augmentation on the classification performance of the models are shown by the results. Accuracy, precision, recall, and F1-score all increased as the dataset was expanded, demonstrating the models’ improved capacity for accurate prediction.

5.2.7 Ablation study 3: how majority voting enhances classification predictions

In all experiments, using the majority voting method in the classification module produced notable benefits. The ensemble learning approach outperformed the individual models, achieving an accuracy of 0.8316 without any data augmentation, as shown in Table 8. The ensemble attained values of 0.8784, 0.8945, and 0.8864, respectively, as shown in Fig. 7 for precision, recall, and F1-score, showing that this improvement was consistent across a number of metrics. These findings show how the various models perform individually and how the majority voting method can contribute to each model’s advantages. When the

dataset was expanded to 500 images, ensemble learning with majority voting maintained its significant advantages. As compared to the individual models, the accuracy increased to 0.7500, demonstrating an improvement in classification performance, as shown in Table 5. Additionally, the precision and recall rates were improved, as evidenced by precision, recall, and F1-score values of 0.8164, 0.8204, and 0.8184, respectively, as shown in Fig. 7. The majority voting method’s advantages were evident in its ability to take into account different points of view from different models, obtaining predictions that were more precise and reliable. As the dataset was expanded to 1364 and 2000 images, ensemble learning with majority voting consistently outperformed the individual models. In the experiment with 1364 augmented images, the ensemble attained an accuracy of 0.9425, demonstrating a substantial enhancement in performance, as shown in Table 6. Outstanding precision and recall rates could be seen by looking at the precision, recall, and F1-score values of 0.9878, 0.9503, and 0.9687, respectively, as shown in Fig. 7. The ensemble also achieved an accuracy of 0.8667 in the experiment with 2000 augmented images, demonstrating the continued benefit of majority voting, as shown in Table 7. The ensemble’s capacity for making accurate and reliable classifications was shown by the precision, recall, and F1-score values of 0.9757, 0.8647, and 0.9168, respectively, as shown in Fig. 7. In conclusion, the majority voting method outperformed the individual models in all experiments in terms of accuracy, precision, recall, and F1-score when compared to the use of

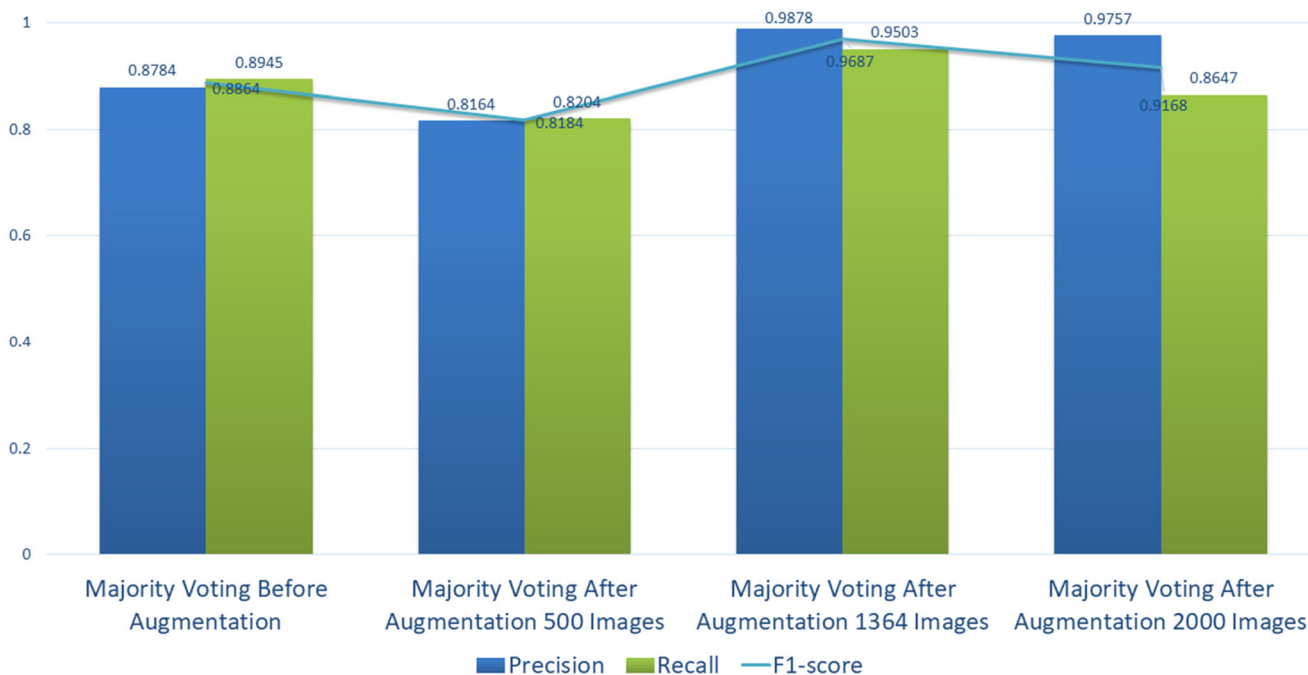


Fig. 7 Comparison of majority voting throughout various experiments

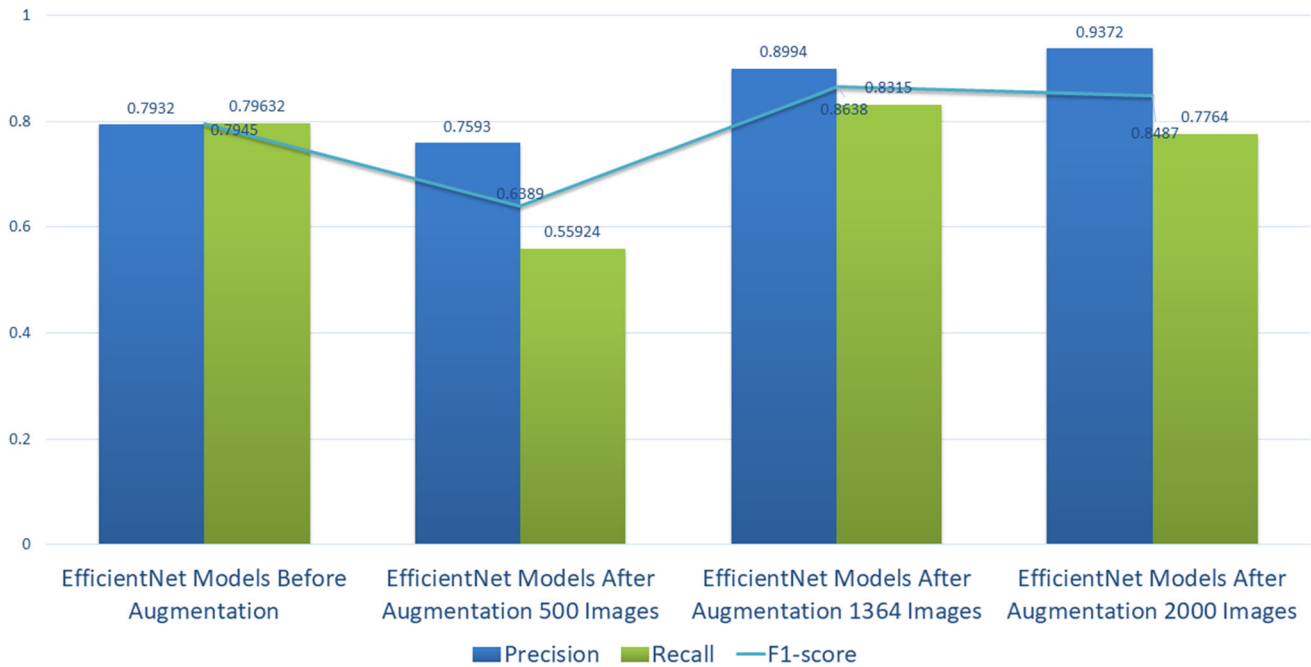


Fig. 8 Comparison of EfficientNet models for average precision, recall, and f1-score across multiple experiments, excluding majority voting

EfficientNet separately, as shown in Fig. 8. Comparing the previous results, the ensemble learning approach improved the performance of the classification module, producing more precise and reliable predictions by taking advantage of the collective decisions made by the individual models.

5.2.8 Ablation study 4: investigate the impact of using segmentation module prior to the classification operation

Prior to the classification process, the use of a segmentation module and the use of an outline contour to highlight polyps are crucial for correctly identifying and classifying polyps. The segmentation module is essential for clearly defining the polyp boundaries. Medical professionals and automated systems can benefit from enhanced classification performance and greater insight into polyp characteristics by successfully excluding and focusing on the polyp regions of interest. So, we chose the experiment with 1364 augmented images, which achieved the highest classification performance. The accuracy of the classification results

Table 14 Results of the ablation study experiment performed on 500 augmented images before and after using the segmentation module and EfficientNet

Model	Average accuracy
EfficientNet	0.5731
EfficientNet + segmentation module	0.6160

Table 15 Results of the ablation study experiment performed on 1364 augmented images before and after using the segmentation module and EfficientNet

Model	Average accuracy
EfficientNet	0.7522
EfficientNet + segmentation module	0.8151

Table 16 Results of the ablation study experiment performed on 2000 augmented images before and after using the segmentation module and EfficientNet

Model	Average accuracy
EfficientNet	0.7156
EfficientNet + segmentation module	0.7958

throughout the three experiments has significantly improved with the combination of a segmentation module with EfficientNet. On the 500 augmented images experiment, EfficientNet’s average accuracy was 0.5731 without the segmentation module. The average accuracy improved to 0.6160, as shown in Table 14, after the segmentation module was added, indicating a clear improvement in the model’s capacity to correctly classify the augmented images. Similar to this, EfficientNet achieved an average accuracy of 0.7522 in the 1364 augmented images experiment without the segmentation module. However, the average accuracy significantly increased to 0.8151, as

shown in Table 15, after the segmentation module was added. This significant improvement demonstrates how well the segmentation module works to improve EfficientNet’s classification performance. Last but not least, EfficientNet achieved an average accuracy of 0.7156 in the experiment with 2000 augmented images without the segmentation module. However, the segmentation module’s addition increased the average accuracy to 0.7958, as shown in Table 16, highlighting its beneficial influence on the model’s classification abilities, as shown in Tables 14, 15, and 16.

6 Graphical user interface (GUI) for the PolyDSS model

A GUI was developed to implement the PolyDSS and make it practical for clinical use, assisting medical professionals in segmenting and classifying polyps quickly and effectively. The user of the GUI does not need to be concerned about the complexity of the deep learning models because they can handle them as a black box, and for further information, the user can view a brief tutorial video illustrating how the system operates by hitting the help button. The upload Image button can be used to upload a patient image sample, as shown in Fig. 9, and the output will be two images: segmented and classified image. The following steps show the detailed sequence of the PolyDSS model:

- steps 1: By clicking the upload button and browsing for the file’s location, the user loads the patient’s image.

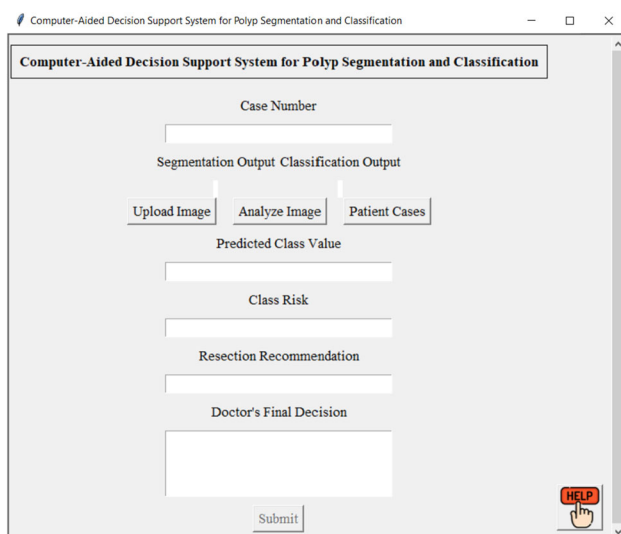


Fig. 9 GUI design to aid medical professionals in segmenting and classifying polyps

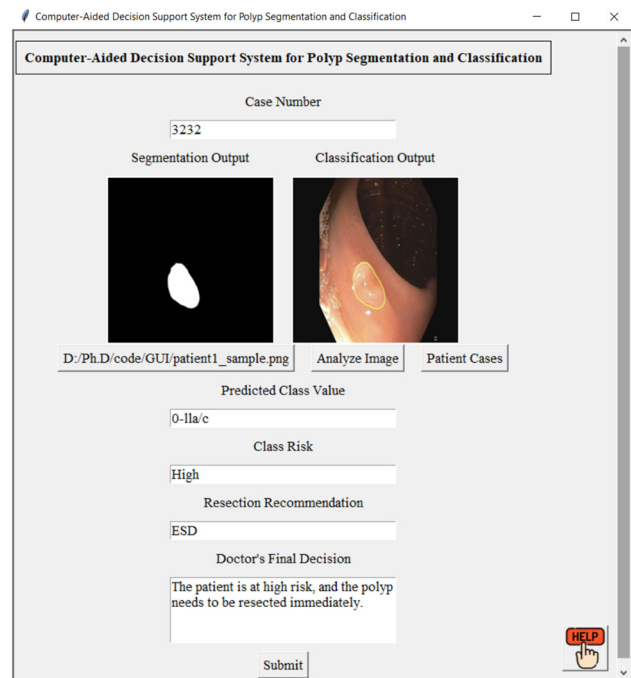


Fig. 10 GUI in action (output) where the segmented and classified image appears supplemented by the predicted class value, class risk, resection recommendation, and doctor’s final decision



Fig. 11 The final step that displays the user’s recorded decision on the final image

- steps 2: The input image is sent in sequence to the segmentation module and then to the classification module after pressing the analyze button.

- steps 3: Following model processing, the user is provided with four important pieces of information that are not editable: the case number (randomly generated by the system), predicted class, class risk, and resection recommendation, and a folder is created on the local disk with the patient case number containing the processed images.

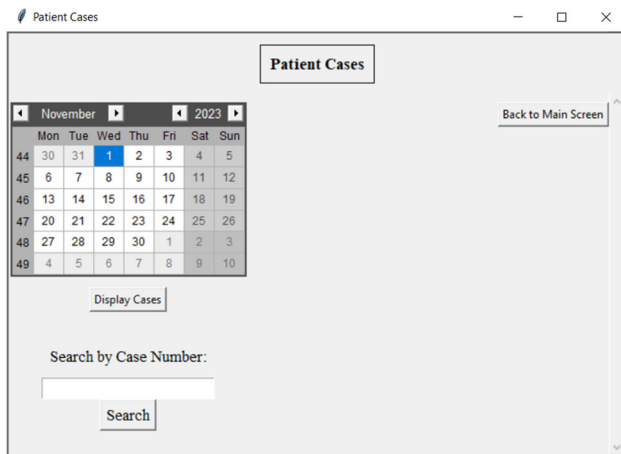


Fig. 12 Patient cases screen, from which the user can retrieve patient files after selecting a specific date or search by patient case number

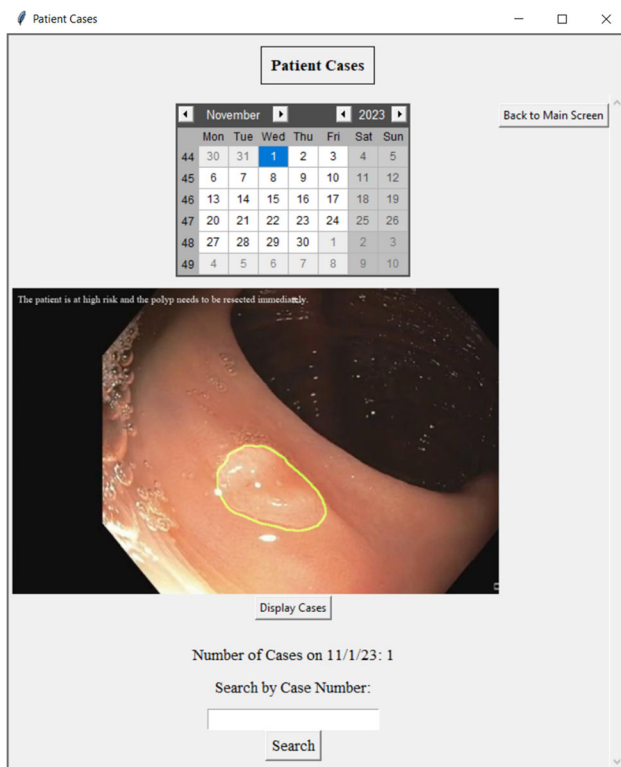


Fig. 13 Patient cases screen output showing the retrieved case after selecting a specific date

- steps 4: The user takes the system's recommendations into account before making his final decision.
- steps 5: The final decision is then recorded on the output image and saved in the patient's file, as shown in Figure 11, and the complete GUI is shown in Figs. 9 and 10.
- steps 6: Finally, the user can check the patient history by going to the patient cases screen and

selecting a certain date to see all cases on that date, or by searching for a specific case using the patient case number as shown in Figs. 12 and 13.

6.1 GUI evaluation

GUI evaluation is essential in the healthcare field, particularly in computer-aided decision support systems like the one under consideration. We conducted a thorough evaluation of the Graphical User Interface (GUI) developed for PolyDSS model in this study. The GUI is an important component of the system, allowing user interaction and affecting the system's overall effectiveness in clinical practice.

6.1.1 Survey structure and respondent profile

The survey was sent to a varied number of medical specialists with specialized knowledge, including consultant gastroenterologists and hepatologists, internist endoscopists, and prominent professors of medicine and surgery. This diversity guaranteed that the feedback collected represented a wide range of opinions from people with different levels of medical and technological knowledge. The survey framework addresses a variety of issues concerning GUI usability and effectiveness, such as ease of use, satisfaction with functionality, and areas for sustainable development and improvement. For the purpose of collecting thorough input from respondents, the survey combines both quantitative (eight select one question) and qualitative (two open-ended questions) methodologies.

6.1.2 Survey results and discussions

The survey was prepared with Google Forms and distributed electronically to respondents. The survey results were analyzed, and a summary of the findings is shown in Figs. 14 and 15.

Figure 14, sub-figures (a) and (b), show the demographics of our participants in the survey, which include gastroenterologists, hepatologists, internists, and surgeons, as well as their years of experience. These data are critical in contextualizing the comments we received on our GUI. Notably, we found a broad range of responses, with 50% identified as consultant gastroenterologists, 30% as internist endoscopists, and a smaller percentage as hepatologists and professors of medicine and surgery. This range of professional backgrounds increases both the depth and breadth of our survey's insights, guaranteeing that the feedback we obtained encompasses perspectives from different fields within the health care sector. The distribution

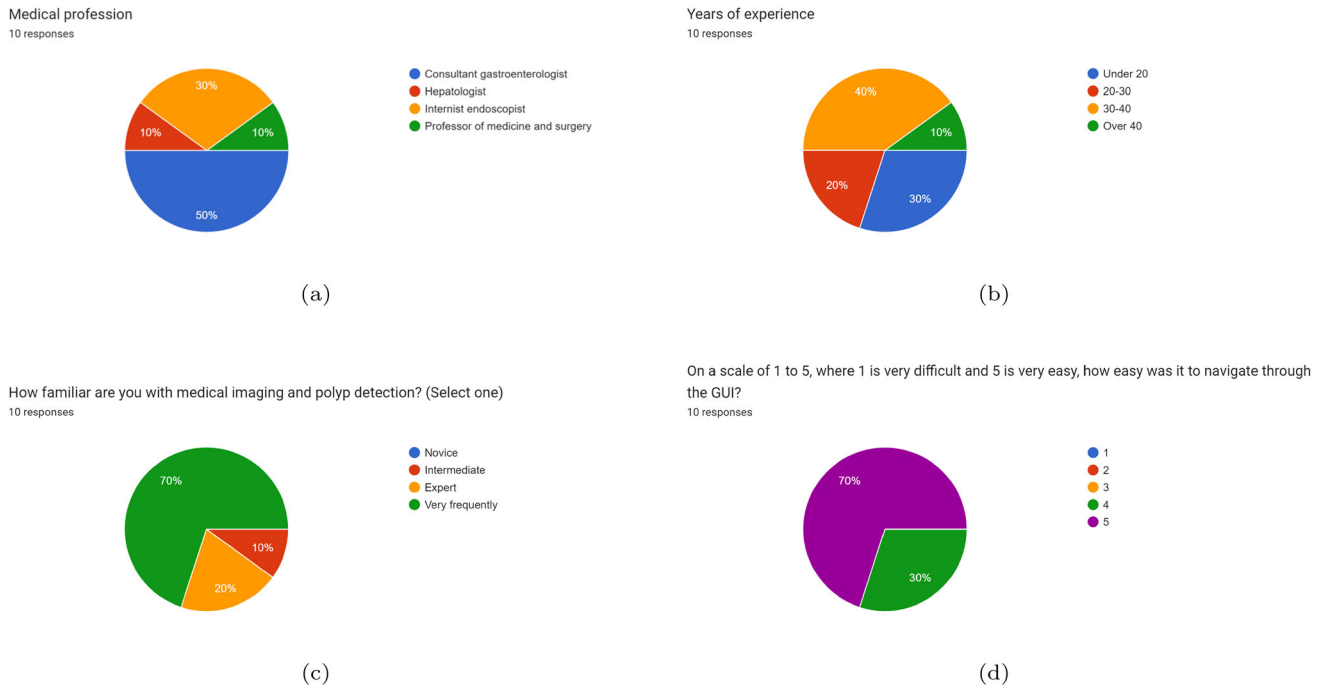


Fig. 14 a Respondents medical profession b Respondents years of experience c Polyp detection familiarity d GUI navigation

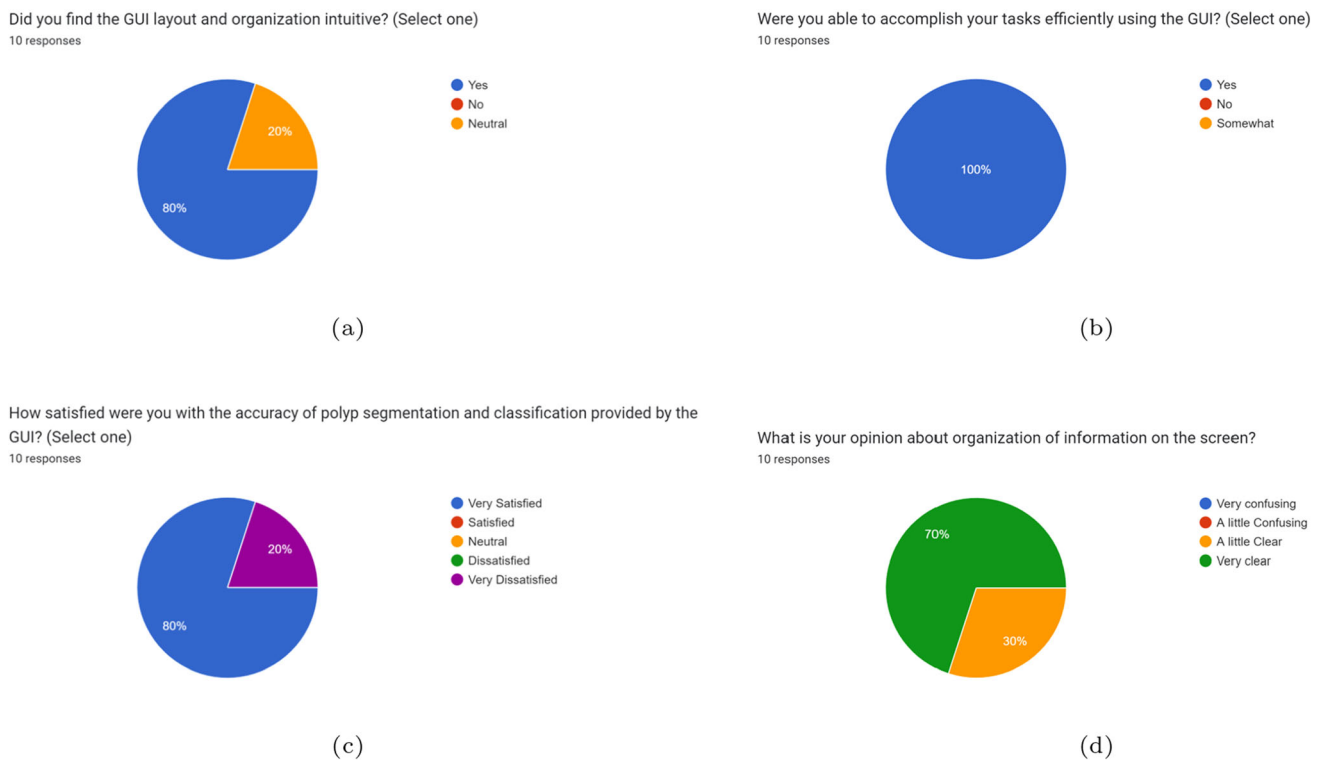


Fig. 15 a GUI and layout organization b Task efficiency c Polyp segmentation and classification accuracy satisfaction d Information organization

of years of experience among our respondents is also worth mentioning. It displays an evenly distributed blend, with 30% having less than 20 years of experience, 20% having

between 20 and 30 years, 40% having between 30 and 40 years, and 10% having more than 40 years of experience. This distribution emphasizes the inclusiveness of our study,

which includes the perspectives of experienced specialists as well as individuals who are new to their respective areas of study.

Moving on to Fig. 14, sub-figures (c) and (d) demonstrate our respondents' familiarity with polyp identification as well as their experiences with GUI navigation. An impressive 70% of respondents said they were quite familiar with polyp identification, highlighting the utility and relevance of our GUI in their professional setting. Furthermore, 70% of respondents gave the GUI navigation a flawless 5 rating, indicating an exceptional level of user-friendliness and effective system interaction.

Sub-figures (a) and (b) of Fig. 15 investigate respondents' satisfaction with GUI layout and organization, as well as their efficiency in accomplishing medical tasks using the system. 80% of users were highly impressed with the structure and organization of the GUI, emphasizing its intuitiveness and effectiveness. Also, 100% of respondents indicated good task completion efficiency, demonstrating that our GUI successfully supports users' decision-making processes.

Finally, Fig. 15 including sub-figures (c) and (d) provide information about the accuracy of polyp segmentation and classification, as well as the readability of information presented on the screen. A remarkable 80% of respondents indicated great satisfaction with the GUI's polyp segmentation and classification accuracy, confirming its potential to improve clinical decision assistance. Furthermore, 70% believed the arrangement of information on the screen was extremely obvious, while 30% believed that it was barely so, indicating space for additional improvement in information display in the future. Based on the findings of the two open-ended questions, respondents advised that offering a basic tutorial video for first-time users would be advantageous, as well as considering enlarging the displayed information on screen and accessing the history of the patient cases. All the previously mentioned findings were considered and the GUI developed carefully and iteratively.

In conclusion, our respondents' diversified backgrounds and considerable experience, combined with their high familiarity with polyp identification and positive feedback on GUI usability and accuracy, highlight the promising potential of our computer-aided decision support system. These findings emphasize the necessity of taking into account user perspectives and needs, proving the relevance of comprehensive GUI evaluation in the creation of medical technologies that meet the specific needs of healthcare professionals.

7 Limitations and future work

In this section, we explore the specific limitations found during the course of this research and present a plan for further investigation.

7.1 Limitations

We have discussed our suggested model's design, implementation, and evaluation in the sections above, emphasizing its promising performance and prospective applications. Even the most promising research initiatives have inherent limitations and potential for improvement, which must be acknowledged. As indicated earlier in Table 4 where it is observed that most contemporary models contain a significant number of layers and parameters, one constraint of our current technique is that it may not fully harness the benefits of even larger models. The B5, B6, and B7 models could not be investigated due to computational limitations (GPU memory). Another data restriction that PolyDSS model shares with many others, is the availability and quality of training data, which may have an impact on how well the proposed deep learning models perform. For the deep learning models used in the study, there are not many training samples available, hence the dataset (PICCLO) suffers from data scarcity. Additionally, there is a class imbalance in the dataset, with more members of the minority classes than those from the majority class. Furthermore, we focused on one ensemble learning strategy, such as majority voting.

7.2 Future work

Consider larger model versions like the B5, B6, and B7; as a first limitation, we were unable to try them due to a lack of computational resources. In addition to optimizing our training pipeline, our future plans to address the mentioned limitation include applying model pruning approaches, which entail removing less important weights and connections from the neural network in order to reduce its size without significantly degrading performance. Another strategy is to take into account distributed or cloud-based computing platforms. The on-demand availability of GPU and TPU resources from cloud providers is common and can be particularly helpful for training larger models like Amazon SageMaker. The second limitation was the small number of training samples, which was handled in our study by using several augmentation techniques. In the future, we would like to work with medical institutions to get more diverse and high-quality datasets to help with model generalization enhancement. Finally, we will

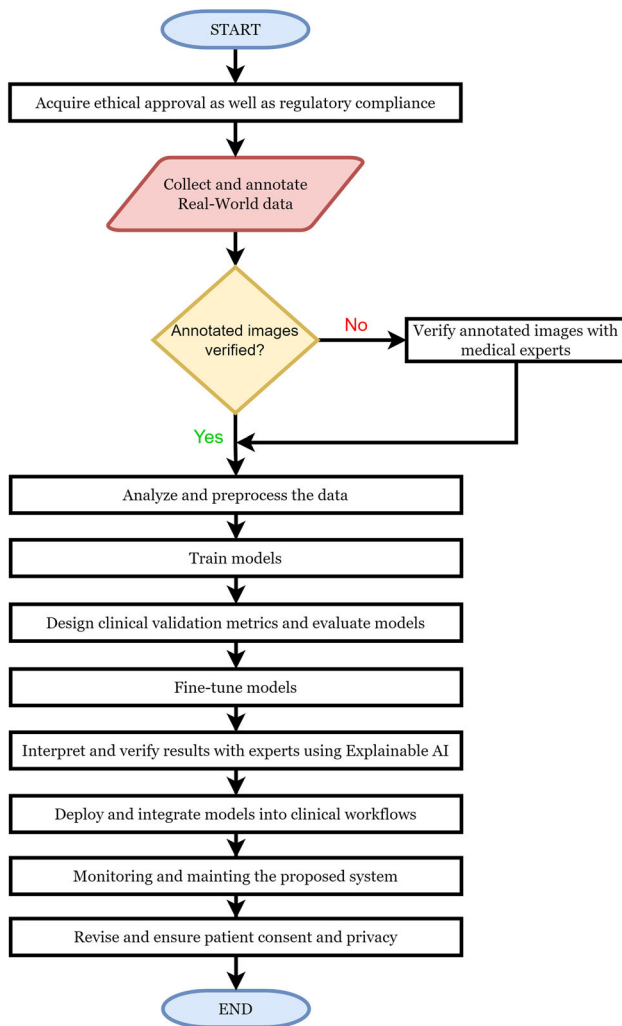


Fig. 16 Flowchart of clinical validation steps

investigate more ensemble learning techniques, such as weighted average and bootstrap aggregation.

7.2.1 Clinical validation plan

The PolyDSS model uses deep learning to increase the accuracy and efficiency of polyp segmentation and classification in endoscopic images, which is an important part of early gastrointestinal disease diagnosis. However, the conversion of this type of technology from research to clinical practice demands a thorough clinical validation plan. This plan is critical for ensuring the model's safety and usefulness in real-world medical scenarios. A detailed plan for clinical validation is shown in Fig. 16. Clinical validation is a critical step in determining the practical usability of a polyp segmentation and classification model in medical contexts. It starts with gaining ethical approvals and compiling a broad, representative dataset of endoscopic images and patient data that has been thoroughly

annotated by medical specialists. The model's performance is evaluated using conventional metrics after data preparation and model training. After that, the clinical utility of the model is improved via an iterative feedback loop that includes fine-tuning based on expert input, and then clinical specialists actively contribute, providing domain-specific insights and assessing model predictions with the help of Explainable AI, such as heat and saliency maps, which make the model's decisions understandable to medical professionals. Once a model has been verified, it is integrated into clinical workflows (ex. the Olympus CV290 machine) with continuous monitoring and updates, as well as full compliance with patient consent and privacy rules.

8 Ethical considerations

In this study, we have placed considerable focus on principles of ethics that support the proper use of data and models. In this section, critical ethical considerations are highlighted regarding data privacy, informed consent, and possible biases in our dataset and model.

8.1 Data privacy

Data privacy is critical in the digital age. This research relied on various types of data sources, and we took precautions to ensure that the data was anonymized. Any personally identifiable information that might jeopardize the privacy of individuals in the dataset is deleted.

8.2 Informed consent

In situations in which data gathering concerned human individuals, we sought informed consent while following standard ethical principles. Participants in the survey were sent a document detailing the goal of the study, as well as a link to the survey. This document offered information on the study's nature, focused on GUI evaluation in accordance with the study's objectives. All parties' informed consent are valued and prioritized transparency and ethical data handling throughout the research process in the following sect. 8.2.1.

8.2.1 Survey on GUI evaluation

The survey was conducted to evaluate the Graphical User Interface (GUI) of the PolyDSS medical system. This survey's ethical considerations included:

- Ensuring that participants submitted their informed agreement to participate in the GUI evaluation survey by signing at the end of the survey using their names.

Before taking part in the survey, participants were told of its objectives and willingly gave their consent.

- Protecting survey respondents' privacy by masking data collected throughout the survey. To protect survey participants' privacy, all replies were treated discreetly, and personally recognizable data was removed.
- Following ethical guidelines to guarantee that the survey serves respondents while avoiding harm or anguish. Steps were taken to ensure that the survey was performed in a respectful and minimally disruptive manner, putting survey participants' psychological well-being first throughout the review process.

8.3 Potential biases

Considering bias in data and models is a serious ethical issue, we admitted that the number of image samples in the PICCLO dataset differs across classes. Such disparities in class distribution might result in model bias, as classes with more samples may dominate minority classes. This issue may result in biased model performance and decision-making, particularly in applications where all classes are of similar relevance. To reduce the possibility of bias, we used a variety of approaches, including data augmentation and conducting several experiments. Data augmentation techniques were used to augment the minority classes, balancing the dataset. In addition, a number of experiments were performed to assess the model's performance under various class distributions, confirming that the model's predictions were both accurate and impartial across all classes. These steps were taken to mitigate potential bias and ensure that all classes in the dataset and the model's predictions were fairly represented.

9 Conclusion

In this paper, a model is introduced to segment and classify colorectal polyps. Colorectal cancer can spread less quickly if polyps are identified and classified correctly. Segmentation and classification are the two main components of the suggested model. The ResUNet and ResUNet++ models use a segmentation module followed by a classification module to perform precise and reliable image analysis. Prior to the classification step, segmentation performance is improved using a locally shared feature mechanism. The output of the segmentation module is then passed to the classification module, which uses five versions of EfficientNet (B0, B1, B2, B3, and B4) to classify images into six main polyp classes (0-Ip, 0-Ips, 0-Is, 0-IIa, 0-IIa/c, and 0-IIb). According to experimental findings, our suggested model successfully segmented polyps using

ResUNet++ and LSF, with DSC, precision, and recall scores of 0.9244, 0.8839, and 0.8827. Experiments were conducted using the PICCLO dataset, which has a noticeable class imbalance. To address the class imbalance, data augmentation techniques were used. After data augmentation, results significantly improved, and the ResUNet++ + LSF model's DSC, precision, and recall increased from 0.8071 to 0.9244, 0.7836 to 0.8839, and 0.7694 to 0.8827, respectively. B4 had the highest accuracy and f1-score for classifying polyps, emerging at 0.9034 and 0.9437, respectively. EfficientNet models were made more reliable and efficient by combining the strengths and feature representations of each model when ensemble learning with majority voting was used. By using ensemble learning majority voting, the 1364 augmented image level yielded the highest accuracy, precision, recall, and f1-score values of 0.9425, 0.9878, 0.9503, and 0.9687, respectively. We believe that the suggested model and positive findings from the experiments will assist endoscopists in classifying polyps and provide recommendations in the case of resection using EMR or ESD in accordance with Paris classification guidelines, besides developing an easy-to-use GUI that puts the model into practice in the field of coloscopy.

Author contributions AIS formulated the paper's main ideas, conducted the experiments, and drafted the manuscript. FA and OB participated in completing the analysis. All authors contributed to the research study. All authors reviewed the findings and approved the final manuscript version.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB). The Science, Technology, and Innovation Funding Authority (STDF) provided open access funding in collaboration with The Egyptian Knowledge Bank (EKB).

Data availability The data that support the findings of this study are available with permission from Basque Biobank at <https://www.bio.bancovasco.bioef.eus/en/Sample-and-data-catalog/Databases/PD178-PICCOLO-EN.html>.

Declarations

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this work.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted

use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Siegel RL, Miller KD, Wagle NS, Jemal A (2023) Cancer statistics, 2023. *CA Cancer J Clin* 73(1):17–48. <https://doi.org/10.3322/caac.21763>
- Castelo M, Paszat L, Hansen BE, Scheer AS, Faught N, Nguyen L, Baxter NN (2023) Comparing time to diagnosis and treatment between younger and older adults with colorectal cancer: a population-based study. *Gastroenterology*. <https://doi.org/10.1053/j.gastro.2023.02.024>
- Kara OC, Venkatayogi N, Ikoma N, Alambeigi F (2023) A reliable and sensitive framework for simultaneous type and stage detection of colorectal cancer polyps. *Ann Biomed Eng*. <https://doi.org/10.1007/s10439-023-03153-w>
- Senthakumaran T, Moen AEF, Tannæs TM, Endres A, Brackmann SA, Rounge TB, Bemanian V, Tunsjø HS (2023) Microbial dynamics with CRC progression: a study of the mucosal microbiota at multiple sites in cancers, adenomatous polyps, and healthy controls. *Eur J Clin Microbiol Infect Dis* 42(3):305–322. <https://doi.org/10.1007/s10096-023-04551-7>
- Singh LK, Khanna M, Singh R (2023) Artificial intelligence based medical decision support system for early and accurate breast cancer prediction. *Adv Eng Softw* 175:103338. <https://doi.org/10.1016/j.advengsoft.2022.103338>
- Markus T, Saban M, Sosna J, Assaf J, Cohen D, Vaknin S, Luxenburg O, Singer C, Shaham D (2023) Does clinical decision support system promote expert consensus for appropriate imaging referrals? chest-abdominal-pelvis CT as a case study. *Insights Imag* 14(1):45. <https://doi.org/10.1186/s13244-023-01371-3>
- Puyal JG-B, Brandao P, Ahmad OF, Bhatia KK, Toth D, Kader R, Lovat L, Mountney P, Stoyanov D (2023) Spatio-temporal classification for polyp diagnosis. *Biomed Opt Express* 14(2):593–607. <https://doi.org/10.1364/BOE.473446>
- Mansour NM (2023) Artificial intelligence in colonoscopy. *Curr Gastroenterol Rep*. <https://doi.org/10.1007/s11894-023-00872-x>
- Volkan G (2023) A novel deep-learning-based cadx architecture for classification of thyroid nodules using ultrasound images. *Comput Life Sci Interdiscip Sci*. <https://doi.org/10.1007/s12539-023-00560-4>
- Hanevelt J, Moons LM, Hentzen JE, Wemeijer TM, Huisman JF, de Nederveen Cappel WH, van Westreenen HL (2023) Colonoscopy-assisted laparoscopic wedge resection for the treatment of suspected t1 colon cancer. *Ann Surg Oncol*. <https://doi.org/10.1245/s10434-022-12973-4>
- Pedro Celard EL, Iglesias JMS-F, Rubén Romero A, Vieira S, Borrajo L (2023) A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. *Neural Comput Appl* 35(3):2291–2323. <https://doi.org/10.1007/s00521-022-07953-4>
- Eldem H, Ülker E, Işıklı OY (2023) Alexnet architecture variations with transfer learning for classification of wound images. *Eng Sci Technol Int J* 45:101490. <https://doi.org/10.1016/j.jestch.2023.101490>
- Park CH, Yang DH, Kim JW, Kim JH, Min YW, Yi SH, Bae JH, Jung HS, Choi KD, Park JC et al (2020) Clinical practice guideline for endoscopic resection of early gastrointestinal cancer. *Korean J Gastroenterol* 75(5):264–291. <https://doi.org/10.4166/kjg.2020.75.5.264>
- Saraiva S, Rosa I, Fonseca R, Pereira AD (2022) Colorectal malignant polyps: a modern approach. *Ann Gastroenterol* 35(1):17. <https://doi.org/10.20524/aog.2021.0681>
- Endoscopic Classification Review Group (2005) Update on the Paris classification of superficial neoplastic lesions in the digestive tract. *Endoscopy* 37(06):570–578. <https://doi.org/10.1055/s-2005-861352>
- Ribeiro H, Libânio D, Castro R, Ferreira A, Barreiro P, Carvalho PB, Capela T, Pimentel-Nunes P, Santos C, Dinis-Ribeiro M (2019) Reliability of Paris classification for superficial neoplastic gastric lesions improves with training and narrow band imaging. *Endosc Int Open* 7(05):E633–E640. <https://doi.org/10.1055/a-0828-7541>
- Fujiyoshi MRA, Inoue H, Fujiyoshi Y, Nishikawa Y, Toshimori A, Shimamura Y, Tanabe M, Ikeda H, Onimaru M (2022) Endoscopic classifications of early gastric cancer: a literature review. *Cancers* 14(1):100. <https://doi.org/10.3390/cancers14010100>
- Marginean EC, Dhanpat J (2020) Pathologic assessment of endoscopic resection specimens with superficial carcinoma of the esophagus: current practice and practical issues. *Ann N Y Acad Sci* 1482(1):130–145. <https://doi.org/10.1111/nyas.14455>
- Hsu C-M, Hsu C-C, Hsu Z-M, Shih F-Y, Chang M-L, Chen T-H (2021) Colorectal polyp image detection and classification through grayscale images and deep learning. *Sensors* 21(18):5995. <https://doi.org/10.3390/s21185995>
- Jorge Bernal F, Javier S, Gloria F-E, Debora G, Cristina R, Fernando V (2015) WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. *Comput Med Imaging Graph* 43:99–111. <https://doi.org/10.1016/j.compmedimag.2015.02.007>
- Lo C-M, Yeh Y-H, Tang J-H, Chang C-C, Yeh H-J (2022) Rapid polyp classification in colonoscopy using textural and convolutional features. *Healthcare* 10:1494. <https://doi.org/10.3390/healthcare10081494>
- Krenzer A, Heil S, Fitting D, Matti S, Zoller WG, Hann A, Puppe F (2023) Automated classification of polyps using deep learning architectures and few-shot learning. *BMC Med Imaging* 23(1):59. <https://doi.org/10.21203/rs.3.rs-2106189/v1>
- Misawa M, Kudo S, Mori Y, Hotta K, Ohtsuka K, Matsuda T, Saito S, Kudo T, Baba T, Ishida F et al (2021) Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointest Endosc* 93(4):960–967. <https://doi.org/10.1016/j.gie.2020.07.060>
- Krenzer A, Banck M, Makowski K, Hekalo A, Fitting D, Troya J, Sudarevic B, Zoller WG, Hann A, Puppe F (2023) A real-time polyp-detection system with clinical application in colonoscopy using deep convolutional neural networks. *J Imaging* 9(2):26. <https://doi.org/10.21203/rs.3.rs-1310139/v1>
- Yue G, Wei P, Yun Liu Yu, Du Luo J, Wang T (2023) Automated endoscopic image classification via deep neural network with class imbalance loss. *IEEE Trans Instrum Meas* 72:1–11. <https://doi.org/10.1109/tim.2023.3264047>
- Borgli H, Thambawita V, Smedsrud PH, Hicks S, Jha D, Eskeland SL, Randel KR, Pogorelov K, Lux M, Nguyen DTD et al (2020) Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci Data* 7(1):283. <https://doi.org/10.31219/osf.io/mkzqc>
- Shen M-H, Huang C-C, Chen Y-T, Tsai Y-J, Liou F-M, Chang S-C, Phan NN (2023) Deep learning empowers endoscopic detection and polyps classification: a multiple-hospital study. *Diagnostics* 13(8):1473. <https://doi.org/10.3390/diagnostics13081473>
- Pogorelov K, Randel KR, Griwodz C, Eskeland SL, de Lange T, Johansen D, Spampinato C, Dang-Nguyen DT, Lux M, Schmidt

- PT, et al (2017), Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on multimedia systems conference, pp. 164–169, <https://doi.org/10.1145/3193289>
29. Lewis J, Cha Y-J, Kim J (2023) Dual encoder-decoder-based deep polyp segmentation network for colonoscopy images. *Sci Rep* 13(1):1183. <https://doi.org/10.1038/s41598-023-28530-2>
 30. Jha D, Smedsrud PH, Riegler MA, Halvorsen P, de Lange T, Johansen D, Johansen HD (2020) Kvasir-seg: a segmented polyp dataset. In: MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26, pp. 451–462, Springer, https://doi.org/10.1007/978-3-030-37734-2_37
 31. Tajbakhsh N, Gurudu SR, Liang J (2015) Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans Med Imaging* 35(2):630–644. <https://doi.org/10.1109/tmi.2015.2487997>
 32. Silva J, Histace A, Romain O, Dray X, Granado B (2014) Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Int J Comput Assist Radiol Surg* 9:283–293. <https://doi.org/10.1007/s11548-013-0926-3>
 33. David V, Jorge Bernal F, Javier S, Gloria F-E, López Antonio M, Adriana R, Michal D, Aaron C (2017) A benchmark for endoluminal scene segmentation of colonoscopy images. *J Healthc Eng.* <https://doi.org/10.1155/2017/4037190>
 34. Zhu J, Ge M, Chang Z, Dong W (2023) Crnet: global-local context and multi-modality cross attention for polyp segmentation. *Biomed Signal Process Control* 83:104593. <https://doi.org/10.1016/j.bspc.2023.104593>
 35. Sánchez-Peralta LF, Blas Pagador J, Picón A, Calderón AJ, Polo F, Andracka N, Bilbao R, Glover B, Saratxaga CL, Sánchez-Margallo FM (2020) Piccolo white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets. *Appl Sci* 10(23):8501. <https://doi.org/10.3390/app10238501>
 36. El Karazle K, Raman V, Then P, Chua C (2023) Detection of colorectal polyps from colonoscopy using machine learning: a survey on modern techniques. *Sensors* 23(3):1225. <https://doi.org/10.3390/s23031225>
 37. Zhang Z, Liu Q, Wang Y (2018) Road extraction by deep residual u-net. *IEEE Geosci Remote Sens Lett* 15(5):749–753. <https://doi.org/10.1109/lgrs.2018.2802944>
 38. Jha D, Smedsrud PH, Riegler MA, Johansen D, De Lange T, Halvorsen P, Johansen HD (2019) Resunet++: an advanced architecture for medical image segmentation. In: 2019 IEEE international symposium on multimedia (ISM), pp. 225–2255, IEEE, <https://doi.org/10.1109/ism46123.2019.00049>
 39. Zhan B, Song E, Liu H, Xiangyang X, Li W, Hung C-C (2023) Segmenting medical images via explicit-implicit attention aggregation. *Knowl-Based Syst* 279:110932. <https://doi.org/10.1016/j.knosys.2023.110932>
 40. Junhu F, Lin S, Zhou P, Guo Y, Wang Y (2022) M3resu-net: a deep residual network for multi-center colorectal polyp segmentation based on multi-scale learning and attention mechanism. *Phys Med Biol* 67(20):205005. <https://doi.org/10.1088/1361-6560/ac92bb>
 41. Salpea N, Tzouveli P, Kollias D (2023) Medical image segmentation: a review of modern architectures. In: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII, pp. 691–708, Springer, <https://doi.org/10.5220/0010851300003123>
 42. Hung NB, Duc NT, Van Chien T, Sang DV (2021) Ag-resunet++: an improved encoder-decoder based method for polyp segmentation in colonoscopy images. In: 2021 RIVF international conference on computing and communication technologies (RIVF), pp. 1–6, IEEE, <https://doi.org/10.1109/rivf51545.2021.9642070>
 43. Kang J, Gwak J (2020) KD-ResUNet++: automatic polyp segmentation via self-knowledge distillation. In: *MediaEval*. <https://ceur-ws.org/Vol-2882/paper74.pdf>
 44. Yang Z, Hongshan Yu, Sun W, Mao Z, Sun M (2018) Locally shared features: an efficient alternative to conditional random field for semantic segmentation. *IEEE Access* 7:2263–2272. <https://doi.org/10.1109/access.2018.2886524>
 45. Koonce B, Koonce B (2021) Efficientnet, convolutional neural networks with swift for tensorflow: image recognition and dataset categorization, pp. 109–123, https://doi.org/10.1007/978-1-4842-6168-2_10
 46. Nayak DR, Padhy N, Mallick PK, Zymbler M, Kumar S (2022) Brain tumor classification using dense efficient-net. *Axioms* 11(1):34. <https://doi.org/10.3390/axioms11010034>
 47. Wang J, Liu Q, Xie H, Yang Z, Zhou H (2021) Boosted efficientnet: detection of lymph node metastases in breast cancer using convolutional neural networks. *Cancers* 13(4):661. <https://doi.org/10.3390/cancers13040661>
 48. Marques G, Agarwal D, De la Torre Díez I (2020) Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. *Appl Soft Comput* 96:106691
 49. Zhou Z-H, Zhou Z-H (2021) Ensemble learning. Springer, Berlin. https://doi.org/10.1007/978-0-387-73003-5_293
 50. Dong X, Zhiwen Yu, Cao W, Shi Y, Ma Q (2020) A survey on ensemble learning. *Front Comp Sci* 14:241–258. <https://doi.org/10.1007/s11704-019-8208-z>
 51. Omer Sagi and Lior Rokach (2018) Ensemble learning: a survey, Wiley interdisciplinary reviews. *Data Min Knowl Disc* 8(4):e1249. <https://doi.org/10.1002/widm.1249>
 52. Chehade AH, Abdallah N, Marion J-M, Oueidat M, Chauvet P (2022) Lung and colon cancer classification using medical imaging: a feature engineering approach. *Phys Eng Sci Med* 45(3):729–746. <https://doi.org/10.21203/rs.3.rs-1211832/v1>
 53. Rahmani AM, Azhir E, Naserbakht M, Mohammadi M, Aldalwie AH, Majeed MK, Taher Karim SH, Hosseinzadeh (2022) Automatic covid-19 detection mechanisms and approaches from medical images: a systematic review. *Multimed Tools Appl* 81(20):28779–28798. <https://doi.org/10.1007/s11042-022-12952-7>
 54. Mohammed A, Kora R (2023) A comprehensive review on ensemble deep learning: opportunities and challenges. *J King Saud Univ Comput Inf Sci.* <https://doi.org/10.1016/j.jksuci.2023.01.014>
 55. Alzubaidi L, Fadhel MA, Al-Shamma O, Zhang J, Santamaría J, Duan Y (2022) Robust application of new deep learning tools: an experimental study in medical imaging. *Multimed Tools Appl.* <https://doi.org/10.1007/s11042-021-10942-9>
 56. Huang Q-X, Lin G-S, Sun H-M (2023) Classification of polyps in endoscopic images using self-supervised structured learning. *IEEE Access.* <https://doi.org/10.1109/ACCESS.2023.3277029>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.