Polyps Detection in Colonoscopy Images Based on a Combination of Deep Feature Extraction and Ensemble Learning Techniques

Win Sheng Liew, Tong Boon Tang, and Cheng-Kai Lu*

Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia
*chengkai.lu@utp.edu.my

ABSTRACT

Colorectal cancer (CRC) mostly arises from adenomatous polyps, and early detection and endoscopic resection of lesions have been claimed to be effective in decreasing the incidence of CRC and its mortality rate. Nonetheless, the current visual assessments for polyp detection are time-consuming and uncertain due to the subjectivity of individual evaluation. In this paper, a novel approach to automatically detect colonic polyps is proposed. The methodology exploits image thresholding, median filter, and normalization techniques to keep interference to a minimum. A new combination of several techniques, namely a residual learning framework as a feature extractor, principal component analysis for feature reduction, and ensemble learning for optimizing a predictive algorithm, is used to detect the colonic polyps. We merged three publicly available databases, Kvasir, ETIS-LaribPolypDB, and CVC-ClinicDB, to train and evaluate the proposed model, which includes images with and without polyps. With the combination of deep learning and ensemble learning, we developed a supervised machine learning solution for detecting the colonic polyps in colonoscopy images automatically with a low error rate of 0.92%. Our method achieved an accuracy level of 99.08%, with an F1 score, sensitivity, and specificity of 99.08%, 99.41%, and 98.75%, respectively, which significantly outperform the existing works.

Keywords: Colorectal cancer (CRC), polyp, principal component analysis, residual learning framework, deep learning, ensemble learning, colonoscopy

Introduction

Colorectal cancer (CRC) has become the second most common malignancy and cancer-related deaths in the United States[1,2]. According to Colorectal Cancer Statistics 2020, it reported that the population of CRC patient in older age groups has a declining incidence, but in contrast there was an increasing incidence in younger individuals[2]. Fortunately, screening early to detect CRC can significantly reduce the incidence of CRC and its mortality rate. On the other hand, computer-aided diagnosis (CAD) has been widely used in screening, medical diagnosis, and therapeutic systems for various cancer diseases including CRC over the past decade. By using CAD technology, radiologists can focus on smaller sub-volumes rather than on the entire volume, which significantly helps doctors make accurate decisions regarding the removal of polyps at an early stage, which in turn benefits the curative interventions[3-5]. Artificial Intelligence (AI) is considered as a machine emulation of human thinking processes, which can make the system become intelligent, able to learn, and self-organizing[6]. Recently, the use of AI techniques, such as deep learning (DL), in CAD has had a considerable impact on the interpretation of medical images, helping radiologists carry out diagnoses by acting as a second reader in detecting cancer diseases.

Generally, the methods of building an image classification model are divided into three parts: pre-processing, feature extraction, and classification. All the images need to undergo image pre-processing and normalization to reduce image degradation and resize the image's dimensions. Apart from that, every image has different features such as colour, density, shape, and texture[7]. Therefore, the features extracted from the dataset can be trained through supervised machine learning (ML) algorithms to learn the classification part. However, the feature extraction part still needs expert engineering support from a human being[8]. In the recent past, DL has been widely utilized in the image and video domains, because it can learn the entire model, which has two parts (feature extraction and classification), by using convolutional neural networks (CNNs), unlike the early conventional handcrafted methods[9,10]. In DL, CNN is a compelling ML technique and is part of deep neural networks. Besides, CNN has recently shown remarkable results in polyp detection[10-26]. In 2015, CNN features outperformed handcrafted features in a polyp detection challenge[11], because CNNs can learn rich feature representations automatically from the large numbers of diverse images and carry out the detection task[10].

In this direction, the present study attempts to investigate for the first time the performance of a combination of CNN with an ML algorithm in recognition of abnormalities, that is, the occurrence of polyps in endoscopy images. In ML, ensemble learning is mainly utilized to enhance a classifier's efficiency[27]. It combines multiple learning algorithms/classifiers

to categorize new samples to obtain a better predictive accuracy[28]. The most common types of ensemble techniques include bagging, boosting, stacking, and voting. The techniques of bagging and boosting can improve the accuracy of classification vastly by combining weak classifiers to improve the overall performance[27]. Furthermore, in this research, three endoscopic datasets have been merged, where the first dataset is the Kvasir dataset[18], the second dataset is ETIS-LaribPolypDB[29], and the third dataset is CVC-ClinicDB[30]. The classification is assessed based on several performance evaluation criteria, namely accuracy, error rate, F1-score, sensitivity, specificity, and area under the receiver operating characteristic (AUROC) curve.

The ResNet-50 is applied as a feature extractor in this work to extract features before applying principal component analysis (PCA) and AdaBoost ensemble learning. The remainder of the paper is organized as follows: Section of "Related work" presents an overview of endoscopy and discusses the use of AI for the analysis of endoscopy images. Section of "Methods" provides details of the methodology employed. Sections of "Results" and "Discussion" present the results obtained and the discussion. Finally, the conclusion of the study is presented in Section of "Conclusion."

Related work

Endoscopy

Endoscopy is a non-surgical procedure that allows a physician to examine and observe the inner organ or tissue of a person[31]. Depending on the diagnosed diseases of the body part, each type of endoscopy has its own special term[32], such as oesophagoscopy (oesophagus), gastroscopy (stomach), colonoscopy (colon), and hysteroscopy (uterus). In endoscopy, there is a slender and flexible tube with an optical sensor and light capsule attached to it. Based on Fig. 1, in upper endoscopy, the endoscope is introduced through the mouth, whereas for examination of the intestinal area, the endoscope can be introduced through the anus. Additionally, physicians may use biopsy forceps in the endoscope to remove suspicious growths of tissue for further investigation.

Typically, during an endoscopy screening, more than 55,000 sensor images of the gastrointestinal (GI) tract are captured for each patient, yet abnormalities of the GI tract may turn up in only a few of them[35]. Consequently, it is time-consuming to analyze all the sensor images, because the low contrast, complex background, lesion shape, and the color of each sensor image may affect the results of segmentation and classification[36]. These issues have complicated the diagnosis job, and consequently, multiple opinions from experts are required to avoid misdiagnosis[37]. Thus, there is a high demand for the development of an alternative method to automatically detect abnormalities in the GI tract to reduce the workload of inspecting and analyzing the endoscopy data[35].

Artificial Intelligence-based Polyp Detection Techniques in Endoscopy

Due to the major attention to CNN, it has been utilized in the field of medical image analysis. Several robust methods of automated detection of polyps in colonoscopy have been developed for the early detection of colonic polyps in recent studies. For instance, Shin et al.[10] presented a region-based CNN (Faster R-CNN) incorporating Inception ResNet as a transfer learning (TFL) scheme for the detection system. Their detection system is integrated with post-learning techniques, and several publicly available datasets are utilized in their study, such as CVC-ClinicDB, ETIS-LaribPolypDB, CVC-ClinicVideoDB, and the ASU-Mayo Clinic Colonoscopy video dataset. Besides, another group[18] used the Kvasir database to perform a detection experiment by using different configurations: global feature classification, DL, and TFL. The results showed that the CNN outperformed the random and majority class baseline, while TFL outperformed the DL-based approaches because the parameters of CNN are not optimized and the number of epochs is small. Lin et al.[19] applied a region-based CNN (R-CNN) along with data augmentation, feature extraction, and a fine-tuning model with pre-trained weights from ImageNet and Microsoft COCO (Common Objects in COntext) datasets to detect polyps. The study confirms that R-CNN and the ResNet CNN model perform better, with improved localization on most polyp images compared to other state-of-the-art CNN algorithms.

Mohammed et al.[20] proposed a Y-Net architecture by combining two encoders into a decoder network. The pre-trained VGG-19 is transferred to the first encoder and the images from the ASU-Mayo polyp database are fed into both encoders concatenated with the decoder. Additionally, Zobel et al.[21] exploited the Mask R-CNN architecture with ResNet-101 to extract the image features. Due to the small number of datasets, they also used the TFL approach to train a complex network architecture on the "Bayreuth"DB, CVC-ClinicDB, and ETIS-LaribPolypDB databases. Another group of authors[22] have proposed a single-shot detector (SSD) framework with InceptionV3 as a feature extractor, in which the SSD uses a feed-forward CNN to create a fixed-size boundary box for each object on different feature maps. The SSD model is evaluated using the CVC-ColonDB, CVC-ClinicDB, and ETIS-LaribPolypDB datasets. Besides, Vani et al.[23] conducted an analysis to compare the performance of different DL techniques. They utilized different image datasets from CVC-ColonDB, images captured from real-time wireless capsule endoscope (WCE) and endoscopy images from Endoatlas and Shaily. Nadimi et al.[24] used an optimized ZF-Net algorithm with stochastic gradient descent with momentum (SGDM), which combines data augmentation, pre-processing, and TFL techniques. Furthermore, Patino-Barrientos et al.[25] proposed a DL model based on Kudo's classification scheme, using a VGG-16 as a feature extractor on a private dataset from the University of Deusto.

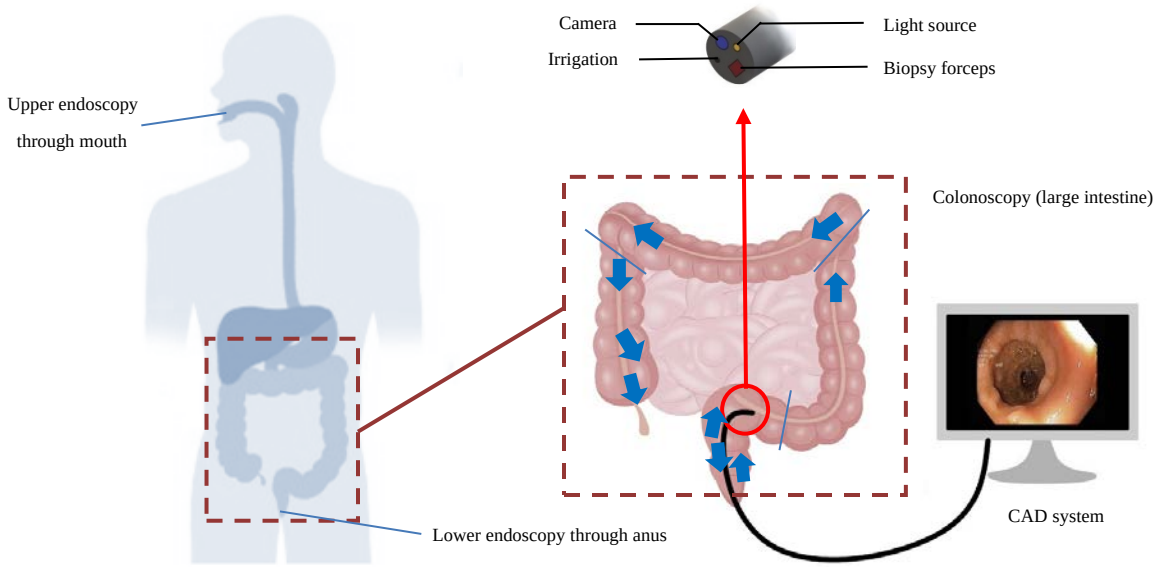

Figure 1. Upper and lower endoscopic examination and a suggestion of systematic screening from European Society of Gastrointestinal Endoscopy in colonoscopy[33,34].

Description	Database		
	Kvasir (Dataset A)	**ETIS-LaribPolypDB (Dataset B)**	**CVC-ClinicDB (Dataset C)**
Type of data	Colour image	Colour image	Colour image
Type of format	.jpg	.tif	.tif
Data dimensions	720 x 576	1225 x 966	384 x 288
Number of data	P: 375; NP: 742	P: 100	P: 300

Table 1. Selected datasets to be used in the research. P – polyp, NP – non-polyp.

Methods

Materials and Tools

Three publicly available datasets (Kvasir, ETIS-LaribPolypDB, and CVC-ClinicDB) with class labels of "polyp" and "non-polyp" were merged to train, validate, and test the detection model on a Windows 10 64-bit operating system with an Intel ® Core i7-2600 CPU at 3.40 GHz with 16GB RAM and an AMD Radeon HD 6450 GPU.

Within the Kvasir database[18], there are eight classes of GI images, but only two classes (polyps and normal colon) were selected to conduct the research. This is because the rest of the classes included oesophagitis, Z-line, pylorus, caecum, rectum, and dyed landmarks, which are not related to this research work. Based on their research[18], they conducted multi-class GI disease detection, which is not the same goal as our work. From the databases in Table 1, there are a total of 1517 colour images including images with and without polyps. Of these, 768 were used to train the detection model, 329 for validation, and the remaining 420 to test the model. Within these huge datasets, the images have different qualities such as brightness, distribution of intensity, position, and size of polyp, allowing the detection model to learn various types of features. Fig. 2 and 3 show some pathological and normal findings of sample images used with different features from the datasets.

Image Pre-Processing

During the endoscopic screening, the camera's optical properties will be affected by different visual qualities, such as artefacts[26], vignettes[26], and illuminations[7]. Therefore, the image needs to undergo pre-processing before the feature-extraction stage to reduce image degradation. A median filter is used to filter out unnecessary information. The median filter[38] replaces all the image pixels at the same time with the pre-defined (3 × 3) neighborhood median of image pixels. Equation (1) below represents a generalized function for any neighborhood:

$$f'(m,n) = Med \mid (-k \leq u,v \leq k)\{F(m+u,n+v)\}, \quad (1)$$

where k = 1 and the median is computed over a 3 × 3 filter, leading to less noise in the image $f'(m,n)$.

Figure 2. Sample colonoscopy images with polyps from datasets: (**a**) Kvasir; (**b**) ETIS-LaribPolypDB; (**c**) CVC-ClinicDB.

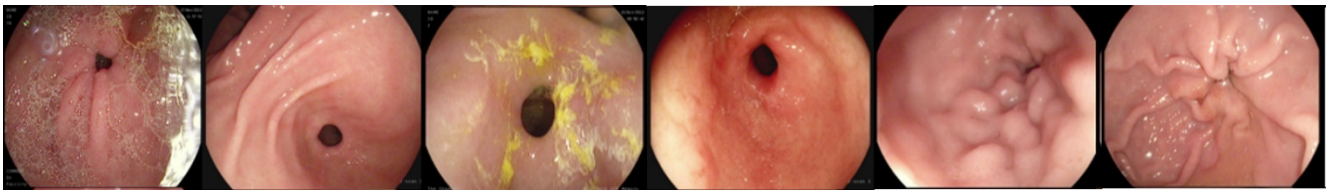

Figure 3. Sample colonoscopy images without polyps from Kvasir dataset only.

In addition, image thresholding is a process of separating the desired object from the background. When applying this method, the image has two classes of pixels: foreground and background[39]. For example, the thresholding method will automatically specify a thresholding value (T), where the pixel values below T are considered as foreground and those above T as background. Apart from that, the image's contrast level might vary due to poor illumination or an improper setting on the capsule's onboard light source. Hence, the contrast of the image must be manipulated to compensate for the image's acquisition by changing the dynamic range of the image[40]. Besides, the pixel values of a low-contrast image and a high-contrast image can be stretched by increasing the dynamic range across the spectrum of the image[40] using Equation (2).

$$O(x,y) = O_1 + \left(\frac{O_2 - O_1}{I_2 - I_1}\right)[I(x,y) - I_1], \tag{2}$$

where O_1 is equivalent to 0 and O_2 is equivalent to the number of desired levels; I_1 and I_2 are the minimum and maximum values of the grey level.

Besides, normalization is a process used to resize the image dimensions to the same scale (224 × 224). It is an important step which ensures that each image has a similar pixel distribution to fit into the neural network. The image normalization is calculated as follows:

$$x' = \frac{x - \bar{x}}{\sigma}, \tag{3}$$

where x is the original feature vector, $\bar{x}$ is the mean of the feature vector, and σ is its standard deviation. The images in Fig. 4 underwent the pre-processing and normalization processes.

Feature Extraction

In ML, feature extraction is required for every image to build the derived values based on the content of the object in the image. In the recent past, many common feature extraction techniques have been used for compact representation of image data, such as histogram of oriented gradients (HOG)[41], speeded-up robust features (SURFs)[41,42], scale-invariant feature

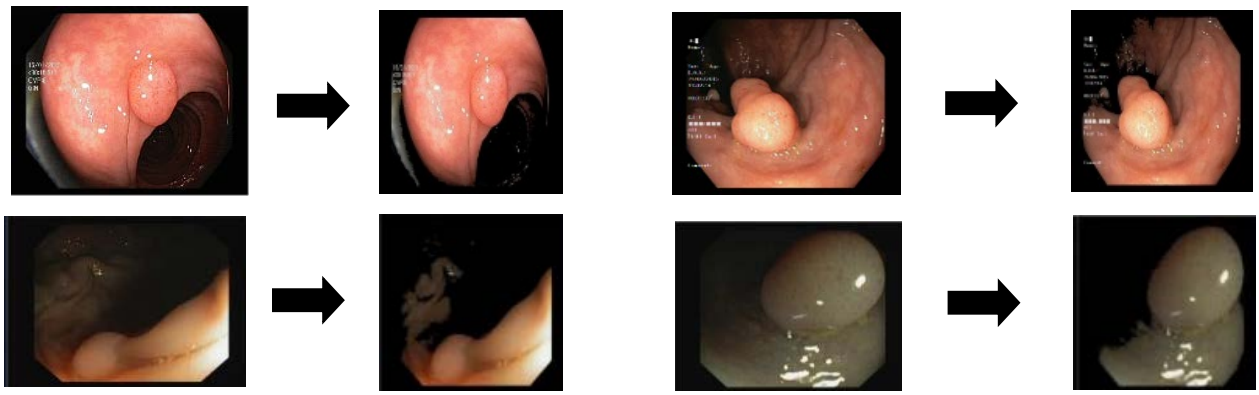

Figure 4. Examples of images that underwent pre-processing and normalization operations.

transform (SIFT)[7,42], and local binary patterns (LBPs)[43]. Nevertheless, in this work, an automatic feature extraction algorithm, DL, is used as a powerful ML technique that uses a CNN. A massive collection of different images can be trained using CNNs. Based on these large collections, CNNs can learn rich feature representations for a vast number of images[10]. These feature representations often transcend the conventional handcrafted features[11,43].

Moreover, the operation in parallel computing allows the matrix operation to be speeded up. The calculation of the output of the neural network is:

$$y = \sigma(\omega^L \dots \sigma(\omega^2 \sigma(\omega^1 x + b^1) + b^2) \dots + b^L), \tag{4}$$

where x is the input, ω is one set of network parameters, b is the bias, σ is the activation function, and y is the output neuron. The activation function plays an important role in DL to perform a multiple combination transformation. For example, the ReLU nonlinear activation function is defined as follows:

$$f(x) = \max(0, x), \tag{5}$$

Other than that, the Softmax function on the fully connected layer turns the logit scores into probabilities that sum to 1. The equation is as follows:

$$S(y_i) = \frac{e^{y_i}}{\sum_{k=1}^{K} e^{y_k}}, \qquad for\ i = 1, \dots, K, \tag{6}$$

where y is the input logit that takes the i^{th} vector value and K is the amount of real numbers for the probability distribution.

Creating a DL model always requires a massive amount of training data, which makes it necessary to build and train a deep neural network from scratch. In this case, TFL is an efficient and practical solution to the very limited availability of data, especially in the medical research field[44]. The employment of TFL can improve the accuracy while reducing the training time[45]. Currently, there are many existing pre-trained CNNs such as LeNet, SqueezeNet, GoogLeNet, AlexNet, and VGG. In this research, ResNet-50 is rebuilt by replacing the final three layers with new features that are specific for the image dataset of interest with little modification of the original network architecture. Once the network is created, it is ready for training. This neural network is well-suited for the classification task[46] as it can learn to make decisions like a human being and plainly follows the variations of circumstance within the database[47]. After performing the TFL using ResNet-50, the feature representations from the training and testing images are obtained through the activations in the fully connected (FC) layer, at the end of the ResNet-50 network. The FC layer pools the input features over all spatial locations, resulting in 2048 extracted features in total.

However, different neural networks have different performances in terms of accuracy and training time. This is associated with the number of deeper layers in the network and the size of the training dataset. If the dataset is huge, TFL might not be faster than training from scratch. In addition, a research study was done to analyze the performance of several pre-trained networks by fitting colonoscopy images from our datasets into each network for training. The performance of each network is presented in Table 2. Inception ResNet-V2 has the highest accuracy, followed by ResNet-50. However, different neural networks are suitable for different tasks. For example, Kornblith et al.[50] claimed that ResNet was better than shallow fine-tuning for artistic data analysis and that it is the best generic visual feature extractor. Due to the cost of the high computational complexity, ResNet-50 is utilized as a feature extractor for the model in the latter classification work and Inception ResNet-V2 is not chosen because of hardware constraints in implementing such a complex deep neural network in future work. However, the computation time depends on the size of the neural network[46] and the amount of data; a deep and complex neural network requires both longer processing time and training time[51].

CNNs	Depth of network	Parameters (millions)	Accuracy (%)	Estimated computational time (mins)
AlexNet	8	61.0	96.10	22
GoogLeNet	22	7.0	96.92	67
ResNet-50	50	25.6	99.02	209
ResNet-101	101	44.6	98.85	288
VGG-19	19	144.0	98.40	492
Inception ResNet-V2	164	55.9	99.30	953

Table 2. Comparison of different pre-trained networks[48,49].

Type of classifiers	Accuracy (%)
Decision tree	95.77
Naïve Bayes	97.50
KNN	97.99
SVM	98.51
AdaBoost	98.70

Table 3. Comparison of different classifiers.

Ensemble Classification

From ML and statistics, supervised learning is one of the classification tasks in learning a function based on the labeled training data. There are some common algorithms for classification, such as support vector machine (SVM), Naïve Bayes classifier, decision tree, k-Nearest Neighbors (KNN), and AdaBoost. In order to identify a best classifier for our proposed model, an experiment was conducted by using HOG[41] as a feature extractor for the colonoscopy images before fitted into the classifiers. The results are demonstrated in Table 3, from which it is found that AdaBoost has the highest accuracy, followed by SVM. AdaBoost has thus been selected for the classification task to distinguish the polyp and non-polyp images. AdaBoost is an ensemble learning algorithm that is used to classify the unseen data instead of just a single classifier. In contrast with other classifiers such as SVM, AdaBoost takes less time to attain a similar learning accuracy[52].

Before the classifier training, PCA is used to reduce the dimensionality of data which consists of many variations while retaining the present variation (eigenvalues) in the dataset to the maximum extent[53]. PCA is a statistical process that is widely used for data analysis and dimension reduction[54]. Based on the eigenvectors in PCA, the original N-dimensional data are transformed to M-dimensional data[55]. In the new M-dimensional data, the number of features is typically less than or equal to the number of original features; they will never be higher in number than the original ones. When the 2048 extracted features with high dimensions are reduced using PCA, 90% of the data variance (with 227 features) is described and most of the significant features remain while the computation and information redundancy is kept minimal. After the PCA analysis, the vector of feature input data and the vector of class labels will be used to train the classifier model[56]. Furthermore, the ensemble combines the decisions from the multiple classifier[56] to improve the overall performance by boosting the weak classifiers, as shown in Fig. 5. During each training iteration, the weight of each sample will be altered based on the classifier error rate obtained[57]. The pseudocode of AdaBoost is described as in Fig. 6. Fig. 7 demonstrates the combination of classifiers for the final classifier *Y(x)*.

Performance Measures and Evaluation

The performance of the proposed CAD system is measured by computing a confusion matrix, which is widely used in statistical classification problems and is a table layout with two rows and two columns that visualizes the following parameters:

True Positive (TP): A polyp is detected in a frame that contains a polyp.
False Positive (FP): A polyp is detected in a frame without a polyp image.
True Negative (TN): No polyp is detected in a frame without a polyp image.
False Negative (FN): A polyp is missed in a frame that contains a polyp.

Based on the parameters above, the usual performance derivations such as accuracy, F1 score, positive predictive value (PPV), sensitivity (TPR), specificity, and false positive rate (FPR) are defined:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%, \quad (7)$$

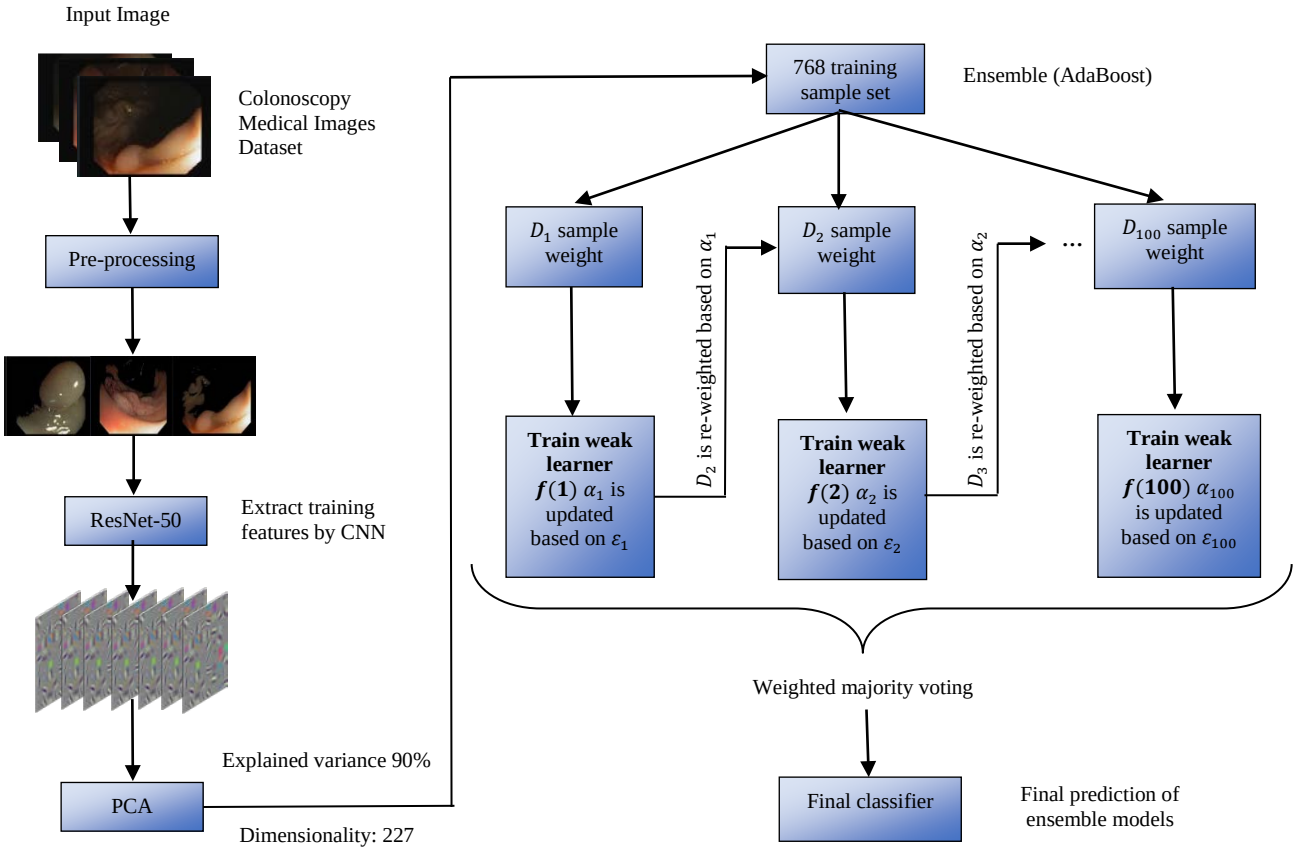

Figure 5. Overall methodology of proposed ensemble learning by boosting. The weak learners are trained iteratively to correct the mistakes made by previous models until the final model with the lowest bias is obtained.

Algorithm 1 AdaBoost

Input: a set of data $\{(x_1, y_1), \ldots, (x_{768}, y_{768})\}$

initialize $w_1^i = 1/n$ for $i = 1,2, \ldots,768$

there are 100 classifiers

for $k = 1{:}100$

train weak classifier $f_k(x)$ using w_k^i

obtain error rate ε_k for 100 classifiers

calculate $\alpha_k = 0.5 * \ln((1 - \varepsilon_k)/\varepsilon_k))$

re-weight, for $i = 1,2, \ldots,768$:

$$w_{k+1}^i = \begin{cases} w_k^i * e^{\alpha_k} & if\ f_k(x_i) \neq y_i \\ w_k^i * e^{-\alpha_k} & if\ f_k(x_i) = y_i \end{cases}$$

get $f_1(x), f_2(x), \ldots, f_{100}(x)$ classifiers

weighted-sum 100 classifiers

Output: final classifier,

$$Y(x) = sign\left(\sum_{k=1}^{100} \alpha_k f_k(x)\right)$$

Figure 6. The boosting algorithm - AdaBoost.

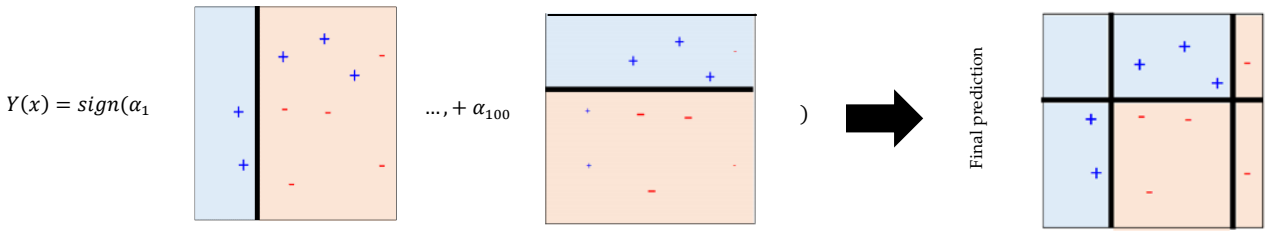

Figure 7. The weighted majority vote for each prediction.

$$Error\ rate = \frac{FP + FN}{TP + TN + FP + FN} \times 100\%\,, \tag{8}$$

$$PPV\ (precision) = \frac{TP}{TP + FP} \times 100\%\,, \tag{9}$$

$$Sensitivity\ (TPR, recall) = \frac{TP}{TP + FN} \times 100\%\,, \tag{10}$$

$$F1\ score = 2 \times \left(\frac{precision \times recall}{precision + recall}\right), \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \times 100\%\,, \tag{12}$$

$$FPR = \frac{FP}{FP + TN}\,, \tag{13}$$

The F1 score is the weighted mean of Equations (9) for precision and (10) for recall. Sensitivity refers to the ratio of correctly predicted positives to the total number of positive cases, while specificity is the ratio of correctly predicted negatives to the total number of negative cases. In addition, a receiver operating characteristic (ROC) curve can be obtained through Equations (10) and (13).

Results

Performance of Computer-Aided Diagnosis (CAD) System

The performance evaluation of the proposed algorithm for CAD system are shown in the following section. Table 4 compares the performance of the best performers and latest research works with our proposed method; for the combination of three datasets, our method achieved a better performance than all the other methods. However, an experiment was done using only the Kvasir dataset with our proposed algorithm (70% of images were selected randomly for training; the remaining 30% were used for testing). Similarly, the same model was tested on the other two datasets, ETIS-LaribPolypDB and CVC-ClinicDB. Due to the unavailability of non-polyp images (normal colon) in both datasets, *FP* and *TN* are zero.

In medical decision-making, the ROC curve is used to check or visualize the performance of classifiers[58]. The ROC is a probability curve and is plotted with sensitivity (true positive rate, TPR) against the FPR (1 – true negative rate, TNR) at various thresholds, while the AUC (area under the curve) is a measure of divisibility to prove the capability of the model to distinguish the classes. In general, a higher AUROC (area under the receiver operating characteristic) curve indicates better performance of the model. The ROC curve is plotted to demonstrate the performance of the proposed CAD system in Fig. 8a. Three databases were grouped together to train the CAD model. Nevertheless, the ROC curve in Fig. 8b shows the model trained by one dataset (Kvasir) only, because ETIS-LaribPolypDB and CVC-ClinicDB do not have the images without polyps. As a result, the most accurate model for the CAD system was obtained by combining three databases which consist of more features learned from ETIS-LaribPolypDB and CVC-ClinicDB. For this model, the accuracy, error rate, F1 score, sensitivity, and specificity of the colonic polyp classification task were 99.08%, 0.92%, 99.08%, 99.41%, and 98.75%, respectively.

Moreover, the performance of the model is estimated through the plot of the generalization error values during the learning process in Fig. 9. The generalization error and overfitting are closely related, as the generalization error is computed to measure the ability of the model to predict the outcome values for the formerly unseen data. The amount of overfitting is tested using the five-fold cross-validation method[59]: the smaller the generalization error, the less overfitting has occurred. According to Fig. 9, the cumulative generalization error decreases to approximately 1.5% when 43 weak learners compose the ensemble classifier.

Method	Database	Technique	Performance Evaluation Criteria					
			Acc	Er	F1	Sen	Spec	FPR
Proposed	Dataset A	• Preprocessing • CNN TFL ResNet-50 (FE) • PCA • Ensemble (AdaBoost)	98.64	1.36	98.62	97.90	99.38	0.0062
	Dataset B		46.00	N/A	N/A	46.00	N/A	N/A
	Dataset C		100	N/A	N/A	100	N/A	N/A
	Dataset A, B, and C		99.08	0.92	99.08	99.41	98.75	0.0125
Wittenberg et al., (2019) [20]	Dataset B, C, and "Bayreuth"DB	• Mask R-CNN • ResNet-101 (FE) • TFL	N/A	N/A	83.33	87.00	N/A	N/A
Liu et al., (2019) [21]	Dataset B, C, and CVC-ColonDB	• Preprocessing • SSD framework • InceptionV3 (FE)	N/A	N/A	76.80	80.30	N/A	N/A
Vani et al., (2019) [22]	CVC-ColonDB , WCE video frames and endoscopy images from Endoatlas and Shaily	• Data augmentation • DL technique (VGG-19) using Keras framework	94.45	5.55	93.00	94.00	N/A	N/A
Nadimi et al., (2020) [23]	Images captured from colon capsule endoscopy	• Data augmentation • CNN TFL (ZF-Net) • SGDM	98.00	2.00	N/A	98.10	96.30	N/A
Patino-Barrientos et al., (2020) [24]	Private dataset from University of Deusto	• Preprocessing • 4 layers base CNN model • VGG-16 (FE)	83.00	17.00	83.00	86.00	N/A	N/A

Table 4. Performance comparison between previous works and our ensemble learning method with different sets of colonoscopy/endoscopy images. Acc – accuracy, Er – error rate, F1 – F1-score, Sen – sensitivity, Spec – Specificity, FPR – false positive rate, Dataset A – Kavsir, Dataset B – ETIS-LaribPolypDB, Dataset C – CVC-ClinicDB, CNN – convolutional neural network, TFL – transfer learning, FE – feature extraction, PCA – principal component analysis, SSD – single shot detection, DL – deep learning, SGDM – stochastic gradient descent with momentum.

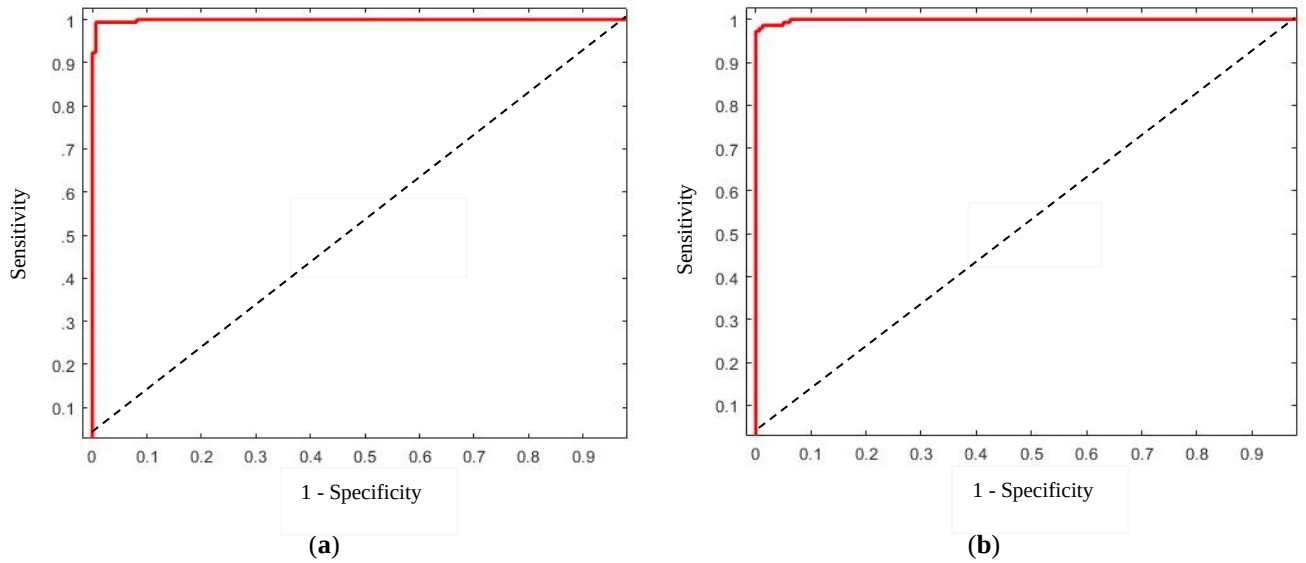

Figure 8. ROC curve for polyp classification of (**a**) a combination of three databases: Kvasir (Dataset A), ETIS-LaribPolypDB (Dataset B), and CVC-ClinicDB (Dataset C); (**b**) Kvasir (Dataset A) database only.

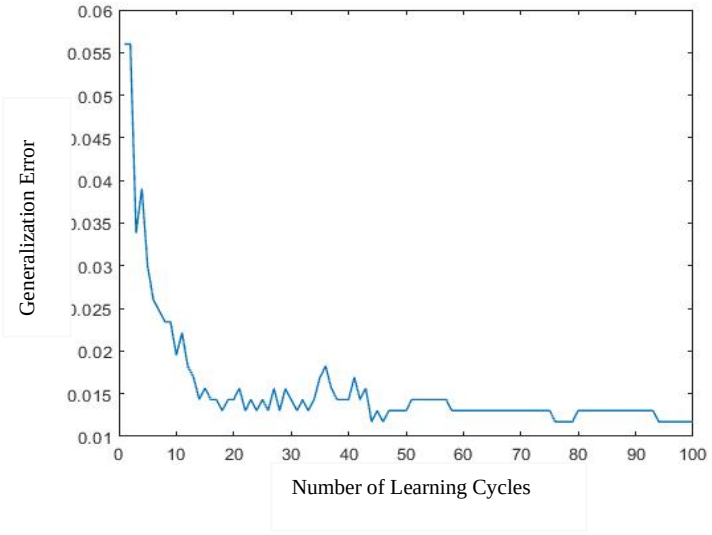

Figure 9. Generalization error over numbers of learning cycles.

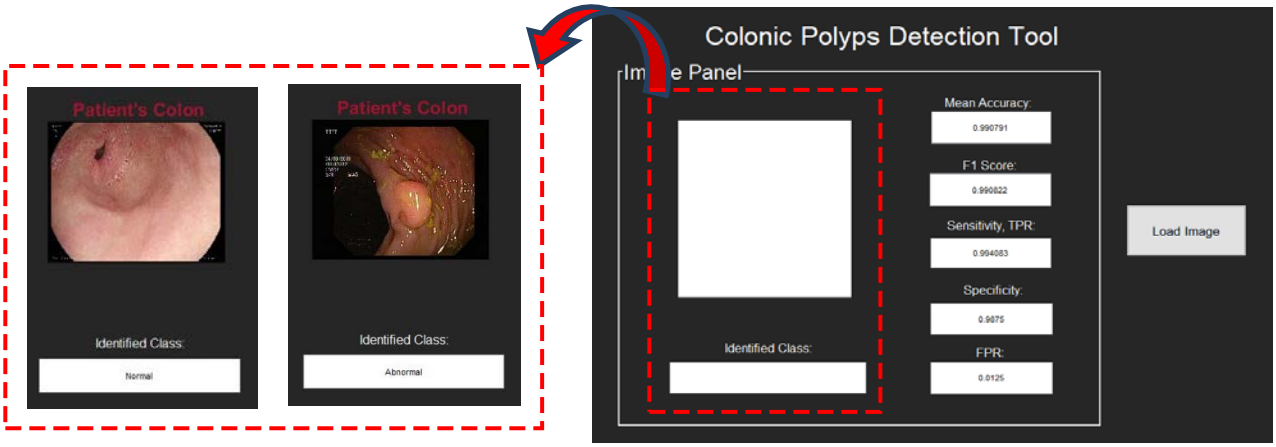

Figure 10. The GUI for the colonic polyp detection tool. A normal output class label is shown when a non-polyp image is loaded and an abnormal one is shown when a polyp image is loaded.

Graphical User Interface (GUI)

The network parameters from training were saved and a graphical user interface (GUI) was created to load the images. The GUI was designed as shown in Fig. 10. All the parameters (accuracy, F1 score, sensitivity/ TPR, specificity, and FPR) will be shown in the GUI and it is ready to load the image from the testing dataset for a classification task. The classification result (normal or abnormal cases) will be shown in the identified class column of the GUI.

Datasets and Misjudgments by Detection System

Due to the limited availability of public databases, three datasets with a total of 1097 images are merged to train and validate the model. However, among all the datasets, there are slightly more images with than without a polyp. Hence, an analysis was performed, as shown in Table 5, and indicates that the model was not biased towards the majority class. The two different classes (polyp and non-polyp) are separated into three different ratio configurations, where i) there are more images with a polyp than without a polyp; ii) the number of images with a polyp is equal to those without a polyp; iii) there are fewer images with a polyp than without a polyp. For the first verification, there are 30% more images with a polyp than without a polyp; for the second, the number of images with a polyp is the same as the number without; finally for the third, there are 50% fewer images with a polyp than without a polyp.

Evaluation of model	Number of images		
	Balanced datasets	Imbalanced dataset	
	P: 532; NP: 532	P: 565; NP: 396	P: 266; NP 532
Accuracy	99.06	96.78	93.68
Error rate	0.94	3.22	6.32
Sensitivity	98.12	98.82	91.11
Specificity	100	94.74	96.25
AUC	0.9992	0.9994	0.9368

Table 5. Performance of the proposed model with balanced and imbalanced datasets. P – polyp, NP – non-polyp.

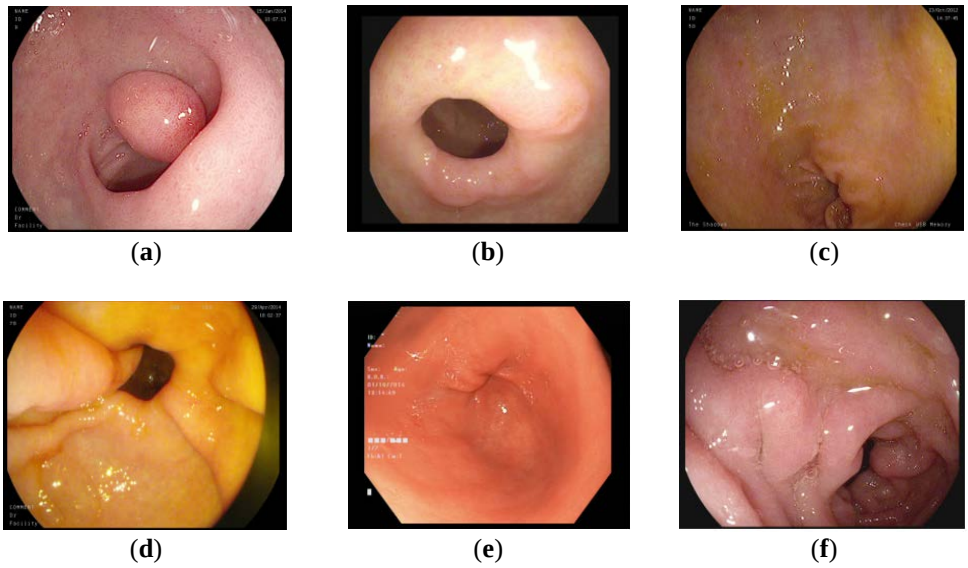

Figure 11. Misclassified images. (**a**) Image misjudged as non-polyp; (**b**)-(**f**) Images misjudged as polyps.

However, the automated detection model still suffers from inaccuracy or misjudgements of endoscopic images. All images were tested, and it was found that several images were misjudged due to the texture, density, and shape information with the image context of the polyp, causing the system to detect irrelevant objects that resemble polyps such as ulcers, inflammation, or bleeding. This problem is very challenging, as polyps and non-polyps have very similar characteristics, such as shapes, colors, and surfaces of structures. Thus, it is tough to discriminate the features between them. Fig. 11 demonstrates those images which are misclassified. In Fig. 11a, the polyp is wrongly detected due to its context; the polyp is separated, as an unwanted object was filtered out during pre-processing. On the other hand, the images in Fig. 11b–f are misclassified, as polyps are detected. This is because the structure of the colon is uneven and folded, causing it sometimes to resemble polyps based on the texture and shape.

Discussion

Based on the plots and results above, the proposed approach achieved a good performance by selecting ResNet-50 and AdaBoost as a new combination technique to automatically classify the polyps. Different combinations will result in different performance. Using a deeper layer for feature extraction improves the classification performance compared to the shallower layers, because deeper layers contain higher-level features, which are built using the lower-level features of earlier layers. Nevertheless, classifiers also play an important role in distinguishing features in classification problems. Based on Table 3, AdaBoost outperformed the other classifiers as it is adaptive in the sense that the subsequent classifiers constructed are tuned in support of those instances misclassified by previous classifiers. After executing the AdaBoost algorithm, the final “strong” classifier has improved the classification performance. Furthermore, more data can improve the overall performance. According to Table 4, a research was done by training the model using Dataset A only, and tested on Dataset B. It was found that the results were quite poor. This is because the images of the abnormal cases in Dataset B (refer to Fig. 2b) has very similar color with the images of the normal cases in Dataset A (refer to Fig. 3), a lot of polyp images was therefore misjudged as non-polyp. Besides, the same model was tested on Dataset C. Due to the similar characteristics of images on both Datasets A and C, we obtained good results with 100% of accuracy and sensitivity. With the combination of three datasets, the model can learn more features, which can improve the classification performance of the model.

Conclusion

In this paper, a novel and automated polyp-detection system was developed by using integrated TFL and ensemble learning. ResNet-50 was used as the feature extractor, and the last few layers were transferred and rebuilt as fully connected layers with new features specifically from the image dataset. Therefore, the performance of the model can be improved while reducing training time. Other than that, an adaptive boosting-based ensemble classifier was used to learn and categorize the images based on the principal component of feature extraction with the class labels (non-polyp or polyp) from the training dataset. The presence of polyps is detected with an accuracy of 99.08% in the trial of 1517 images from the combination of three free publicly accessible databases. The proposed detection system achieved a low error rate of 0.92%, with sensitivity, specificity, and F1 score of polyp detection are 99.08%, 99.41%, and 98.75%, respectively. The methods developed so far are therefore promising as the basis for subsequent expert human assessment, and they dramatically outperformed all the existing works. Additionally, the proposed approach is reliable and effective, as it can greatly reduce the polyp miss-detection rate, which caused by the subjectivity of assessment during the visual colonoscopy evaluation. Due to the naturally irregular structure of the colon, the detection system will still misclassify some images. This might be a limitation of the system and should be improved in the future if it is to be utilized in colonoscopy for real-time detection. In the future, the effectiveness of the proposed detection system will be tested and compared in assisting physicians in detecting polyps.

References

1. Ozawa, T. *et al.* Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. *Therap Adv Gastroenterol* **13**, (2020).
2. Siegel, R. L. *et al.* Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians* **70**, 145–164 (2020).
3. Paik, D. S. *et al.* Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT. *IEEE Transactions on Medical Imaging* **23**, 661–675 (2004).
4. Tajbakhsh, N., Chi, C., Gurudu, S. R. & Liang, J. Automatic polyp detection from learned boundaries. in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* 97–100 (2014). doi:10.1109/ISBI.2014.6867818.
5. Stoitsis, J. *et al.* Computer aided diagnosis based on medical image processing and artificial intelligence methods. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **569**, 591–595 (2006).
6. Bose, B. K. Expert system, fuzzy logic, and neural network applications in power electronics and motion control. *Proceedings of the IEEE* **82**, 1303–1323 (1994).
7. Gueye, L., Yildirim-Yayilgan, S., Cheikh, F. A. & Balasingham, I. Automatic detection of colonoscopic anomalies using capsule endoscopy. in *2015 IEEE International Conference on Image Processing (ICIP)* 1061–1064 (2015). doi:10.1109/ICIP.2015.7350962.
8. Lowe, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* **60**, 91–110 (2004).
9. Chao, W.-L., Manickavasagan, H. & Krishna, S. G. Application of Artificial Intelligence in the Detection and Differentiation of Colon Polyps: A Technical Review for Physicians. *Diagnostics (Basel)* **9**, (2019).
10. Shin, Y., Qadir, H. A., Aabakken, L., Bergsland, J. & Balasingham, I. Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches. *IEEE Access* **6**, 40950–40962 (2018).
11. Bernal, J. *et al.* Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results From the MICCAI 2015 Endoscopic Vision Challenge. *IEEE Transactions on Medical Imaging* **36**, 1231–1249 (2017).
12. Tajbakhsh, N., Gurudu, S. R. & Liang, J. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)* 79–83 (2015). doi:10.1109/ISBI.2015.7163821.
13. Zhang, R., Zheng, Y., Poon, C. C. Y., Shen, D. & Lau, J. Y. W. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. *Pattern Recognition* **83**, 209–219 (2018).
14. Tajbakhsh, N., GURUDU, S. R. & Liang, J. System and methods for automatic polyp detection using convolutional neural networks. (2016).
15. Brandao, P. *et al.* Towards a Computed-Aided Diagnosis System in Colonoscopy: Automatic Polyp Segmentation Using Convolution Neural Networks. *J. Med. Robot. Res.* **03**, 1840002 (2018).
16. Yu, L., Chen, H., Dou, Q., Qin, J. & Heng, P. A. Integrating Online and Offline Three-Dimensional Deep Learning for Automated Polyp Detection in Colonoscopy Videos. *IEEE Journal of Biomedical and Health Informatics* **21**, 65–75 (2017).
17. Zhang, R. *et al.* Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level CNN Features

From Nonmedical Domain. *IEEE Journal of Biomedical and Health Informatics* **21**, 41–47 (2017).

18. Pogorelov, K. *et al.* KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. in *Proceedings of the 8th ACM on Multimedia Systems Conference* 164–169 (ACM, 2017). doi:10.1145/3083187.3083212.

19. Lin, T.-Y. *et al.* Microsoft COCO: Common Objects in Context. in *Computer Vision – ECCV 2014* (eds. Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) 740–755 (Springer International Publishing, 2014). doi:10.1007/978-3-319-10602-1_48.

20. Mohammed, A., Yildirim, S., Farup, I., Pedersen, M. & Hovde, Ø. Y-Net: A deep Convolutional Neural Network for Polyp Detection. *arXiv:1806.01907 [cs]* (2018).

21. Wittenberg, T., Zobel, P., Rathke, M. & Mühldorfer, S. Computer Aided Detection of Polyps in Whitelight-Colonoscopy Images using Deep Neural Networks. *Current Directions in Biomedical Engineering* **5**, 231–234 (2019).

22. Liu, M., Jiang, J. & Wang, Z. Colonic Polyp Detection in Endoscopic Videos With Single Shot Detection Based Deep Convolutional Neural Network. *IEEE Access* **7**, 75058–75066 (2019).

23. Vani, V. & Prashanth, K. V. M. Polpy Detection in Endoscopy Image Using Deep Learning. *EC Gastroenterology and Digestive System*, 663-672 (2019).

24. Nadimi, E. S. *et al.* Application of deep learning for autonomous detection and localization of colorectal polyps in wireless colon capsule endoscopy. *Computers & Electrical Engineering* **81**, 106531 (2020).

25. Patino-Barrientos, S., Sierra-Sosa, D., Garcia-Zapirain, B., Castillo-Olea, C. & Elmaghraby, A. Kudo's Classification for Colon Polyps Assessment Using a Deep Learning Approach. *Applied Sciences* **10**, 501 (2020).

26. Mamonov, A. V., Figueiredo, I. N., Figueiredo, P. N. & Richard Tsai, Y.-H. Automated Polyp Detection in Colon Capsule Endoscopy. *IEEE Transactions on Medical Imaging* **33**, 1488–1502 (2014).

27. Sharon, H., Elamvazuthi, I., Lu, C.-K., Parasuraman, S. & Natarajan, E. Development of Rheumatoid Arthritis Classification from Electronic Image Sensor Using Ensemble Method. *Sensors* **20**, 167 (2020).

28. Alpaydin, E. *Introduction to machine learning.* (MIT Press, 2010).

29. Polyp - Grand Challenge. *grand-challenge.org* https://polyp.grand-challenge.org/EtisLarib/.

30. Cvc-Clinicdb - Polyp - Grand Challenge. https://polyp.grand-challenge.org/CVCClinicDB/.

31. Endoscopy: Types, preparation, procedure, and risks. https://www.medicalnewstoday.com/articles/153737 (2017).

32. Types of Endoscopy. *Cancer.Net* https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/tests-and-procedures/types-endoscopy (2011).

33. Marques, S., Bispo, M., Pimentel-Nunes, P., Chagas, C. & Dinis-Ribeiro, M. Image Documentation in Gastrointestinal Endoscopy: Review of Recommendations. *GE Port J Gastroenterol* **24**, 269–274 (2017).

34. Rey, J. F., Lambert, R. & ESGE Quality Assurance Committee. ESGE recommendations for quality control in gastrointestinal endoscopy: guidelines for image documentation in upper and lower GI endoscopy. *Endoscopy* **33**, 901–903 (2001).

35. Alaskar, H., Hussain, A., Al-Aseem, N., Liatsis, P. & Al-Jumeily, D. Application of Convolutional Neural Networks for Automated Ulcer Detection in Wireless Capsule Endoscopy Images. *Sensors* **19**, 1265 (2019).

36. Liaqat, A. *et al.* Automated ulcer and bleeding classification from wce images using multiple features fusion and selection. *J. Mech. Med. Biol.* **18**, 1850038 (2018).

37. Charfi, S. & Ansari, M. E. Computer-aided diagnosis system for colon abnormalities detection in wireless capsule endoscopy images. *Multimed Tools Appl* **77**, 4047–4064 (2018).

38. Nelikanti, A. Colorectal Cancer MRI Image Segmentation Using Image Processing Techniques. in (2014).

39. Jeyavathana, R. B., Balasubramanian, D. R. & Pandian, A. A. A Survey: Analysis on Pre-processing and Segmentation Techniques for Medical Images. /paper/A-Survey%3A-Analysis-on-Pre-processing-and-Techniques-Jeyavathana-Balasubramanian/9a2becc0f82325c0ff4d3aee625dbe5c58e7e3b8 (2016).

40. Alginahi, Y. Preprocessing Techniques in Character Recognition. *Character Recognition* (2010) doi:10.5772/9776.

41. Kibria, S. B. & Hasan, M. S. An analysis of Feature extraction and Classification Algorithms for Dangerous Object Detection. in *2017 2nd International Conference on Electrical Electronic Engineering (ICEEE)* 1–4 (2017). doi:10.1109/CEEE.2017.8412846.

42. Simon, A. State of the Art of Object Recognition Techniques. *Sci. Semin. Neurocientific Syst. Theory* (2016).

43. Zhu, R., Zhang, R. & Xue, D. Lesion detection of endoscopy images based on convolutional neural network features. in *2015 8th International Congress on Image and Signal Processing (CISP)* 372–376 (2015).

doi:10.1109/CISP.2015.7407907.

44. Shie, C.-K., Chuang, C.-H., Chou, C.-N., Wu, M.-H. & Chang, E. Y. Transfer representation learning for medical image analysis. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2015) doi:10.1109/EMBC.2015.7318461.

45. Felzenszwalb, P. F., Girshick, R. B., McAllester, D. & Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**, 1627–1645 (2010).

46. Ansari, A. & Bakar, A. A. A Comparative Study of Three Artificial Intelligence Techniques: Genetic Algorithm, Neural Network, and Fuzzy Logic, on Scheduling Problem. in *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology* 31–36 (2014). doi:10.1109/ICAIET.2014.15.

47. Franklin, S. Learning high quality decisions with neural networks in" conscious" software agents. *WSEAS Transactions on Systems.*

48. Canziani, A., Paszke, A. & Culurciello, E. An Analysis of Deep Neural Network Models for Practical Applications. (2016).

49. Pretrained Deep Neural Networks - MATLAB & Simulink - MathWorks United Kingdom. https://uk.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html.

50. Kornblith, S., Shlens, J. & Le, Q. V. Do Better ImageNet Models Transfer Better? in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2656–2666 (IEEE, 2019). doi:10.1109/CVPR.2019.00277.

51. Whitley, D. Genetic Algorithms and Neural Networks. in *Genetic Algorithms in Engineering and Computer Science* 191–201 (John Wiley, 1995).

52. Vink, J. P. & de Haan, G. Comparison of machine learning techniques for target detection. *Artif Intell Rev* **43**, 125–139 (2015).

53. Lu, H., Yang, L., Yan, K., Xue, Y. & Gao, Z. A cost-sensitive rotation forest algorithm for gene expression data classification. *Neurocomputing* **228**, 270–276 (2017).

54. Ibrahim, M. F. I. & Al-Jumaily, A. A. PCA indexing based feature learning and feature selection. in *2016 8th Cairo International Biomedical Engineering Conference (CIBEC)* 68–71 (2016). doi:10.1109/CIBEC.2016.7836122.

55. Zhu, M. *et al.* PCA and Kernel-based extreme learning machine for side-scan sonar image classification. in *2017 IEEE Underwater Technology (UT)* 1–4 (2017). doi:10.1109/UT.2017.7890275.

56. Lu, H., Meng, Y., Yan, K. & Gao, Z. Kernel principal component analysis combining rotation forest method for linearly inseparable data. *Cognitive Systems Research* **53**, 111–122 (2019).

57. Pang, S., Zhang, Y., Ding, M., Wang, X. & Xie, X. A Deep Model for Lung Cancer Type Identification by Densely Connected Convolutional Networks and Adaptive Boosting. *IEEE Access* **8**, 4799–4805 (2020).

58. Tartar, A. & Akan, A. Ensemble learning approaches to classification of pulmonary nodules. in *2016 International Conference on Control, Decision and Information Technologies (CoDIT)* 472–477 (2016). doi:10.1109/CoDIT.2016.7593608.

59. Generalization error. *Wikipedia* (2020).

Author contributions

T.B.T. and C.K.L. conceived the research, W.S.L. and C.K.L. conducted the research, W.S.L., T.B.T., and C.K.L. analysed the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.