Informatics in Medicine Unlocked 24 (2021) 100603

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Computerized classification of gastrointestinal polyps using stacking ensemble of convolutional neural network

Mohammad Motiur Rahman *, Md. Anwar Hussen Wadud, Md. Mahmodul Hasan

Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Santosh, Tangail, 1902, Bangladesh

ARTICLE INFO

Keywords:
Polyp classification
Convolutional neural network
Stacking ensemble
Gastrointestinal polyp
Medical diagnosis

ABSTRACT

In this paper, we have proposed a classification method of gastrointestinal polyps using the stacking ensemble technique. The ensemble method consisted of three fine-tuned deep convolutional neural network architectures (Xception, ResNet-101, and VGG-19), and the network weights were initialized from the ImageNet dataset. Besides, this paper presented a multi-attribute decision-making technique-based frame selection method utilizing several measures of a suitable frame. The frame selection procedure reduces the processing overhead of the system and attained better classification results. Moreover, this study applied a set of image enhancement operations to remove specular reflection, clipping unnecessary regions, contrast enhancement, and noise reductions. The specified classification method of polyps showed significant improvement in performance metrics on available public datasets. The five-fold cross-validated performance of the study has an accuracy of 98.53 ± 0.62%, recall score of 96.17 ± 0.87%, a precision value of 92.09 ± 4.62%, a specificity score of 98.97 ± 0.36%, and an AUC score of 0.9912. This method can be helpful for endoscopists to make rigid decisions.

1. Introduction

Along with the changing lifestyle of humankind, many different diseases come out drastically. Colorectal cancer (CRC) is one of those terrifying diseases that have an immersive impact as a result of the modern lifestyle. The number of CRC patients is rising on the contrary mortality rate is steadily decreasing comparatively over the last decade [1]. The slight decrease in mortality is possible due to the rise in the number of colonoscopies increases per annum. A colonoscopy is a particular form of endoscopy while endoscopy, a generalized term of medical procedure for seeing the inner organs of the human body, performed using a long tube with a tiny camera inserted into respective human cavities. Colonoscopy is mainly responsible for inspecting the lower abdominal organs for instance the rectum and large intestine. The entire exhausted colonoscopy procedure is the procedure of diagnosing abnormalities in digestive organs. This complex process leads to failure at recognizing polyps for several unavoidable reasons like expertise, human factors, the variability of abnormalities, and many more. Real-time detection and classification will reduce the gaps in skill and experience of endoscopists during colonoscopy. Moreover, endoscopists can be more observant to inspect classified polyp regions. Thus, this will significantly decrease the risk of misdiagnosis. On average, more than 27% of colonoscopies failed to diagnose polyps during colonoscopy [2]. These misdiagnosed polyps may turn into CRC at later stages of life.

CRC is a relatively common disease of the gastrointestinal system that is curable at its early stages. At earlier stages, the polyp can be harmlessly removed which reduces the risk of conversion into fatal CRC. Despite this tremendous opportunity of lessens the effect of CRC, diagnosis of polyps is precarious. Thus, it's an active research area for many years and growing, and researchers from diverse backgrounds work on the automatic diagnosis of polyps. The variability in shape, texture, color, and other factors make it firm to diagnose polyps correctly. Prior studies have concentrated primarily on polyp detection, localization, and segmentation [3–5]. In contrast, not many studies on the classification of polyps have been done. However, the classification of polyp is an essential aspect as all types of polyps are not hazardous. Identification of non-cancerous polyps and cancerous polyps is desirable to reduce the cost and complication of endoscopic operation. We have proposed a polyp classification method to classify polyps into three categories (adenomas, hyperplastic and serrated) and a non-polyp category.

The previous researches on polyp classification used traditional computer vision, machine learning, and deep learning techniques. Recent studies using advanced deep learning techniques have achieved success in many different fields on medical diagnosis and especially in

* Corresponding author.
E-mail addresses: mm73rahman@gmail.com (M.M. Rahman), mahwadud@gmail.com (Md.A.H. Wadud), mahmodul.mbstu@gmail.com (Md.M. Hasan).

https://doi.org/10.1016/j.imu.2021.100603
Received 23 December 2020; Received in revised form 11 May 2021; Accepted 15 May 2021
Available online 27 May 2021

polyp detection, localization, and segmentation. We have proposed a deep neural network approach for polyp classification. The procedure consists of three different convolutional neural networks (CNN) architectures. The models accumulated together to form a stacking ensemble model for the final classification decision. The CNN models perform feature extraction and classification on their own, which can extract both lower- and higher-level features from images. Hence, preprocessing schemes play a vital role to have better performance. We have applied different processes to enhance pictures to achieve boosted performance. The system included removing specular reflection, cropping polyp frames, contrast enhancement, selecting better frames, and others. Besides, the tuning of CNN models with different parameters and hyperparameters was examined to find the optimum values. This study proposed an ensemble using three different kinds of CNN architecture for the classification of polyps. The VGG-19 extracts feature using regular convolutions and VGG net as underlying building blocks of 19 layers. The ResNet-101 is a profound network of 101 layers and uses the residual technique for training the network and captures features from images. The model Xception uses another form of convolution mechanism that speeds up model training and extracts features with 71 deep-network layers. The combination of these three networks introduces the diversity of feature extraction, heterogeneity of features that significantly improves the performance. The contributions of the study are:

- A stacking ensemble classification technique that utilizes a set of deep CNN models of three distinct architectures. The ensemble consists of VGG net (VGG19), Res net (ResNet-101), and Xception. The incorporation of multiple networks helps to eliminate the drawbacks of a single model.
- We developed a frame selection method for video data having reduced processing time and better performance. The frame selection method uses a multi-attribute decision-making technique to select suitable frames from video.
- Also, we proposed a set of preprocessing operations to remove specular reflections from images. The light reflection is a significant obstacle for diagnosing polyps.

The rest of the paper is structured as follows: Section 2 contains similar articles on the classification of polyps and suggested techniques. In Section 3, the method studied is shown with the working theory of the elements of this method. The performance analysis, importance and effects of the proposed approach and comparison with past research are interpreted in Section 4. Section 5 concludes the paper with future scopes.

2. Related works

Many researchers from different perspectives have conducted researches on the diagnosis of gastrointestinal polyps. However, there is a lot of scopes to improve diagnostic performance. The previous works on polyp diagnosis can be broadly classified into three kinds, specifically detection, segmentation, and classification. Classification of polyps is relatively less inspected compared to other domains. The polyp classification works are performed using different technologies such as computer vision, machine learning, and deep learning.

Deep learning methods perform significantly in polyp detection, localization, and segmentation. Researchers also proposed classification methods using the deep learning paradigm. Most of these studies used a particular deep learning model after comparing a few models. The machine learning approach for classifying polyp images from texture, color, and shape features was presented in Ref. [6]. The authors of the study applied support vector machine-based multiclass and binary

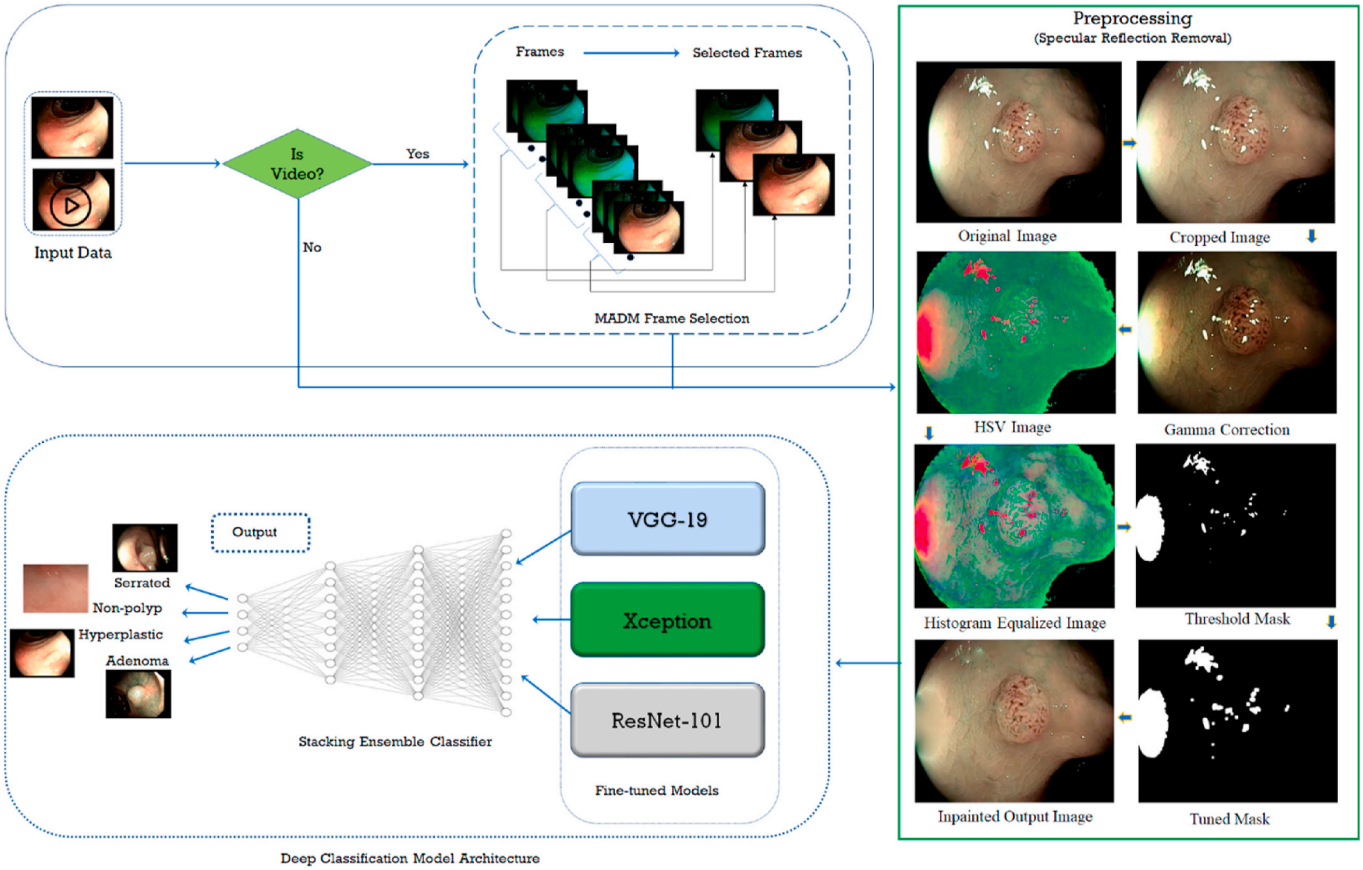


Fig. 1. Block diagram of polyp classification methodology.

classification of polyps from image features of the UCI repository dataset. The images were captured using the narrow-band imaging (NBI) technique. Their research took a huge running time for classification. Mahmood et al. proposed a semi-supervised method using predicted depth information while training their model [7]. Their system performs both segmentation and classification and having 87.24% accuracy on classification.

Classification of serrated and adenoma polyps using Inception-ResNet-v2 was proposed in Ref. [8]. They provided a comparison of white light (WL) and NBI images for diagnosis of polyps where NBI images perform significantly. Their method attained an accuracy of no more than 89% on testing. Patel et al. studied an end-to-end polyp classification method using CNN methods. They have studied six different CNN architectures on two different sets of data in their study. The VGG-19 and VGG-19 with batch-normalization worked comparatively well than other models [9].

Fusion methods of polyp classification are not extensively studied. Bi-dimensional empirical mode decomposition with CNN attained significant performance for five class classification [2]. Their approach included two non-polyp types as categories for classification. They did not test the method on a generalized dataset for performance evaluation. A fusion of color wavelet and CNN with feature selection was studied in Ref. [10]. They applied different classifiers for the classification of polyps from combined features from CNN and color wavelet transformed GLCM features. The fusion methods of [2,10] not specified their real-time application capability. Authors of [11] proposed a deep learning classifier that utilized uncertainty and calibration for a polyp classification on Australian colonoscopy dataset of five polyp classes. The performance of their method was computed only on a private dataset.

The previous polyp classification methods did not use the ensemble technology capabilities of many models for better performance. Besides, the preprocessing operations were not especially related to polyps in prior studies. This article presents a scheme to deal with the limitations of previous studies and achieve better performance advancements in real-time video processing.

3. Methodology

This research contributed to the development of a stacking group-based polyp classification system. A few preprocessing techniques highly related to polyp diagnosis have also been imposed to eliminate discrepancies from endoscopic images. A novel application of multi-attribute decision-making is inherited to process endoscopy video in real-time. The approach diagram is shown in Fig. 1.

3.1. Dataset

The efficiency of deep learning classifications depends remarkably on the well-transformed data collection. On contrary, most of the available datasets are not well-formed for training DL and ML models. Moreover, the evaluation of DL methods is well justified on their result on the publicly open dataset by comparing with other methods worked on the dataset. This study collected most of the data from public datasets. Mesejo et al. from the Department of Electronics, University of Alcala (http://www.depeca.uah.es/colonoscopy dataset/) provide a polyp classification dataset with labels [12]. This video dataset also contains the assessment of four experts and three novices endoscopist doctors. It contains twenty-one hyperplastic, fifteen serrated, and forty adenoma videos in both white light and narrow-band imaging techniques. We have utilized this dataset for training and testing our model. Non-polyp endoscopy videos are collected from Ref. [13]. Hereafter, this paper mentioned the Mesejo et al. dataset as database 1 (DB1) and non-polyp dataset as database 2 (DB2).

The training dataset (TrDS) contained 26,512 images of four classes adenoma, serrated, hyperplastic, and non-polyp. We have applied random splitting of dataset to define training and testing set. The random splitting was performed on video dataset thus TrDS and test set had completely distinct set of images. An imbalanced number of images of multiple classes may introduce biases to models. To resist the imbalanced dataset bias, the training dataset contains an equal number (6628) of images of each type. The model is tested using a portion of DB1 and DB2 that are not used for training (holdout test set). Also, we tested

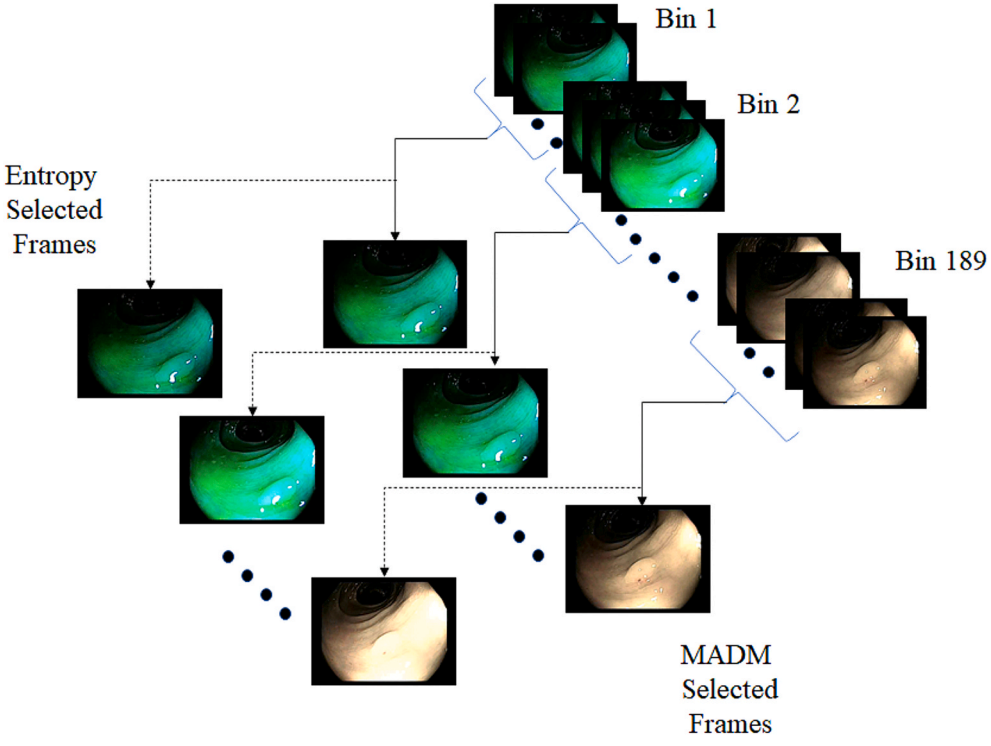


Fig. 2. Multi-attribute decision making technique-based frame selection.

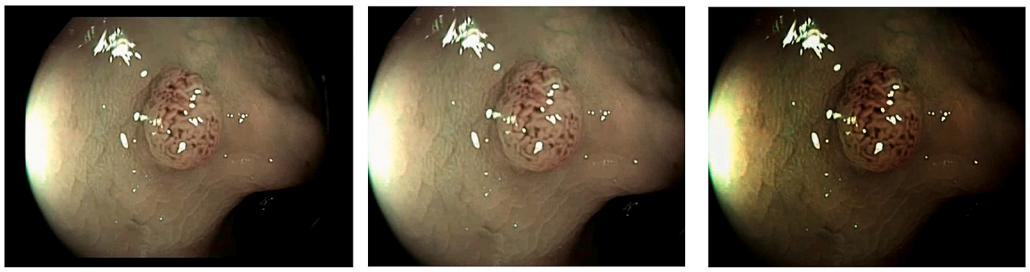


Fig. 3. Image preprocessing; (a) Original image frame, (b) Unnecessary part eliminated image frame and (c) Gamma corrected (0.5) image.

the models using k-fold (specifically 5-fold) cross-validation for precise evaluation of performance.

3.2. Preprocessing

The performance of deep learning models expectedly depends on domain-specific preprocessing operations. Colonoscopy videos are usually affected by light reflection, motion blur, camera angle, and other artifacts. These problems of colonoscopy can be removed or minimized by implementing preprocessing operations. This study proposes a way to remove light reflection from colonoscopy images. Image augmentation techniques are utilized to recreate different variations in endoscopy videos. Motion blur label is calculated to remove blurred images from the video for smooth processing of the model. Besides, a multi-attribute decision-making technique of frame selection from video is applied in this paper to reduce the processing time.

3.2.1. Multi-attributes decision making (MADM)

Processing each frame of a video to find polyps is inefficient. Also, this approach demands an enormous amount of time relative to the duration of a video. To optimize the processing time, we have utilized the multi-attribute decision-making (MADM) technique to find the most effective frame among some consecutive frames. We have formed bins where each bin contains fifteen successive frames. Then the MADM method is applied to each bin to select the suitable frame of that particular bin using specified attributes. Entropy, brightness, contrast, energy, and skewness are the attributes used to rank frames. Fig. 2 shows a visualization of the MADM based frame selection process. The decision-making process starts with attribute normalization using Eq. (1).

$$\textit{Attribute Normalization},\ \overline{V}_{ij} = \frac{V_{ij}}{\sqrt{\sum_{j=1}^{n} V_{ij}^2}} \qquad \text{Eq. 1}$$

Then the normalized values of the attributes are multiplied with the corresponding weight of each feature, and the positive and negative impact of each frame is calculated using Eq. (2) and Eq. (3).

$$\textit{Positive Impact}\ (PI)\ ,\ PI_i = \sqrt{\left[\sum_{j=1}^{m}\left(\overline{V}_{ij} - \overline{V}_j^{+}\right)^2\right]} \qquad \text{Eq. 2}$$

The $\overline{V}_j^{+}$ is the set of the best score of each normalized attribute scores in a bin. The best score can either be the maximum score of an attribute (where higher attribute score is expected such as contrast, brightness) or the minimum of an attribute scores (attribute with minimum score is better). Here, $\overline{V}_j^{-}$ is the set of attribute values taking the worse score of each normalized attribute. The worse score will be the minimum value of an attribute (such as contrast, brightness) if high value of attribute is better whereas the worse score will be maximum for reverse situations. $\overline{V}_{ij}$ is the normalized vector of *i*th frame of *j*th attribute. $\overline{V}_j^{-}$ value was used to compute the negative impact of a frame for *j*th attribute.

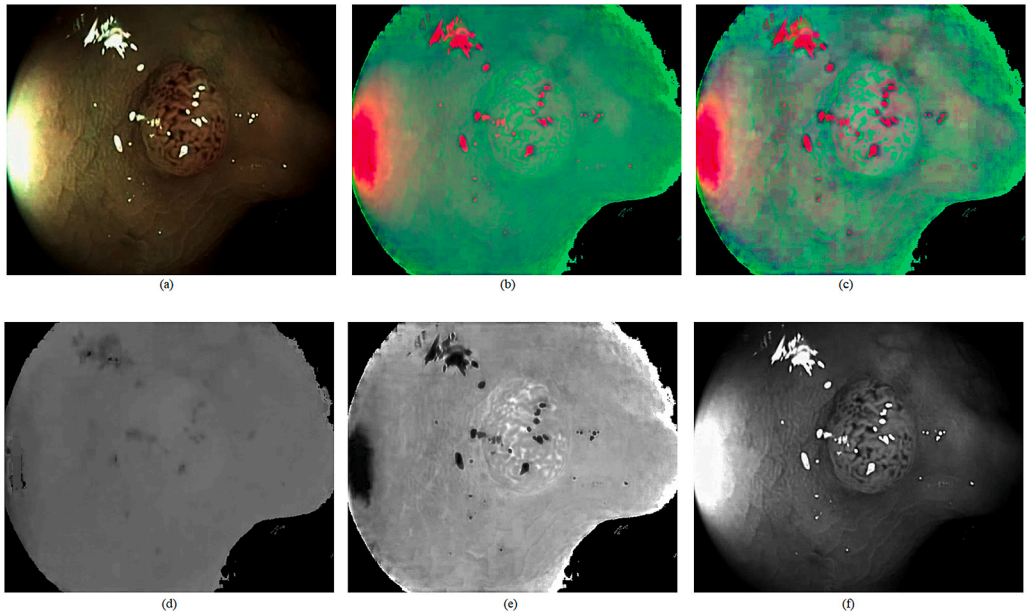


Fig. 4. Image pre-processing; (a) Gamma corrected (0.5) image, (b) HSV space image, (c) Adaptive histogram equalized HSV image, (d) Hue space, (e) Saturation space and (f) Value space of equalized HSV image.

(a) (b) (c)

Fig. 5. Mask of light reflections; (a) cropped input image, (b) image mask from HSV space image and (c) image mask after morphological operations.

$$Negative\ Impact\ (NI)\ ,\ NI_i = \sqrt{\left[\sum_{j=1}^{m}\left(\overline{V}_{ij} - \overline{V}_j^{-}\right)^2\right]} \quad \text{Eq. 3}$$

The decision of frame selection is calculated using Eq. (4). This equation calculates the goodness of each frame (goodness indicates the closeness to the ideal frame however there is no ideal frame in real-life). Then the best-valued frame is selected from each bin for further processing. This technique reduces processing time to one of the sizes of bin time.

$$Effectiveness\ Score\ (ES)\ ,\ ES_i = \frac{Positive\ Impact\ (PI_i)}{Negative\ Impact\ (NI_i) + Positive\ Impact\ (PI_i)} \quad \text{Eq. 4}$$

The values of ES_i always lies between 0 and 1 as the nominator PI_i presents on denominator as an addition with another positive quantity. Afterwards, the frame with the highest ES_i is selected for further processing steps. Here, the highest value of ES_i varies from bin to bin. Therefore, there is no particular optimum value for ES_i. It entirely depends on the frames of the bin. This frame selection process optimizes the required time for classification into very small compared to processing every frame for classification models.

3.2.2. Specular reflection removal

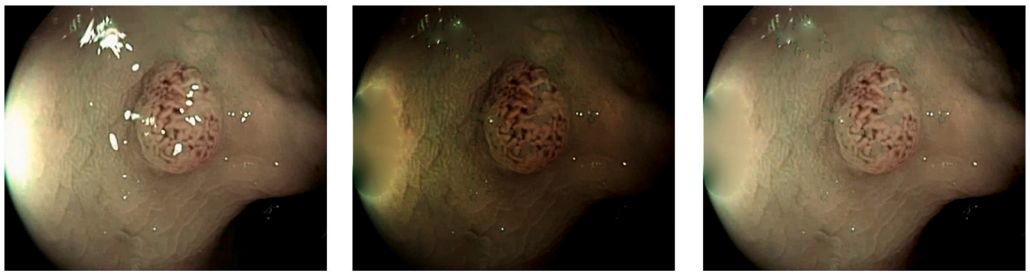
Specular reflection removal is a critical task due to the diverse scene of colonoscopy videos. The presence of specular reflection makes it more complicated to differentiate polyp regions [14]. This study used a sequence of actions to remove the specular reflection. A few prior studies used different reflection removal methods from colonoscopy images. Before performing reflection removal operations, we cut out the dark border areas from images. Herein we proposed a reflection removal scheme from colonoscopy images. The process starts with gamma correction using a gamma value of less than one (0.5). The gamma value was chosen using the trial-and-error technique. This gamma correction reduces the pixel value of pictures and makes them darker relative to the original image.

The gamma correction reduced the intensity value of image pixels and helped to find a better approximation for reflected lights. Fig. 3 shows the gamma-corrected image with an original image and the unnecessary part eliminates suitable image.

High contrast image works significantly well for the human visual system as well for neural networks [15]. We have applied the adaptive histogram equalization technique contrast limited adaptive histogram equalization (CLAHE) to magnify the contrast level of images [16]. The process works reasonably on single-channel images instead of original RGB images. We have used CLAHE on the V-space (value space) of the HSV (hue, saturation, value) image by converting the RGB image into HSV color space. Fig. 4 presents the CLAHE applied images of HSV color space and RGB color space. From the figure, it is evident that the S-space (saturation) and V-space are highly sensitive to reflection.

The high intensity is represented as dark in S-space while the opposite is true for V-space where high intensities are relatively whiter than dark intensity values. The representation of high intensities in the HSV color space component is utilized to find a reflection mask from images. We have calculated the mask by performing thresholding the S-space and V-space component values. The mask contains many scattered small areas. Some of those masked regions have a major impact on performance. We have applied morphological operations (MpOps) on mask elements to find the impactful reflected areas. The morphological process includes two erosions and three dilations. The elliptical shaped 5×5 structuring elements (SE) did well using iterative MpOps of erosions and dilations. The smaller size structuring element preserved the finely detailed images without losing important information. Additionally, the elliptical shape of SE captures the surface polyps significantly well. This iteration of erosions and dilations helps to remove smaller reflected areas and emphasize larger areas of reflection. Fig. 5 shows the mask of reflected areas of the image after morphological operations with the thresholded mask from HSV space images.

The masked areas of the image are inpainted with the pixel values of surrounding pixels. The authors of [17] have proposed a fast-marching algorithm for inpainting using the already known pixels of the image for mask regions. We have applied the algorithm proposed in Ref. [19] to inpainting the masked regions of the image. The algorithm used a heuristic technique of inpainting using available pixels starts from the boundaries and approaches other remaining mask pixels. Besides, the inpainting procedure takes very little time and effort compared with the generative adversarial network (GAN) based inpainting methods.



(a) (b) (c)

Fig. 6. Specular light removed images with original image; (a) Original image (b) Image after inpainting and (c) Gamma corrected image.

GAN-based methods need to train explicitly for each type of scenario and computationally expensive. Thus, the inpainting algorithm of [19] is well suited for this research. Fig. 6 shows the original image and reflection removed image after applying gamma correction on the inpainting image.

3.3. Model architectures

Deep models use a lot of computing power while producing output from inputs. A couple of decades ago, training a neural network model was a time-consuming complicated process. The advent of modern hardware manufacturing technology made it easier to build faster and efficient computing resources such as graphical processing units (GPUs) never than past [18]. This advancement brings enormous advantages to use more complicated and complex computation in a short span of time. This research proposes a stacking ensemble of three different CNN models for polyp classification, thereby we have selected three different types of network architecture for the ensemble. The trio contains a VGG Net (VGG19), Residual Net (ResNet-101), and Xception. The subsequent paragraphs describe the architecture of these models and the stacking ensemble procedure of polyp classification.

3.3.1. Finetuned VGG19 architecture

The VGG net, specifically the VGG19 network contains 19 trainable weighted layers in its architecture [19]. The block structure of the VGG net follows a combination of convolution and max-pooling operations. Fig. 7 shows the fundamental block structure of the VGG net. The VGG19 model architecture was initialized with the weights of ImageNet and utilized as a transfer learning model. However, the neural network part of VGG19 was redefined by implementing keras-tuner to find an optimal number of neurons for two fully connected layers (FCL). The keras-tuner randomly search for the different number of neurons within a predefined range of values to find the optimum number of neurons for FCLs. The optimum value for the VGG19 layer for polyp detection was 394 and 182 after iterating for fifteen random searches. The FCLs followed by dropout that limits the capabilities of the model to decrease overfitting [20]. The output layer of VGG19 contains four (4) units represents the four classes in particular adenoma, serrated, hyperplastic, and non-polyp. The activation function for the output layer was softmax, while the activation function for other FCLs was rectified linear unit (ReLU). The model was trained using Adam as an optimizer with varying learning rates for training started from 0.001 and dropped systematically by 0.1 after every ten epochs. The other two models (Xception and ResNet101) also the similar learning rate dropping technique throughout the training process. Early stopping was also implemented to stop training when the model outputs a constant validation accuracy score for three consecutive epochs [21]. The fine-tuned model was trained for 28 epochs before early stopping criteria stopped the training process. The training procedure also saved the best-performing model weights for future processing.

3.3.2. Finetuned ResNet-101 architecture

Models with a higher number of layers capture the more intuitive features of images and better mapping of input to output. However, training more layers is considerably tricky, and the introduction of more layers may result in poor performances on test data as the excessive parameters fail to map the training data outputs. Moreover, the addition of more layers arises problems like vanishing gradient which prevents effective training of the network by creating zero or nearly zero values during the backpropagation phase. The problem can be countered by adding skip connections over a number (2–3) of layers/hops of the network. The skipped connection or residual connection enables the network to skip training of few layers and prevents the problems like vanishing gradients and also helps to reduce overfitting. The residual

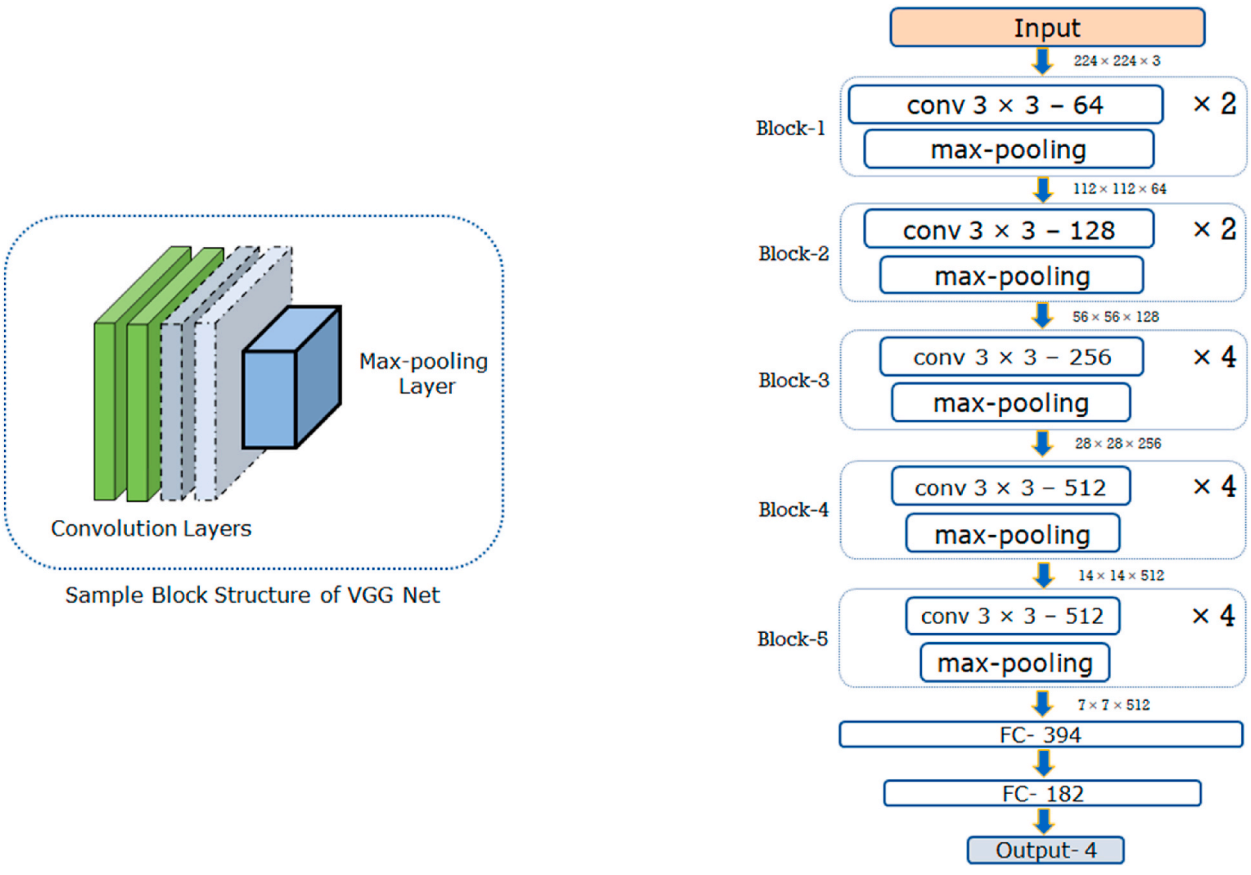


Fig. 7. Network architecture of finetuned VGG19; (a) Sample Block structure of VGG Net (b) Fine-tuned architecture of VGG19 for polyp classification.

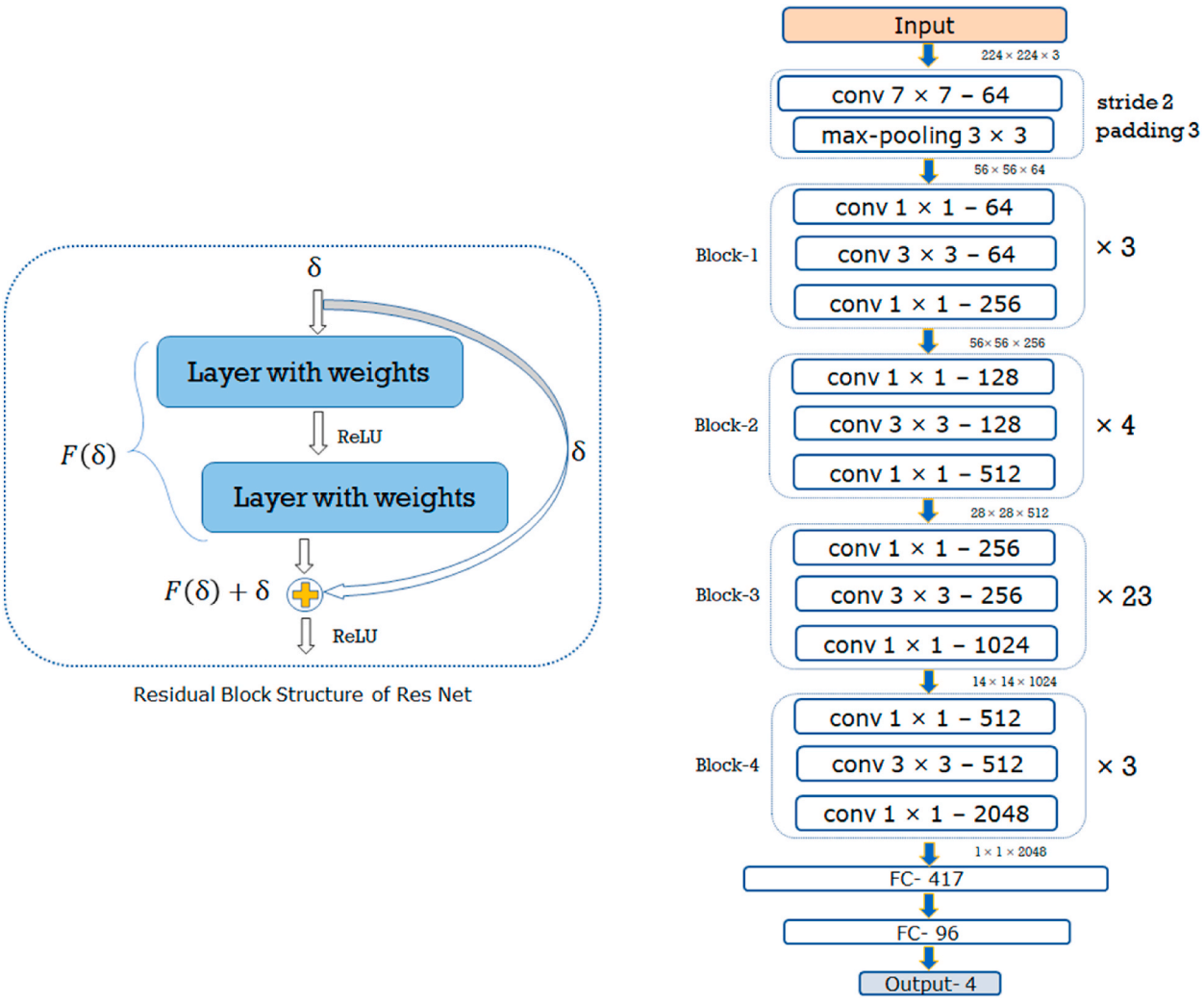


Fig. 8. The fine-tuned architecture of ResNet-101 (without residual connections within blocks) and simple residual block structure of ResNet architecture.

network trains by learning the residue between input and output. Fig. 8 shows a residual block where residual connection over two weighted layers. The output of the block is the summation of residue or difference (F(δ)) and input (δ) after pass through the ReLU activation function. The residual network uses the discussed learning paradigm which helps to train deeper networks with lower suffering from overfitting problems. The FCLs followed by dropout layer to limit the capabilities of the model to decrease overfitting [22]. This study used a pre-trained ResNet-101 model, and transfer learned the weights for polyp classification. The pre-trained weights were trained initially on the ImageNet dataset. The classifier head of the pre-trained network was redesigned with two new FCLs using keras-tuner with 417 units on the first FCL and 96 neurons on the second FCL. The other parameters and hyperparameters (such as, the activation functions, optimizer, loss function, initial learning rates and learning rate dropping) were same to the fine-tuned VGG19 architecture of section 3.4.1. The fine-tuned ResNet101 architecture was trained for 34 epochs before the early stopping mechanism terminates the process. The learning rate of the model also used the varying learning rate technique utilizing the scheduled dropping after every ten epochs. Fig. 8 shows the fine-tuned ResNet-101 without skip-connection for easier visualization.

3.3.3. *Finetuned Xception architecture*

Xception uses a modified hypothesis of the Inception network for its architecture formation [23]. The Xception model was also fine-tuned for polyp classification tasks similar to VGG19 and ResNet-101 models. The model learned and produced improved performance on the validation set for 37 epochs before early stopping forced to break the training loop. The other parameters were utilized similarly to the other two models. Fig. 9 shows a conceptual representation of the Xception model designed for polyp classification. The finetuned Xception architecture includes two fully connected of 286 and 127 neurons. The network also used dropout and batch normalization for training robust model.

3.4. *Stacking ensemble architecture*

Individual CNN model performs significantly well in different computer vision tasks and achieves state-of-the-art performances at them. However, the performance of a single model lacks in several ways such as fixed depth feature extraction, diversity of feature extraction process, difficulty to train models without suffering from under/overfitting. These lacking can be reduced or removed by introducing ensemble techniques [24]. The ensemble can present diversity in feature extraction procedures and helps to achieve better performance matrices.

The stacking ensemble training uses the validation dataset to train the classifiers. Training the classifier using training dataset was not feasible solution. The classifiers trend to overfit with large training data. Each CNN model outputs four different probabilities for each image. The probability values were then feed to classifiers for training and evaluated the performance using test dataset. The stacking ensemble uses the

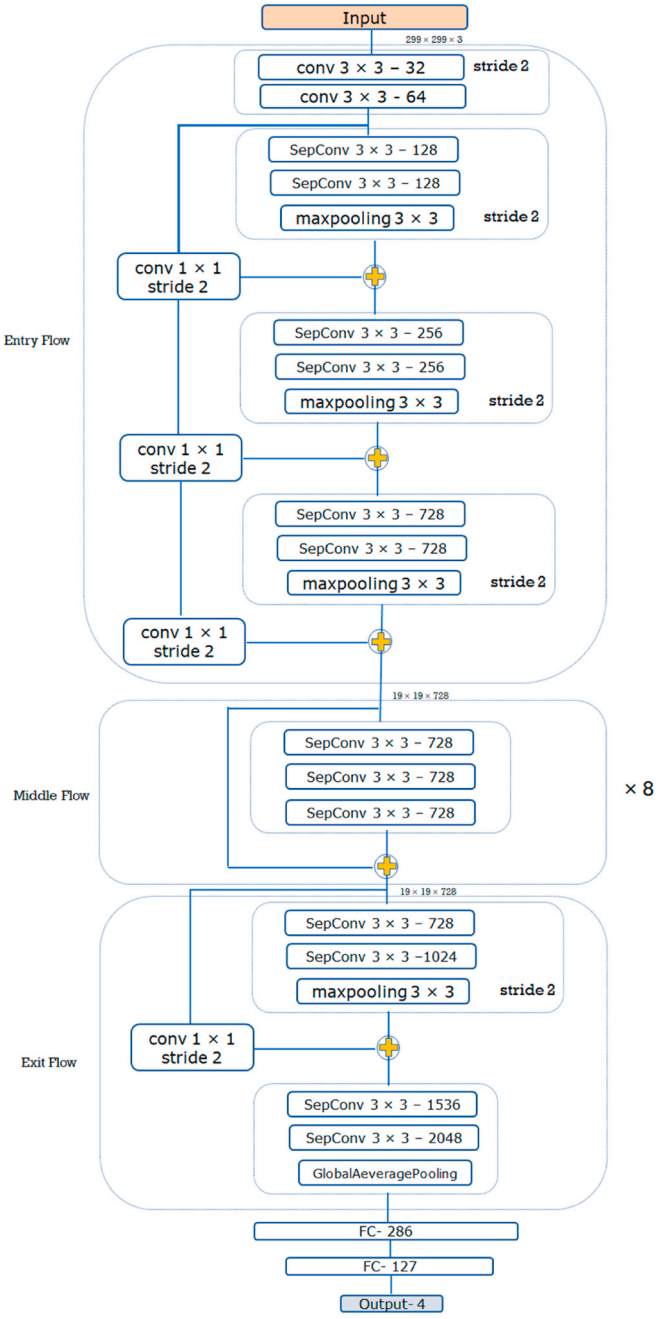


Fig. 9. The fine-tuned architecture of Xception model.

softmax output probability of these models. The stacking ensemble procedure is illustrated in Fig. 10.

The values of three fine-tuned models are then directed to the final layer of the ensemble classifier. We have experimented with different classifiers such as support vector machine (SVM), logistic regression classifier (LRC), decision tree classifier (DTC), and simple two layers NN. After tuned with different architectures, the two-layer hidden neural network with ten (10) and eight (8) neurons gave the best outputs. The output layer has four neurons for four classes with a softmax activation function. The hidden layers have applied ReLU activation to introduce nonlinearity. However, no dropout and regularization were used for training the network. The network was trained for 52 epochs with an early stopping technique before validation accuracy stuck.

4. Result and discussion

The experiments were performed on a cloud machine with 12 GB RAM; an Intel-powered CPU @ 2.30 GHz and Nvidia's Tesla K80 12 GB VRAM. We have implemented the models using python 3.7 with keras framework on TensorFlow 2.0. This section describes the results of our experiments and discusses the significance of metrics.

Section 3 has described the databases used to measure the performances of the proposed method. The augmentation operations on the image were performed to introduce diversity. The augmentation operations help to capture different varieties from the same image and resemble the variation of polyps artificially. Real-time video processing is achieved through the use of MADM technique. The MADM reduces the processing time significantly and selects the best frames for time-consuming operations. Hasan et al. proposed an entropy-based frame

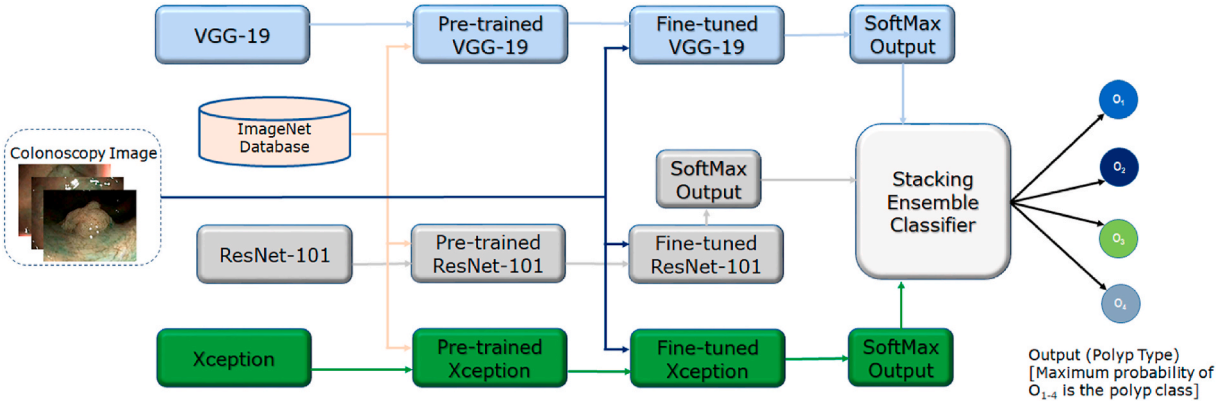


Fig. 10. Block diagram of stacking neural network ensemble technique using three different network architecture.

selection process however entropy-based selected frames comparatively less enlightening than MADM selected frames [13]. The processing time of MADM was slightly higher than the entropy-based technique while MADM selected considerably better frames.

The medical diagnosis method requires rigorous performance analysis to minimize the risk of false diagnosis. The wrong diagnosis encapsulates the false-negative and false-positive diagnosis instances. Herein, adenoma image diagnosis as serrated, non-polyp, or hyperplastic is an instance of false-negative for adenoma class images. Whereas, serrated, non-polyp, or hyperplastic image classified as adenoma is the example of false positive for adenoma. Fig. 11 shows the false positive, false negative, and true positive diagnoses for different classes. The accuracy is a general measure for any classification method. This study used six metrics to measure the performance of the proposed method on other public datasets. Eq. (5) to Eq. (14) mentioned the computing equation for accuracy, precision, sensitivity or recall, specificity, false diagnosis rate (FDR), and matthews correlation coefficient (MCC). Precision is a highly accepted criterion for diagnostic methods.

$$i^{th}\ class\ Precision,\ Precision_{C_i} = \frac{TP_{C_i}}{TP_{C_i} + \sum_{j=1}^{i-1} FP_{C_j} + \sum_{j=i+1}^{n} FP_{C_j}} \quad \text{Eq. 5}$$

We have used Eq. (5) to calculate the precision score of individual classes. TP_{C_i} counts the number of true positive instances of i^{th} class while FP_{C_j} represents the count for false positive instances of i^{th} class means j^{th} class polyp images predicted as i^{th} class polyp. Adenoma, non-polyp or serrated image classified as hyperplastic are the of false positive instances for hyperplastic class. For multiclass classification i takes the value from 1 to n. The value of i varies from 1 to 4 for this research. Eq. (6) illustrates the calculating formula for recall score of individual classes. Herein, FN_{C_j} represents the number of false negative instances of i^{th} class polyp. The number of i^{th} class polyp images predicted as j^{th} class polyp. For instances, hyperplastic images classified as adenoma, non-polyp or serrated are counted as false negative.

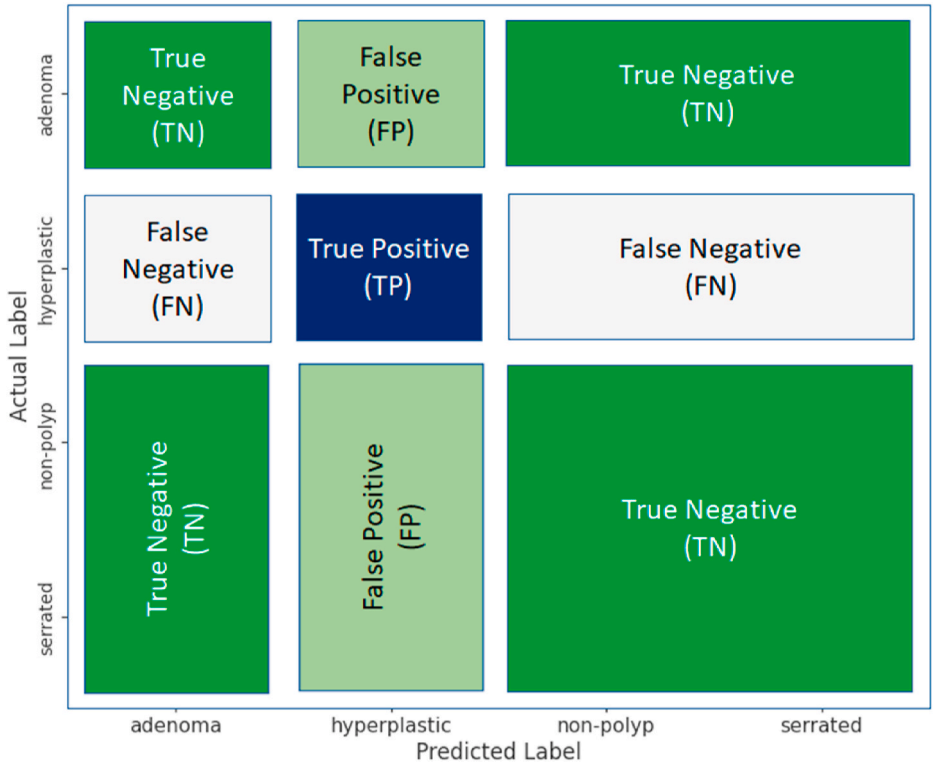


Fig. 11. Pictorial representation of classification measures for multiclass classification of polyps for an actual hyperplastic polyp image.

$$i^{th}\ class\ Recall,\ Recall_{C_i} = \frac{TP_{C_i}}{TP_{C_i} + \sum_{j=1}^{i-1} FN_{C_j} + \sum_{j=i+1}^{n} FN_{C_j}} \quad \text{Eq. 6}$$

$$i^{th}\ class\ Specificity,\ Specificity_{C_i} = \frac{\sum_{j=1}^{i-1} TN_{C_j} + \sum_{j=i+1}^{n} TN_{C_j}}{\sum_{j=1}^{i-1} FP_{C_j} + \sum_{j=i+1}^{n} FP_{C_j} + \sum_{j=1}^{i-1} TN_{C_j} + \sum_{j=i+1}^{n} TN_{C_j}} \quad \text{Eq. 7}$$

$$i^{th}\ class\ Accuracy,\ Accuracy_{C_i} = \frac{TP_{C_i} + \sum_{j=1}^{i-1} TN_{C_j} + \sum_{j=i+1}^{n} TN_{C_j}}{TP_{C_i} + \sum_{j=1}^{i-1} FP_{C_j} + \sum_{j=i+1}^{n} FP_{C_j} + \sum_{j=1}^{i-1} FN_{C_j} + \sum_{j=i+1}^{n} FN_{C_j} + \sum_{j=1}^{i-1} TN_{C_j} + \sum_{j=i+1}^{n} TN_{C_j}} \quad \text{Eq. 8}$$

Eq. (7) and Eq. (8) was used to score the accuracy and specificity of individual class. Where TN_{C_j} illustrates the true negative instances of predictions exemplifies j^{th} class images predicted not predicted as i^{th} class. The metrices of complete test dataset were calculated using Eq. (9) to Eq. (14) respectively to calculate accuracy, precision, recall, specificity, false diagnosis rate, and MCC.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Eq. 9}$$

$$Precision = \frac{TP}{TP + FP} \quad \text{Eq. 10}$$

$$Recall = \frac{TP}{TP + FN} \quad \text{Eq. 11}$$

$$Specificity = \frac{TN}{TN + FP} \quad \text{Eq. 12}$$

$$FDR = \frac{FP}{FP + TN} \quad \text{Eq. 13}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad \text{Eq. 14}$$

The performance of the ensemble model essentially depends on each CNN model's success and supports the lack of each model. Collectively CNN models perform better within the ensemble organization. The output of the model VGG19 on the test set is shown in Table 1. The test data set has pictures of four classes with different number of instances of classes. The model achieved an overall accuracy score of 92.37%. The network performed better at classifying hyperplastic and non-polyp images with precision score over 97%. The false diagnosis rate was highest for serrated type polyps. MCC score of serrated polyp images were the lowest of 79.07% and over 97% for hyperplastic and non-polyp images. The nature of adenoma and serrated polyps was responsible for the high false diagnosis rate and low precision of the serrated and adenoma class compared to other classes. Non-polyp images were classified correctly almost all cases. The VGG-19 has taken approximately 174 min for training on our system. The confusion matrix of fine-tuned VGG net model is in Appendix A.

Performance of ResNet-101 on test set is presented in Table 2. The model attained accuracy score of 93.97% which was slightly better than fine-tuned VGG19. There was performance improvement at classifying serrated polyps. Hyperplastic and non-polyp images classified properly with negligible error diagnosis rate below 1.5%. The model performed less false negative results for adenoma polyps which compromised the false diagnosis rate of less than 1% compared to VGG19 model. However, FDR of serrated polyp improved from 15.48% to 9.02% (the less value is better for FDR). ResNet101 has better MCC scores than VGG19 for all type of polyp images with minimum score for serrated and maximum for non-polyp images. Precision score of each class was alike for VGG19 and ResNet-101 model expect serrated category with an improved precision score about 6% for ResNet architecture. The training time of ResNet-101models was around 150 min. Appendix A presents the CM of ResNet-101 model.

The fine-tuned Xception model performed well compared to both ResNet and VGG networks. This finetuned model achieved accuracy of 95.90%. FDR score of has improved both each category with significant improvement for adenoma and serrated polyps. Xception model achieved better precision score each class than VGG19 model while less precision at hyperplastic detections than ResNet-101 model. Xception improves the serrated polyp MCC score significantly higher to above 90%. Table 3 presents the performance scores of fine-tuned Xception

Table 1
Performance of VGG19 on test set.

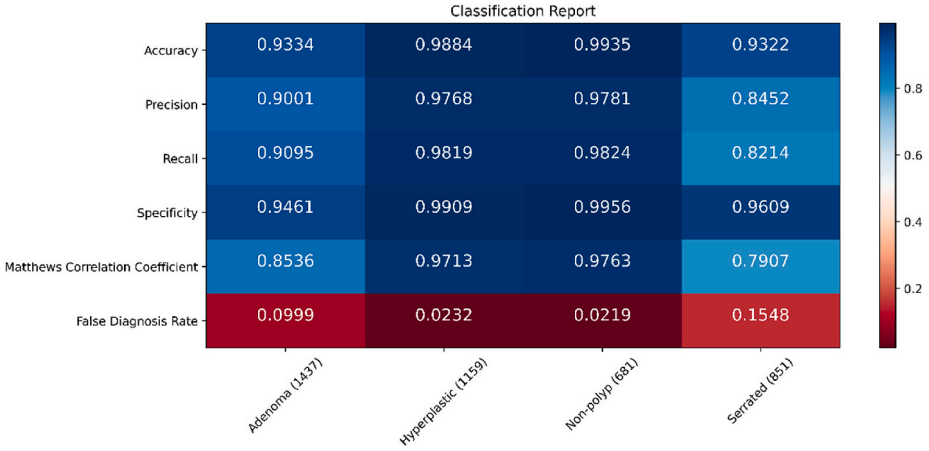


Table 2
Performance of ResNet-101 on test set.

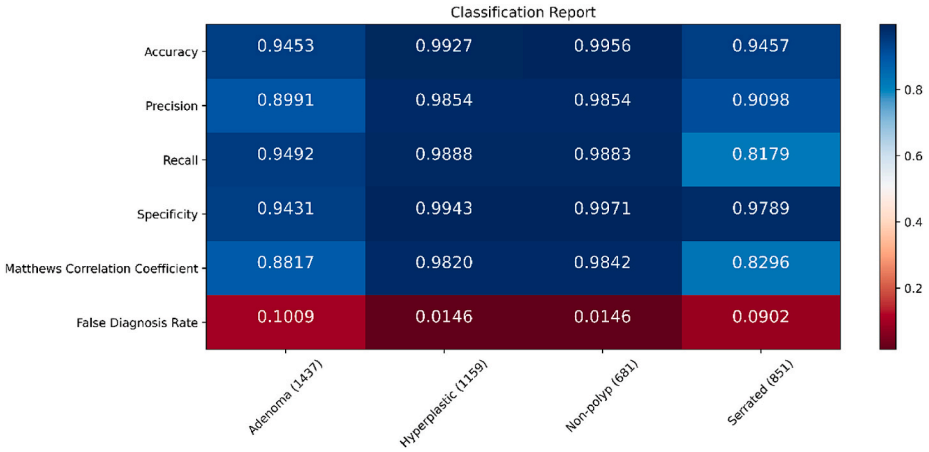


Table 3
Performance of Xception on test data.

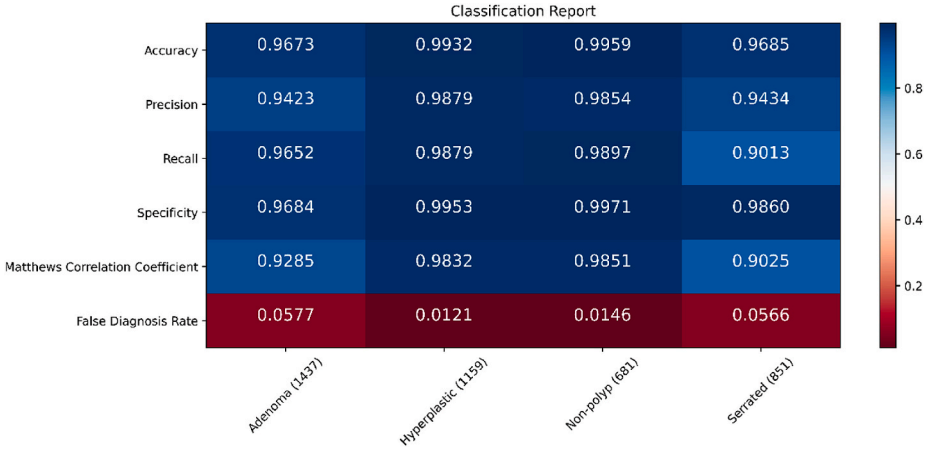
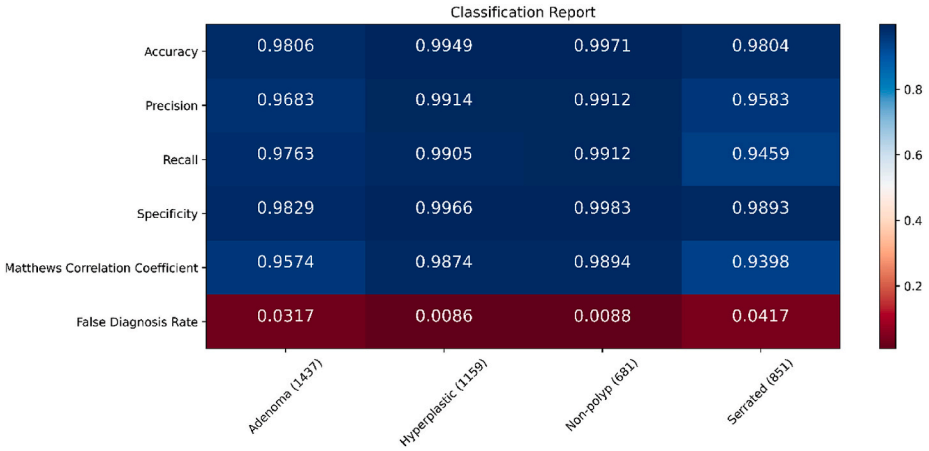


Table 4
Performance of stacking ensemble method on test dataset.



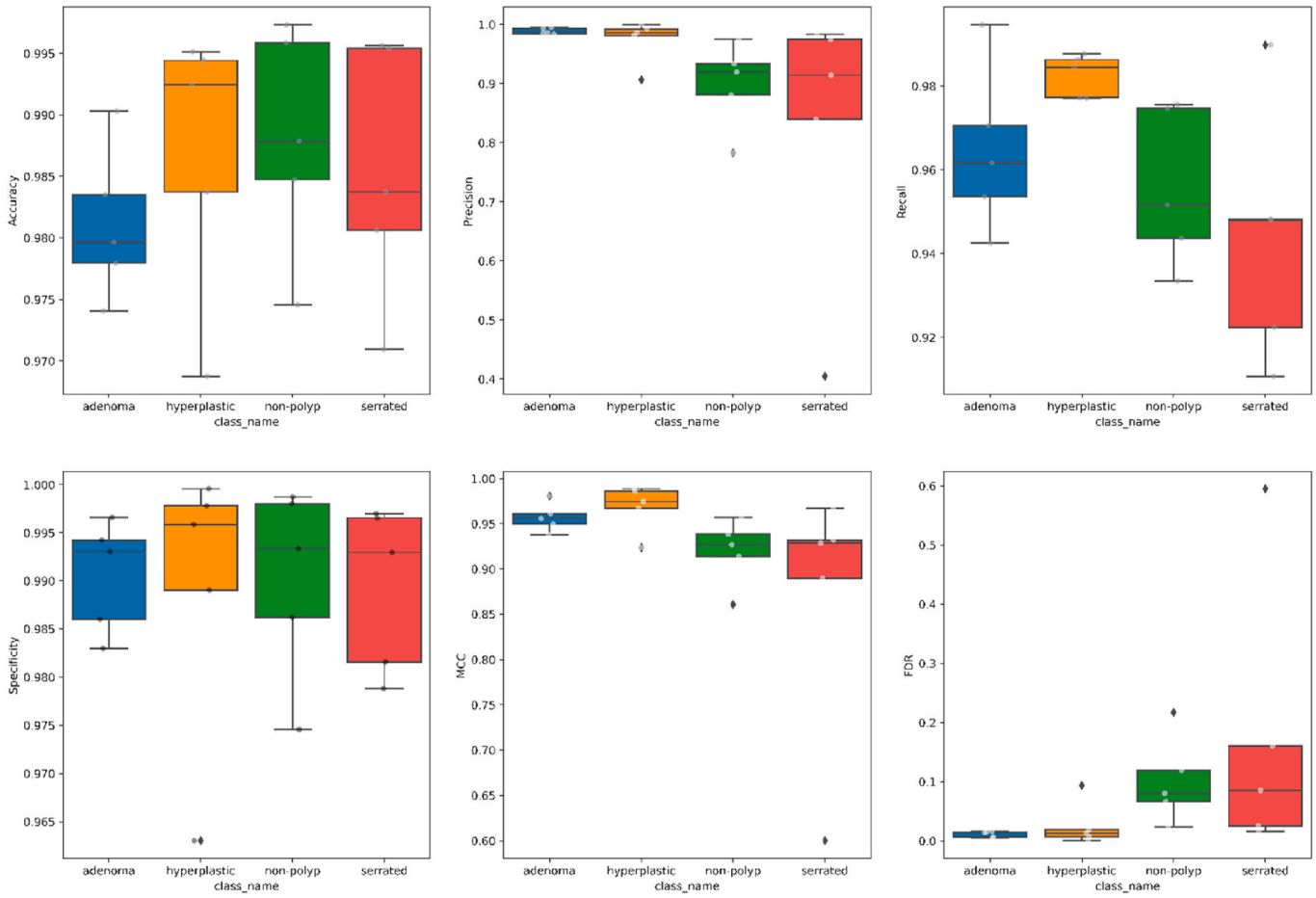


Fig. 12. 5-Fold cross-validated performance of each class.

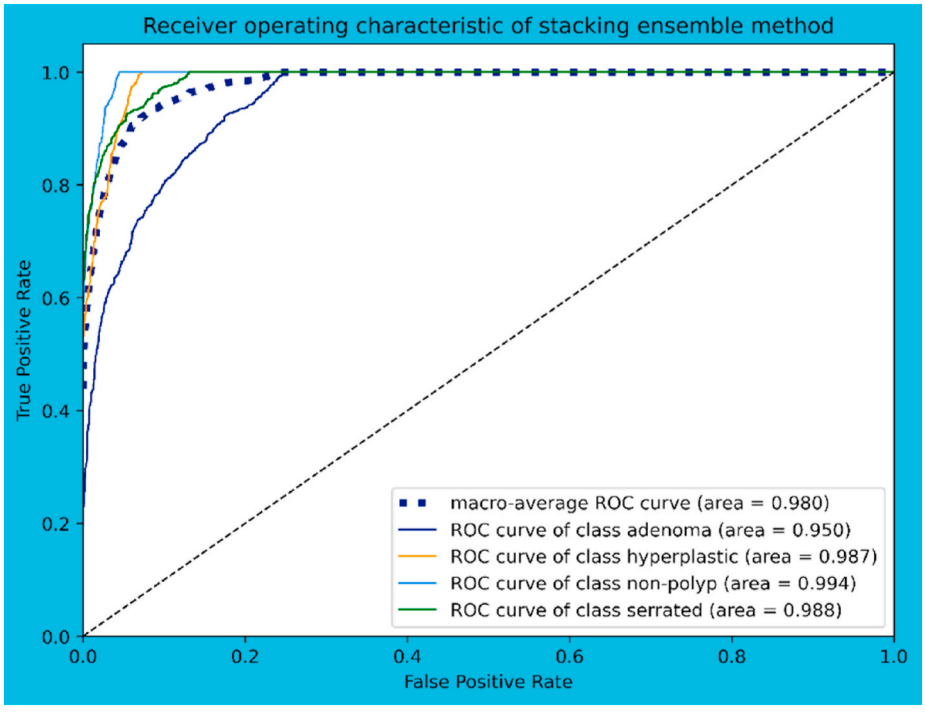
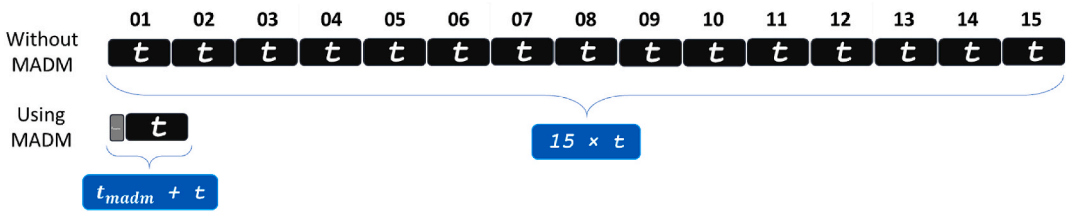


Fig. 13. ROC curve with AUC score different classes of ensemble method.

Table 5
Performance of our proposed method on test data with different polyp classification methods and gastroenterologists.

Method	Accuracy	Precision	Recall	Specificity	FDR
Expert 1	0.7544	0.6169	0.6160	0.7965	0.3830
Expert 2	0.7894	0.6456	0.6623	0.8374	0.3543
Expert 3	0.7631	0.6322	0.6688	0.8173	0.3677
Expert 4	0.7192	0.5666	0.5868	0.7727	0.4333
Beginner 1	0.7280	0.5867	0.5856	0.7872	0.4132
Beginner 2	0.7807	0.7272	0.5828	0.7920	0.2727
Beginner 3	0.6666	0.4772	0.4887	0.7482	0.5227
Mesejo et al. [12]	0.8246	–	0.7278	0.8588	–
Singh et al. [6]	0.9824	–	0.9619	0.9815	–
Ours	**0.9882**	**0.9772**	**0.9759**	**0.9918**	**0.0227**

Let,
t = processing time of proposed pipeline of ensemble method of CNN models
t_{madm} = processing time of frame selection using MADM technique
The processing time of t_{madm} is considerably really small compared to t, the processing time of ensemble pipeline.
Time needs to process a bin of 15 frames.
Without MADM:
$t_{total} = 15 \times t$
Using proposed MADM technique:
$t_{total\_madm} = t_{madm} + t$



network for four class polyp classification. Training Xception architecture took least time compared to other two models. The training time was approximately around 95 min. Appendix A includes the CM of Xception model.

The performance of ensemble method on test set is presented on Table 4. From Table 4, it is evident that stacking ensemble model performed significantly than individual models. The training time of the neural network classifier was around 7 min. CM of ensemble method is presented on Appendix A.

Metrices of the proposed model was also evaluated using 5-fold cross-validation (CV). The overall accuracy of the model with other metrics for adenoma, hyperplastic, non-polyp and serrated categories calculated to illustrate each class performance. Fig. 12 illustrated the 5-fold CV performances of individual classes.

From box-plots of Fig. 12 accuracy of adenoma was compressed within a range while other classes performance varies at different folds. The precision score of the CV folds was well distributed with less deviation of mean value. The other metrics also showed compactness of the method performance with a few outliers at box plot. The accuracy score of 5-fold validation was 98.53 ± 0.62%. The accuracy score was slightly better than hold-out test set of data. The method has less precision score compared to test set precision with 92.09 ± 4.62%. The recall and specificity values were similar to hold-out test set with 96.17 ± 0.87% and 98.97 ± 0.36% respectively. The FDR score was interesting with better performance at some folds while getting mean value of 7.91% with standard deviation of 4.81%.

The processing time of images using deep learning methods predominantly depends on the hardware the operation performed. Our proposed method takes approximately 1.34 s for processing each image on our computing machine. Since we have used MADM based frame selection technique which significantly reduced the processing time of videos. The MADM finds the best frame for performing the computation heavy processes of the ensemble method. Thus, the required time of our method for fifteen frames is constant and is equal to the processing time of a single image. The necessary time of ensemble method for every fifteen frames without MADM based frame selection is 19.38 s. The required time is 14 times higher compared to our proposed method using MADM. Our proposed method can process 670 frames per minute.

Receiver operatic characteristic (ROC) curve of the ensemble method is on Fig. 13. The figure showed the curves of multiple classes with the area measure of each curve. The higher the area under the better the model at differentiating classes with different threshold values. The curve is a representation of false positive rate (FPR) and true positive rate (TPR) at different threshold. The true positive rate increases with the increase of threshold. The lower threshold with higher TPR captures more area under the ROC curve (AUC). From figure, non-polyp curve has the largest area under it with a score of 99.4%. Hyperplastic and serrated polyp images have the AUC score of 98.7% and 98.9% respectively. The adenoma polyp was the hard class to differentiate with less AUC score compared to others covered 95% of the area. The macro average score of the ROC curve was 98%. The ROC curve and AUC of classes showed significance of ensemble method's performance improvement.

Table 5 presents a comparative performance metrics of proposed method with recent studies in polyp classification. Our developed method achieved improved performance than others. The performance metrices on Table 5 of our method was evaluated on hold-out test set. The accuracy of proficient gastroenterologist was 75.44%, 78.94%, 76.31% and 71.92% and trainees successfully classified polyps with an accuracy score of 72.80%, 78.07% and 66.67%. The mean accuracy of expert and beginner doctors was 75.65% and 72.51% respectively.

The overall accuracy score of seven doctors was 74.30% with a standard deviation of 3.91%. The doctors classified the images into three different classes however we have included non-polyp category. Mesejo et al. proposed method on DB1 has an accuracy score of 82.46%, SENS of

72.74% and PREC score of 85.88% [12]. Their method utilized traditional feature extraction methods with machine learning classification algorithms for classification.

5. Conclusion

The proposed stacking ensemble method for polyp classification with a novel frame selection method did significantly well for polyp classification. Our proposed and used techniques for preprocessing polyp images were highly associated with colonoscopy images and showed a substantial impact on classification. The multi-attribute decision-making technique reduced the processing time significantly and produced nearly real-time polyp classification with improved performance metrics. The stacking ensemble technique delivered a comparative better diagnosis of polyp classification than single CNN based classifications on public datasets. The ensemble model of three models helped to counter the lacking of a single model and boost the combined performance. This study will be exciting use to at real-time colonoscopy to help doctors. However, this research can be extended to segment the polyp regions from images to indicate the exact area of polyps.

Funding

There is no funding involved for this research.

Declaration of competing interest

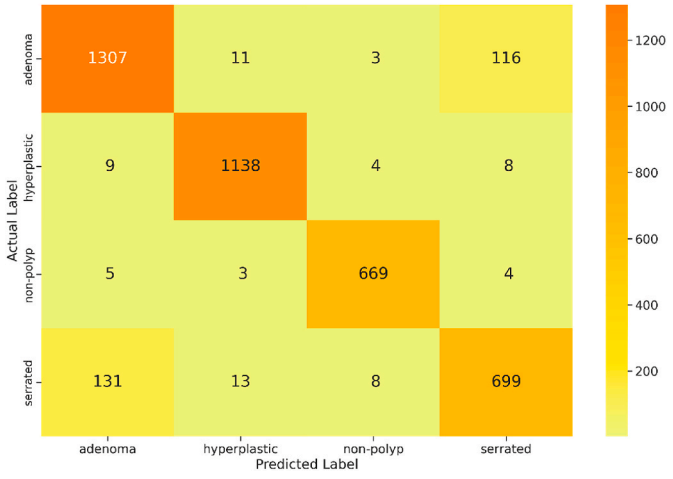
The authors declare that they have no conflicts of interest regarding to the publish the paper.

Acknowledgements

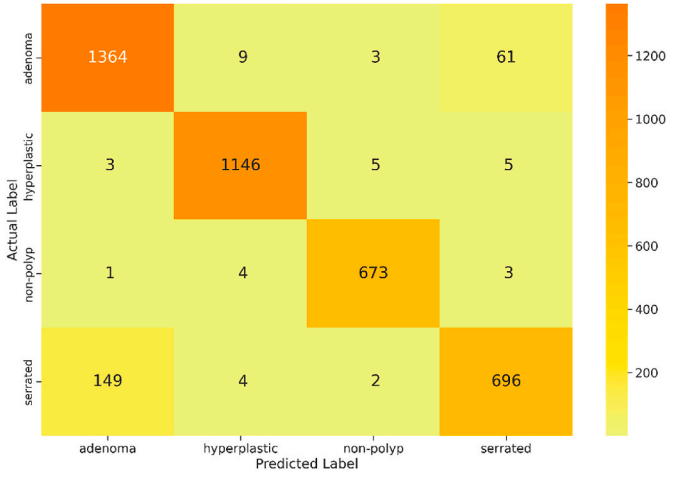
We are very thankful to Dr. Md. Kawsar Ahmed (Chief Medical Officer) and his team, Medical Centre at Mawlana Bhashani Science and Technology University for their invaluable supports and suggestions.

Appendix A

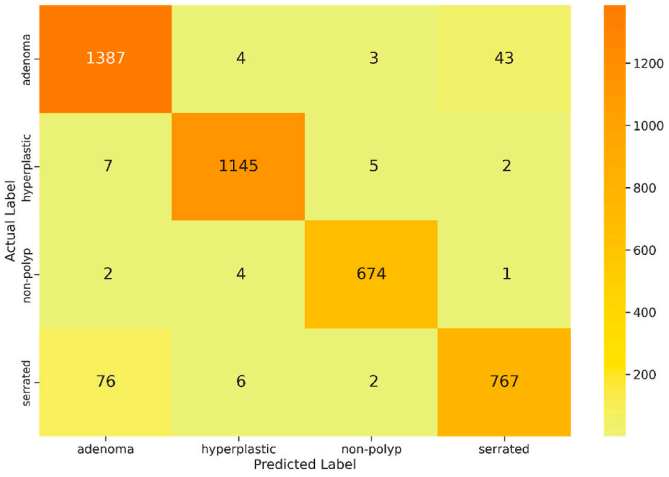
Confusion Matrix of fine-tuned VGG19 model on test data.



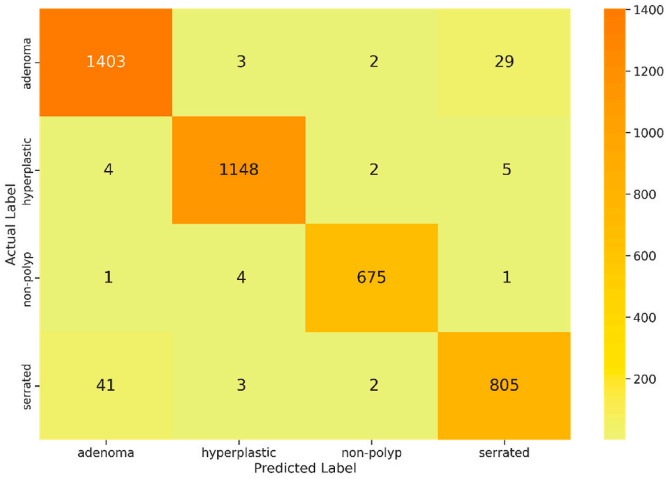
Confusion Matrix of fine-tuned ResNet-101 model on test data.



Confusion Matrix of fine-tuned Xception model on test data.



Confusion Matrix of fine-tuned Ensemble method on test data.



References

[1] Ljubic B, Pavlovski M, Alshehri J, Roychoudhury S, Bajic V, Van Neste C, et al. Comorbidity network analysis and genetics of colorectal cancer. Informatics Med Unlocked 2020;21:100492. https://doi.org/10.1016/j.imu.2020.100492.

[2] Mostafiz R, Rahman MM, Uddin MS. Gastrointestinal polyp classification through empirical mode decomposition and neural features. SN Appl Sci 2020. https://doi.org/10.1007/s42452-020-2944-4.

[3] Hoerter N, Gross SA, Liang PS. Artificial intelligence and polyp detection. Curr Treat Options Gastroenterol 2020;18:120–36. https://doi.org/10.1007/s11938-020-00274-2.

[4] Pacal I, Karaboga D, Basturk A, Akay B, Nalbantoglu U. A comprehensive review of deep learning in colon cancer. Comput Biol Med 2020;126:104003. https://doi.org/10.1016/j.compbiomed.2020.104003.

[5] Sun X, Zhang P, Wang D, Cao Y, Liu B. Colorectal polyp segmentation by U-Net with dilation convolution. Proc. - 18th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2019;2019. https://doi.org/10.1109/ICMLA.2019.00148.

[6] Singh D, Singh B. Effective and efficient classification of gastrointestinal lesions: combining data preprocessing, feature weighting, and improved ant lion optimization. J Ambient Intell Humaniz Comput 2020. https://doi.org/10.1007/s12652-020-02629-0.

[7] Mahmood F, Yang Z, Durr NJ, Chen R, Xu W, Borders D. Polyp segmentation and classification using predicted depth from monocular endoscopy. 2019. https://doi.org/10.1117/12.2513117.

[8] Zachariah R, Samarasena J, Luba D, Duh E, Dao T, Requa J, et al. Prediction of polyp pathology using convolutional neural networks achieves "resect and discard" thresholds. Am J Gastroenterol 2020. https://doi.org/10.14309/ajg.0000000000000429.

[9] Patel K, Li K, Tao K, Wang Q, Bansal A, Rastogi A, et al. A comparative study on polyp classification using convolutional neural networks. PloS One 2020;15:1–16. https://doi.org/10.1371/journal.pone.0236452.

[10] Billah M, Waheed S, Rahman MM. An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features. Int J Biomed Imag 2017. https://doi.org/10.1155/2017/9545920.

[11] Carneiro G, Zorron Cheng Tao Pu L, Singh R, Burt A. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. Med Image Anal 2020. https://doi.org/10.1016/j.media.2020.101653.

[12] Mesejo P, Pizarro D, Abergel A, Rouquette O, Beorchia S, Poincloux L, et al. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE Trans Med Imag 2016. https://doi.org/10.1109/TMI.2016.2547947.

[13] Hasan MM, Islam N, Rahman MM. Gastrointestinal polyp detection through a fusion of contourlet transform and Neural features. J King Saud Univ - Comput Inf Sci 2020. https://doi.org/10.1016/j.jksuci.2019.12.013.

[14] Sánchez FJ, Bernal J, Sánchez-Montes C, de Miguel CR, Fernández-Esparrach G. Bright spot regions segmentation and classification for specular highlights detection in colonoscopy videos. Mach Vis Appl 2017. https://doi.org/10.1007/s00138-017-0864-0.

[15] Badia S, Picchia S, Bellini D, Ferrari R, Caruso D, Paolantonio P, et al. The role of contrast-enhanced imaging for colorectal cancer management. Curr Colorectal Cancer Rep 2019;15:181–9. https://doi.org/10.1007/s11888-019-00443-1.
[16] Reza AM. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. J VLSI Signal Process Syst Signal Image Video Technol 2004. https://doi.org/10.1023/B:VLSI.0000028532.53893.82.
[17] Telea A. An image inpainting technique based on the fast marching method. J Graph Tool 2004. https://doi.org/10.1080/10867651.2004.10487596.
[18] Dai W, Berleant D. Benchmarking contemporary deep learning hardware and frameworks: a survey of qualitative metrics. Proc. - 2019 IEEE 1st Int. Conf. Cogn. Mach. Intell. CogMI 2019;2019. https://doi.org/10.1109/CogMI48466.2019.00029.
[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. 2015.
[20] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014.
[21] Bengio Y. Practical recommendations for gradient-based training of deep architectures. Lect Notes Comput Sci 2012. https://doi.org/10.1007/978-3-642-35289-8_26.
[22] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Comput Soc Conf Comput Vis Pattern Recogn 2016. https://doi.org/10.1109/CVPR.2016.90.
[23] Chollet F. Xception: deep learning with depthwise separable convolutions. Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 2017. https://doi.org/10.1109/CVPR.2017.195.
[24] Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics Med Unlocked 2019;16:100203. https://doi.org/10.1016/j.imu.2019.100203.