Multimedia Tools and Applications (2023) 82:18925–18946
https://doi.org/10.1007/s11042-022-14177-0

TRACK 2: MEDICAL APPLICATIONS OF MULTIMEDIA

An ensemble framework of deep neural networks for colorectal polyp classification

Farah Younas[1] · Muhammad Usman[1,2] · Wei Qi Yan[1]

Received: 22 November 2021 / Revised: 16 March 2022 / Accepted: 27 October 2022 /
Published online: 21 November 2022

Abstract

Colorectal cancer (CRC) is caused by malignant polyps which must be resected and examined for accurate classification. Biopsy, the manual workflow of polyp classification is time-intensive task and requires an automated solution. The objective of this study is to develop an accurate virtual biopsy tool for polyp classification. Moreover, automated assessment of polyps is a challenging task due to the similarities in their patterns, and in contrast to existing studies on binary classification, the outcome of multi-class classification requires evaluation through advanced evaluation measures. The proposed method combined the strength of individual weak learner for an accurate weighted-average ensemble deep learning classification. At first, base-classifiers were pretrained on the ImageNet database. Second, an average ensemble was built and evaluated for enhancing the performance, an appropriate combination of weights was chosen through grid search and assigned to the models. The performance evaluation of the proposed method in terms of F1-micro (0.80), F1-macro (0.81), F1-weighted (0.84) metrics, model reliability using Cohen's Kappa Coefficient (0.60) and Mathew Correlation Co-efficient value (0.49) for binary dataset shows the superiority over existing models. The higher rates of precision and recall show potential usage of the proposed system in the development of a virtual biopsy tool.

Muhammad Usman and Wei Qi Yan contributed equally to this work.

✉ Farah Younas
kpj7505@autuni.ac.nz

Muhammad Usman
musman@aut.ac.nz

Wei Qi Yan
weiqi.yan@aut.ac.nz

[1] Auckland University of Technology, 55 Wellesley Street East, Auckland CBD, Auckland 1010, New Zealand

[2] Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad 44000, Pakistan

Keywords Colorectal cancer deep learning · Prediction · Polypectomy · Transfer learning · Virtual biopsy

1 Introduction

Gastrointestinal (GI) cancer highly contributes towards cancer-related death toll globally. Colorectal cancer (CRC) represents almost 10% of overall cancer cases and has a very high mortality rate, particularly in developed countries [15]. In the United States, CRC is the third most occurring cancer in both men and women. In 2021, CRC accounted for 8% cases in men with a total number of 79,520 cases and 8% cases in women with total reported incidents of 69,983. Furthermore, the number of CRC's deaths reported in 2021 in US was 28,520 and 24,460 in men and women respectively [25]. Removal of precancerous lesions increases the chances of cancer prevention and survival rate is elevated to almost 100% [14]. Regular screening of high-prevalence infectious areas in patients can facilitate the early diagnosis and treatment before the patient becomes symptomatic and lesions becomes cancerous. Development and integration of Computer Aided Diagnosis (CAD) system for virtual biopsy in manual workflow is necessary. In addition, careful evaluation for clinical applicability of these systems is required [27].

In standard procedure, if a small polyp is detected, it is removed through fulguration (burning); or removed through snares (wire loop) or a biopsy instrument. Snare polypectomy technique is used for removal of large or pedunculated polyp. A wire loop through colonoscope is passed in the large intestine and polyp is removed from mucosal lining through electrical current. The main step observed is the preparation of histopathological slides of tumorous tissues for biopsy. Histopathology permits a precise diagnosis and provides better classification of polyp types [4]. However, preparation of glass slides for biopsy and visual identification under microscope is a time-intensive task. In addition, accurate classification of polyp type is highly dependent on pathologists' expertise and experience. The shortcoming of the manual procedure generates the need to develop an automated solution for virtual biopsy to improve the pathologist's decision. The automated solution for biopsy with advancement in Medical Image Analysis and DL approaches has started a new period of computer-aided pathology diagnosis [26].

Efficacy of colorectal lesion diagnosis is based on the Adenoma Detection Rate (ADR); defined as the percentage of colonoscopies with identification of at least one adenomateous lesion [11] Studies show that higher ADR is inversely related to lower CRC interval rates [10] and CRC related deaths [3]. Furthermore, sessile/flat polyps are recurrently missed as compared to larger and pedunculated polyps [12]. In order to improve the ADR, a spate of approaches have been followed. Using varying imaging modalities such as white light (WL) or narrowband imaging (NBI) endoscopy could be beneficial. Narrowband imaging is an advanced imaging technique that lay emphasis on the mucosal surface of the colon and capillary pattern of the polyp resulting in efficient polyp detection and classification [7]. Therefore, advancement in the development of CAD system can potentially improve the patient's diagnosis.

In recent times, artificial intelligence (AI) and deep learning (DL) have made major contribution in medical image analysis [1, 13, 24] and ADR is enhanced significantly through artificially intelligent systems. In colonoscopy, deep learning algorithms have also presented an increased utilization in detection, classification, segmentation and localization methods. Classification methods however are less advanced than detection methods due to lack of availability of large medical datasets [18]; Two of the major factors for further development of deep learning for endoscopy is the availability of high-quality endoscopy images and the increased understanding of technology by endoscopists [17].

In contrast to conventional ML approaches, DL algorithms do not require explicit feature definition. Instead, they utilize data and aggregate high dimensional features which are usually difficult to interpret for achieving the results. Owing to this advanced performance of DL algorithms, conventional ML techniques such as random forest, support vector machine etc. are becoming obsolete and are being replaced by DL approaches. Since these classic ML methods require manual designing and development of colon analysis model, they are not robust, very time consuming and lack flexibility. Therefore, real time applicability and success of these traditional ML models are sub-standard [26]. DL architectures in CRC serve various purposes such as classification in pathology images and polyp classification. Moreover, CNNs have gained a widespread usage in medical image analysis due to their enhanced performance regarding classification tasks.

In this paper, we propose a deep CNN-based heterogeneous weighted ensemble classification method for the analysis of endoscopy images of colon. The class imbalance problem is handled by data augmentation, including rotation, scaling, brightness and flipping of images which are further classified into adenomatous, hyperplastic and adenocarcinoma categories. In this regard, six CNN-based classifiers are trained independently to capture the discriminating features of polyps which are then combined to generate the final decision. The proposed classifier is evaluated of standard unseen test dataset. Block diagrams of proposed method are shown in Figs. 1 and 2. The following contributions are made in this paper:

- A heterogeneous ensemble is proposed for colorectal polyp classification in colonoscopy images. Generalization and robustness of the classification model are enhanced in the proposed ensemble learning by combining the strength of independent CNNs.

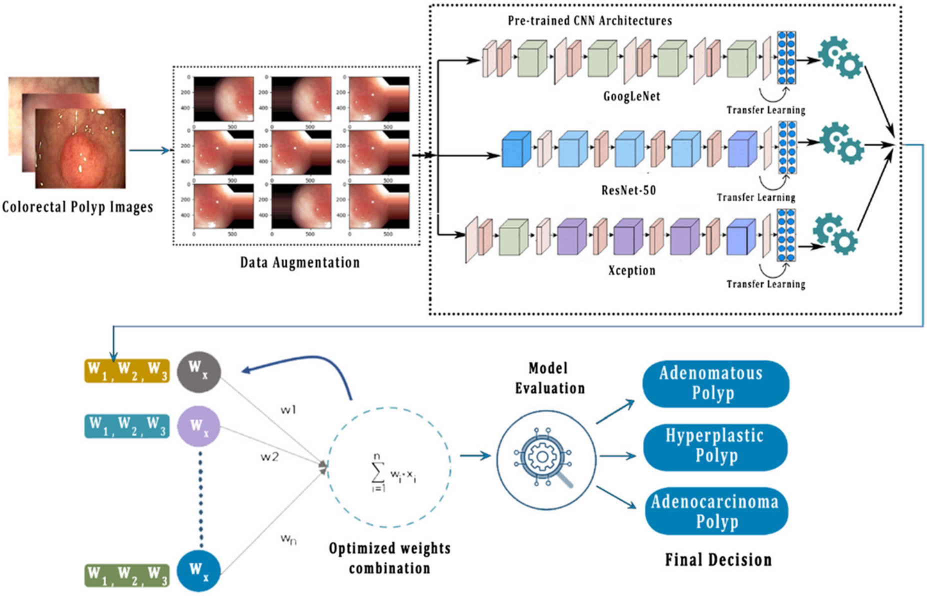

Fig. 1 An overview of the proposed weighted-average ensemble classifier

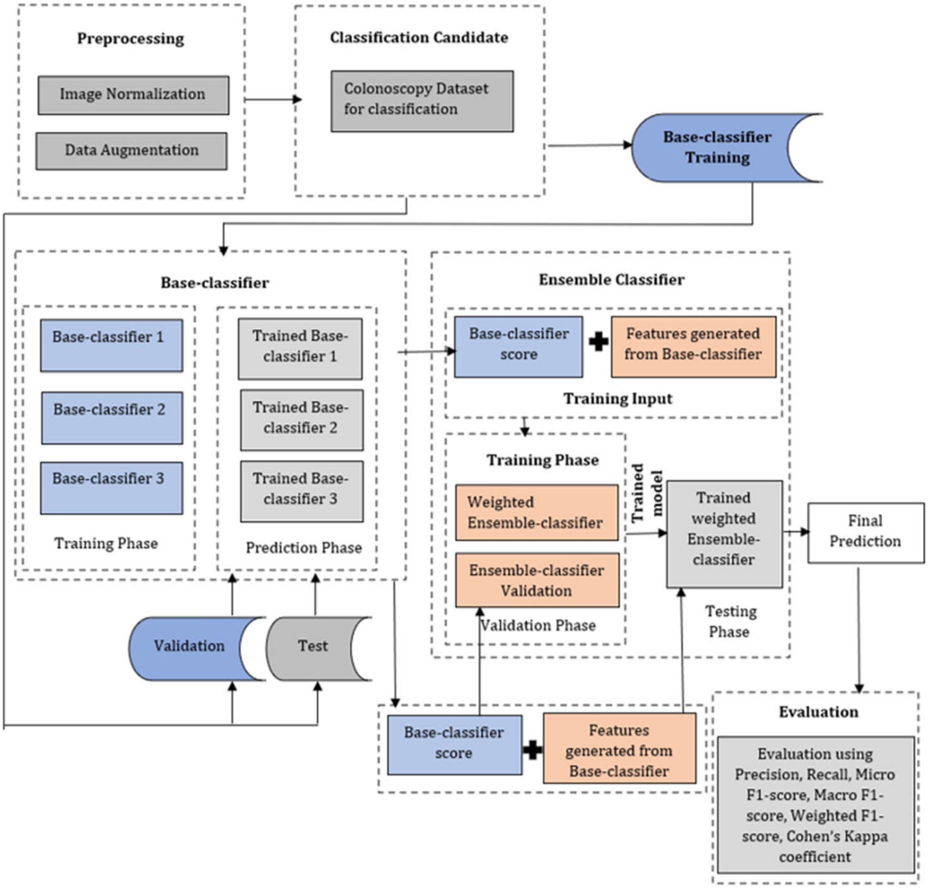
Preprocessing
Image Normalization
Data Augmentation
Classification Candidate
Colonoscopy Dataset for classification
Base-classifier Training
Base-classifier
Base-classifier 1
Base-classifier 2
Base-classifier 3
Training Phase
Trained Base-classifier 1
Trained Base-classifier 2
Trained Base-classifier 3
Prediction Phase
Ensemble Classifier
Base-classifier score
Features generated from Base-classifier
Training Input
Training Phase
Weighted Ensemble-classifier
Ensemble-classifier Validation
Validation Phase
Trained model
Trained weighted Ensemble-classifier
Testing Phase
Final Prediction
Validation
Test
Base-classifier score
Features generated from Base-classifier
Evaluation
Evaluation using Precision, Recall, Micro F1-score, Macro F1-score, Weighted F1-score, Cohen's Kappa coefficient

Fig. 2 An overview of the workflow for the proposed method

- The proposed weighted-average ensemble learning method considerably improves the classification results by assigning the weights to base-classifiers compared to averaging ensemble learning and the state-of-the-art CNNs.

The organization of the paper is as follows: In Section 2, we brief the related work, in Section 3 and 4, we elaborate the dataset and methodology along with implementation details. Results of the proposed method are presented in Section 5 and the paper is concluded in Section 6.

2 Related work

A deep learning model for the classification of polyps, namely, adenomatous polyps and serrated polyps was put forth [29]. The objective of this project is to reduce the cost and time of classifying the polyps, along with assisting the doctor for a more accurate diagnosis. The dataset having 5278 high-quality images was used for training and testing the proposed model.

The proposed CNN model consists of two modules: Base module and head module. The base module takes use of the Inception-ResNetv2 algorithm for automated feature extraction. Next, the head module of the algorithm is employed for transforming the extracted features to a grade scale which can further be applied for classification. The colorectal polyps in this project are classified in adenomatous and serrated polyps. Furthermore, the model is also compared under white light imaging and narrow banded imaging. The results unfold that there was no significant difference in the performance of the model based on white light and narrow banded imaging. The negative prediction for the fresh data was 97% and overall concordance was 94%.

Furthermore, an AI-based detection and classification of colorectal polyps were developed [19] which utilizes a deep neural network architecture. The algorithm is called Single Shot Multibox Detector (SSD), which determines its classes such as adenoma, hyperplastic polyp, sessile serrated adenoma/polyp, cancerous and other polyps after detected the polyp. The dataset for training and testing was acquired from 12,895 patients who underwent colonoscopies at Tada Tomohiro Institute of Gastroenterology and Proctology, Japan. Moreover, 16,418 images were adopted to train the CNN model, among which 3021 images were from polyp patients and 4013 images have normal colorectal. The processing time of CNN was 20 ms per frame. The trained CNN detected 1246 CP with a sensitivity of 92% and a positive predictive value (PPV) of 86%. The sensitivity and PPV were 90% and 83%, respectively; for the white light images, the rates are 97% and 98% for the narrowband images. Among the correctly detected polyps, 83% of the CP were accurately classified. Furthermore, 97% of adenomas were precisely identified under white light imaging. Lastly, the results of the developed system reveal that the accuracy of detection and classification is commendable and has great potential for such AI-based automated systems. However, the optimized hyperparameters are not used which can give better results.

In order to classify the polyps [20], five classes were classified: Adenocarcinoma, adenoma, Crohn’s disease, ulcerative colitis, and normal images. The data was collected from Gill Hospital that contained 3515 images. Furthermore, the KVASIR dataset consisting of 4000 images was also employed for validation of the proposed model. In the model, the deep layers have their spatial information preserved by using diluted convolution for better classification of polyps. Also, the architecture ResNet-50 was considered so as to avoid overfitting. Dropping blocks helps in the regularization of the model. The evaluation metrics include accuracy, recall, precision, and F1-score for evaluations. F1score related to the Colorectal dataset is 0.93 and the F1-score of the KVASIR dataset is 0.88. Lastly, the results of the proposed method are commendable, however, the model should have been compared with more architectures. A network-based transfer learning model was proposed for the improved classification of polyps. The dataset consists of 1000 instances which were collected from Gachen University Gil Hospital, during the colonoscopy of patients. The proposed method was compared to AlexNet along with different databases; Alexnet, Alexnet + SOS, AlexNet + ImageNet, AlexNet + Places, and the proposed method NIN+ ImageNet. Primarily, the Network is the stacking of multilayer perceptron consisting of multiple fully connected layers. Hence, its performance is better than CNN. The accuracy of the proposed method was 18.9%, more significant than AlexNet-based models. The recall rate was 0.92 ± 0.029, the AUC was approximately 0.930 ± 0.020. The measures depict that the proposed model was useful to assist doctors in classifying normal and abnormal polyps more accurately. However, other architectures such as ResNet, DenseNet, etc. should have been compared with the proposed model. Lastly, the classification of types of polyps can also be worked upon. A stacking ensemble method for better performance of polyp classification was proposed [21]. The dataset was collected from the University of Alcala, consisting of 26,512 images of four classes: Hyperplastic, serrates, adenoma, and non-polyp. Removing the reflections from images can hinder the performance of classification. Next, a frame selection method was also employed to reduce the processing time of the model. Lastly, a stacked ensemble learning was applied. The proposed method consists of three convolutional neural networks: Xception, ResNet-101, and VGG-19. The models were fine-tuned and then a softmax classifier was harnessed for the probable outcome of each model. Furthermore, two hidden layers of the neural network gave the best result with 10 and 8 neurons, with ReLU activation function in the hidden layers. The evaluation metrics include accuracy, recall, precision, specificity and AUC with scores 98.53 ± 0.62%, 96.17 ± 0.87%, 92.09 ± 4.62%, 98.97 ± 0.36%, and 0.9912, respectively. Hence, the proposed method performed better than single neural networks, however, more architecture should have been experimented with, for better decision making.

In a paper, Komeda et al. [28] benefited from object classification in Computer Vision to classify the polyps. In addition, in this paper, we combined computer vision and convolutional neural networks (CNNs) for precise classification. The proposed CNN-based CAD model accomplishes real-time image classification and achieved an accuracy of 0.75 with 10-hold cross-validation test. However, the work does not classify the hybrid polyp type serrated adenoma and hyperplastic polyps. Hyperparameter tuning would have contributed towards better model performance. Although the accuracy achieved by this model is not exemplary, CNN-CAD method is still a decent choice as it simplifies the operations and classification.

Based on Kudo classification, a method [23] to classify malignant and nonmalignant polyps, the dataset used in this paper is comprised of 600 images obtained 142 patients. Since the dataset was very small in size, data augmentation was performed to cater data insufficiency. The problem was tackled iteratively by implementing deep neural networks and the comparing the results with VGG-16 network. Evaluation metrics used to validate the results were accuracy, precision, recall and f1-score that achieved 83%,81%, 86% and 83% respectively. Next, the results were compared with other classifiers such as KNN and SVM. The outcomes of SVM and KNN with fifteen neighbors showed the same results. Though, this model proposed a better classification approach which could be further improved if a larger scale of training data is provided that could produce better results.

Another method [9] classified the five polyp images: Tubular adenoma, tubulovillous or villous adenoma, sessile serrated adenoma, and hyperplastic polyps. By using five family members of ResNet with 18, 34, 50, 101 and 152 layers, the dataset used in this paper is divided into 3 subsets: 326 training, 157 testing and 25 validation slides. Furthermore, additional 238 slides were collected from 24 institutes. The evaluation metrics to evaluate the performance of this model were accuracy, sensitivity, specificity. The purpose of this work was to compare the results of proposed method with pathologists' annotated results. Internal dataset achieved the mean accuracy 93.5% and pathologists attained 91.4%. Moreover, the external dataset reached the accuracy up to 87.0% whereas the pathologists obtained an accuracy 86.6%. One of the major issues with this study was data insufficiency which could have been handled using transfer learning and data augmentation. Finally, the results of the proposed model were close to the best, therefore, this model is applied to assist doctors and enhance the polyp diagnosis.

3 Materials

3.1 PICCOLO dataset

The PICCOLO dataset (PICCOLO RGB/NBI Image Collection, 2021) was acquired from Hospital Universitario Basurto, Spain. The dataset consists of clinical metadata and the annotated frames of colonoscopy videos. The frames during colonoscopy were captured through varying lightning technologies: white light (WL) and narrow band imaging (NBI).

- Metadata completed by gastroenterologist includes number of polyps of interest, current polyp ID, polyp size (mm), Paris classification, NICE classification, and preliminary diagnosis.
- Metadata completed by pathologists encapsulate final diagnosis and histological classification.

A systematic procedure was established to acquire the annotated dataset. Colonoscopy video clips were processed for extraction of individual frames. The frames excluded in process based on their lack of sufficient information were frames outside the patient, blurry images, high occurrence of bubbles, high existence of stool, transition frames between NBI and WI.

An analysis was conducted based on the captured frames to identify the type of lightning condition which is used to classify them as polyp or non-polyp images. One frame per second was manually annotated (i.e., one out of 25 frames). The frames were collected and revised to ensure the completeness of dataset.

3.1.1 PICCOLO dataset details

Colonoscopic video frames were recorded at Hospital Universitario Basurto, Spain between October 2017 and December 2019 using Olympus endoscopes (CF-H190L and CF-HQ190L) [22]. The dataset contains 3433 WL and narrow band imaging NBI images from clinical colonoscopy procedure videos in human patients. Total 46 patients were examined, and 76 lesions were included in the dataset. The data was distributed into three sets having 2203 images in training set, 897 in validation set, and 333 in test set. The details of frames in each set are given in Table 1. The dataset contains three classes of polyps: Adenoma, hyperplasia, and adenocarcinoma. Both Wl and NBI are used for the experimentation for proposed model.

Table 1 Frames in each of the sets according to clinical metadata

Dataset	Category	Items	Training Set	Validation Set	Test Set
PICCOLO	Image type	WL	1382	558	192
		NBI	821	340	141
	Diagnosis	Adenocarcinoma	172	166	127
		Adenoma	1552	592	92
		Hyperplasia	435	139	114
		N/A	44	–	–
CPDC	Diagnosis	Adenomatous	700	650	670
		Hyperplastic	400	300	330

3.2 CPDC dataset

The Colonoscopy Polyp Detection and Classification Dataset (CPDC) [6] is a collection of all publicly available endoscopic datasets MICCAI 2017, CVC colon DB, GLRC dataset and KUMC (Kansas Medical Center) dataset. The dataset consists of two classes: Adenomatous and hyperplastic polyps. The training data used for our experimentation includes 1100 training, 1000 validation and 1000 test images. The details of these frames in each dataset are provided in Table 1.

4 Methods

Colorectal polyp classification is a complex problem. Automatic classification using a deep learning network is challenging due to complex pattern of polyps. Single CNNs architectures do not give exemplary results alone however if the strength of individual weak learners in combination can improve the performance. Therefore, we propose a CNNs-based ensemble model for analysing the colonoscopy images. The main steps of the proposed method are: 1) Data augmentation; 2) Ensemble-based colorectal polyp classification. The proposed method is shown in Fig. 1, whereas the overview of workflow is shown in Fig. 2.

4.1 Data augmentation

Colorectal polyp image data from various patients has a high degree of imbalance distribution. Effective classification of images requires a balance between the classes. The cancerous polyp adenocarcinoma possesses carcinoma structure and the availability of such images in the dataset is very limited. Therefore, in order to handle this class, imbalance data augmentation is carried out including flipping, rotation, and brightness. Hyperplasic and adenocarcinoma classes were increased in number to maintain a balance between the three classes. In CPDC dataset, both the classes were augmented to increase the number of training samples.

4.2 Ensemble-based polyp classification

Our motivation is to effectively deal with the complex nature of polyps by improving the generalization of the classification system using ensemble learning. The proposed ensemble-based method exploits deep learning and has two training phases. In the first training phase, three deep pre-trained CNNs with varying architectural designs GoogLeNet, Xception, Resnet-50 are trained independently with ImageNet as base-classifiers. In the second phase of training, averaging-based ensemble learning is utilized for final classification of the input data. The weights from the base-classifiers weighted the averaged to make a final decision.

4.3 Improving generalization through ensemble learning

The motivation for adopting ensemble learning classification is to boost the generalization of the system. A learner may have a limited capability to capture the distribution of data; therefore, an aggregated decision of multiple weak-learners can improve the learning capability of classification system by overcoming the limitation of a single weak-learner. Ensemble learning draws a final decision from multiple diverse learners that may improve the robustness of the system.

A diversity of the base-learners for classification is the basis of ensemble learning which are incorporated in multiple ways. Usually, it is achieved by using: 1) A diversity of learning algorithms or with their configurations; 2) A multitude of features; 3) A group of training instances [5].

In the proposed ensemble learning, a diversity of methods is integrated by combining base learners to learn various features. An averaging method isused to combine the base learners and generate the final decision. Furthermore, weighted averaging method is also implemented, the suitable weights were assigned to the base classifiers for improved classification of task. There are three state-of-the-art deep learning models, namely, GoogLeNet, Xception, and ResNet-50 which are implemented as the benchmark of the proposed method. These CNN models include residual learning and vary in architecture, number of layers, block design.

Algorithm of proposed framework

```
          Inputs:

Step 1.   Training data $X = \{x_1, x_2, ..., x_n\}$
Step 2.   Hyperparameter search space $H = \{h_1, h_2, ..., h_n\}$
Step 3.   Number of trials $(N)$
Step 4.   Number of classes $(C)$
Step 5.   Number of models to be used in the ensemble $(M)$
Step 6.   Testing data $X' = \{x_1', x_2', ..., x_n'\}$
          Training Process:
Step 7.   Generate the set of N random hyperparameter combinations from $H$.
Step 8.   Create N different networks pertaining to $N$ combinations found in Step 7.
Step 9.   Train each pre-trained network $PT_{net}$ on $X$ from Step 8.
Step 10.  Choose the most efficient $PT_{nets}$ with specified accuracy threshold from step 9.
Step 11.  Perform grid search and assign suitable weights
          $w_t = argmax \sum_{i=1}^{n} w_1 , \dots , w_n$ to $PT_{net}$..
Step 12.  Perform the final training for the best-M networks selected from step 11 using entire training dataset.
          Testing Process:
Step 13.  Input testing image $x'$ from the $X'$
Step 14.  Generate output $h(x')$ predictions from each of the chosen $PTnet$ from Step 12.
Step 15.  Perform the final classification by doing an ensemble of predictions from Step 14.
          Output:
Step 16.  Final classification label $h(x')$ for a testing image from $X'$
```

4.4 Implementation

All our experiments were executed based on GPU-Based workstation with i7 processor, 16GB RAM, 1 TB HDD, 3GB GTX 1060 graphics card and Microsoft Windows 10 operating system.

All the CNN models were trained with stochastic gradient descent optimizer, the learning rate was set as 10^{-3} by using 100 epochs and the batch size was assigned as 32. Image resolution was adjusted according to the requirement of pre-trained base-classifier (224 × 224 and 299 × 299). Base classifiers were trained independently, the weights were saved. Averaging ensemble was incorporated for making a final decision. However, for further boosting the results, a variety of weights were assigned to the base classifier according to their performance. However, in order to further improve the results of classification, the weighted average was calculated. A grid search was conducted between [0, 0.5] as the weight to establish the best weight combination so as to maximize the result, assigning a higher weight value to the better classifier. The weighted ensemble assigned the generated weight values to the base classifiers and produced the improved final decision. There are two sets experimented by using the proposed model. The whole process was carried out for both original imbalanced data in first part of experiment and augmented dataset in the second part of the experiment.

Recall rate is considered as the evaluation metric for the classification model. Colorectal polyp classification is a class imbalance problem. Hence, the performance of individual and ensemble model is evaluated based on F1-score metric. F1-score gives an equal weightage to both precision and recall. Therefore, it is considered ideal for unbiased performance evaluation, especially for imbalance dataset as metric.

PICCOLO dataset has a variety of imbalance in distribution. The evaluation of imbalanced data results requires advanced metrics. Furthermore, this project aims at three classes classification. Therefore, in addition to accuracy, recall, precision, and F1-score, the proposed method was evaluated by macro F1-score and weighted F1-score.

Micro f1-score and macro f1-score exemplify two ways of confusion matrix in multiclass classifications. Confusion matrix of every class g_i, $i = 1,\cdots,k$ such that the i-th matrix takes g_i class as the positive class and rest of the classes g_j with $i \neq j$ being the negative classes. Micro average boosts the performance over all the samples, in other words, using the smallest number of units to compute overall performance. Micro-averaged F1-score is computed from micro-averaged recall Rmicro and micro-averaged precision Pmicro. The mathematical equations of these metrics are shown in (1), (2), and (3).

$$P_{micro} = \frac{\sum_{i=1}^{|G|} TP_i}{\sum_{i=1}^{|G|} TP_i + FP_i} \tag{1}$$

$$R_{micro} = \frac{\sum_{i=1}^{|G|} TP_i}{\sum_{i=1}^{|G|} TP_i + FN_i} \tag{2}$$

$$F1_{micro} = 2\frac{Pmicro*Rmicro}{Pmicro + Rmicro} \tag{3}$$

A large value of F1micro indicates a good overall performance of the model. Micro-average was misled for imbalanced data as it is not sensitive to the predictive performance of specific class. However, macro-average takes the averages over the individual class performance. Higher value of F1macro represents a good performance of individual classes. Mathematical equations are shown in Eqs. (4), (5), and (6).

$$P_{macro} = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{TP_i}{TP_i + FP_i} = \frac{\sum_{i=1}^{|G|}P_i}{|G|} \tag{4}$$

$$R_{macro} = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{TP_i}{TP_i + FP_i} = \frac{\sum_{i=1}^{|G|}P_i}{|G|} \tag{5}$$

$$F1_{macro} = 2\frac{Pmacro{*}Rmacro}{Pmacro + Rmacro} \tag{6}$$

$$kappa(k) = \frac{p_o - p_e}{1 - p_e} \tag{7}$$

$$W = \frac{\sum_{i=1}^{\mathrm{n}} w_i X_i}{\sum_{i=1}^{\mathrm{n}} w_i} \tag{8}$$

$$MCC = \frac{TN \times TP - FN \times FP}{(TP + FP)(TP + FP)(TN + FN)} \tag{9}$$

Cohen's Kappa coefficient shows the performance evaluation and reliability analysis in imbalanced class problem. In (7), p_O represents the overall model accuracy and p_e indicates the model prediction and actual class value by using chance agreement. The coefficients are interpreted as follows: No-agreement if values are less than 0, none-to slight agreement for 0.01~0.20, fair agreement when 0.21~0.40, moderate agreement is indicated by values between 0.41~0.60, substantial agreement for 0.61~0.80, almost perfect agreement is presented by 0.81~1.00 [16]. Weighted average is represented in Eq. (8). For binary classification, Matthew Correlation Coefficient (MCC) is taken into account which is used as a measure of the quality of binary classification given in (9).

5 Result analysis

In this paper, a weighted average ensemble-based approach is developed to successfully classify the colorectal polyp images as the classes Adenoma, Hyperplasia and Adenocracinoma. The results of on validation set and test set are shown in Tables 2, 3, 4, 5,

Table 2 Performance of the base-classifier and proposed ensemble model on imbalanced dataset

Models	Dataset	Accuracy	Precision	Recall	F1-score
GoogLeNet	Test	0.72	0.74	0.73	0.73
Xception		0.71	0.75	0.72	0.72
ResNet-50		0.73	0.75	0.69	0.73
Ensemble learning		0.73	0.78	0.73	0.76
Ensemble learning (Weighted Average)		0.74	0.78	0.73	0.75
GoogLeNet	Validation	0.73	0.68	0.77	0.75
Xception		0.70	0.61	0.79	0.70
ResNet-50		0.72	0.63	0.79	0.72
Ensemble learning		0.74	0.64	0.83	0.73
Ensemble learning (Weighted Average)		0.76	0.69	0.84	0.75

6, and 7. The ensemble learning approach shows a strong ability towards classification of polyp. The performance of the baseclassifier, average-based ensemble, and weighted average ensemble is validated based on validation set and tested on test set through evaluation metrics.

5.1 Performance analysis of base classifiers

The experiments were performed based on the original imbalanced dataset in the first phase. The results of base classifiers on validation and test set are shown in Tables 2, 3, and 4. F1-score measure is applied to evaluate the learning capability of base-classifiers. Recall and precision were considered for diversity analysis of the learners. All the base classifiers show the capability to learn in terms of F1-score ranging between 0.71 and 0.74 on validation set, whereas it stays constant for test set (0.73). However, the detailed analysis of results indicates that with the imbalanced dataset, the validation precision is quiet low, only 0.61.

As multiclass classification is accomplished in this paper, the data is not balanced, macro precision, macro recall, macro F1-score, and weighted F1-score were considered to evaluate the performance based on individual classes. The maximum value of macro F1-score achieved based on validation set was 0.70 by GoogLeNet and 0.73 based on test set by using ResNet-50 as shown in Table 4. In the second set of experiment, data augmentation was performed to handle the class imbalance, the results of the base-classifiers based on validation set and test set are shown in Tables 5, 6, and 7. All the base-classifiers show the capability to learn

Table 3 Sensitivity and specificity of the base-classifier and proposed ensemble model on imbalanced dataset

Models	Dataset	Specificity	Sensitivity	TP	TN	FP	FN
GoogLeNet	Test	0.73	0.70	255	593	84	107
Xception		0.70	0.69	265	598	86	110
ResNet-50		0.74	0.71	244	577	72	106
Ensemble learning		0.75	0.72	235	558	68	108
Ensemble learning (Weighted Average)		0.76	0.73	244	572	70	98
GoogLeNet	Validation	0.68	0.81	662	1559	314	156
Xception		0.65	0.73	622	1519	398	152
ResNet-50		0.68	0.75	642	1540	358	154
Ensemble learning		0.69	0.76	665	1562	372	92
Ensemble learning (Weighted Average)		0.70	0.78	695	1622	367	97

Table 4 Performance evaluation of multiclass imbalanced dataset

Models	Dataset	Macro Precision	Macro Recall	Macro F1-score	Weighted F1-score	Cohen's Kappa Coefficient
GoogLeNet	Test	0.71	0.71	0.72	0.72	0.57
Xception		0.69	0.68	0.69	0.70	0.52
ResNet-50		0.72	0.73	0.74	0.70	0.59
Ensemble learning		0.73	0.74	0.73	0.73	0.58
Ensemble learning (Weighted Average)		0.75	0.75	0.74	0.75	0.60
GoogLeNet	Validation	0.69	0.72	0.68	0.75	0.52
Xception		0.66	0.71	0.66	0.70	0.49
ResNet-50		0.67	0.72	0.68	0.74	0.47
Ensemble learning		0.70	0.78	0.73	0.74	0.54
Ensemble learning (Weighted Average)		0.71	0.75	0.71	0.76	0.56

successfully, and F1-score ranged from 0.71 and 0.73. However, for a multiclass classification, macro F1-score is more reasonable as it considers every individual class separately, the maximum value of macro F1-score was 0.75.

Table 5 Performance of the base-classifier and proposed ensemble model on augmented dataset

Models	Dataset	Accuracy	Precision	Recall	F1-score
GoogLeNet	Test	0.73	0.71	0.72	0.72
Xception		0.71	0.70	0.73	0.70
ResNet-50		0.72	0.72	0.72	0.73
Ensemble learning		0.76	0.77	0.74	0.73
Ensemble learning (Weighted Average)		0.80	0.81	0.82	0.80
GoogLeNet	Validation	0.69	0.69	0.68	0.68
Xception		0.73	0.74	0.73	0.73
ResNet-50		0.74	0.76	0.73	0.75
Ensemble learning		0.77	0.76	0.76	0.77
Ensemble learning (Weighted Average)		0.81	0.82	0.80	0.81

Table 6 Sensitivity and specificity of the base-classifier and proposed ensemble model on augmented dataset

Models	Dataset	Specificity	Sensitivity	TP	TN	FP	FN
GoogLeNet	Test	0.72	0.70	240	573	90	96
Xception		0.71	0.72	239	572	98	90
ResNet-50		0.72	0.73	242	575	90	92
Ensemble learning		0.78	0.75	253	591	66	89
Ensemble learning (Weighted Average)		0.82	0.81	270	603	61	65
GoogLeNet	Validation	0.70	0.69	975	2346	396	396
Xception		0.73	0.73	1018	2389	353	353
ResNet-50		0.74	0.74	1039	2410	332	330
Ensemble learning		0.76	0.75	1056	2427	315	315
Ensemble learning (Weighted Average)		0.81	0.80	1122	2493	249	249

Table 7 Performance evaluation of multi-class augmented dataset

Models	Dataset	Macro Precision	Macro Recall	Macro F1-score	Weighted F1-score
GoogLeNet	Test	0.71	0.70	0.72	0.71
Xception		0.70	0.69	0.70	0.70
ResNet-50		0.71	0.72	0.70	0.72
Ensemble learning		0.74	0.74	0.73	0.74
Ensemble learning (Weighted Average)		0.81	0.81	0.84	0.84
GoogLeNet	Validation	0.69	0.70	0.69	0.71
Xception		0.70	0.71	0.72	0.73
ResNet-50		0.72	0.73	0.75	0.76
Ensemble learning		0.75	0.74	0.76	0.75
Ensemble learning (Weighted Average)		0.80	0.79	0.79	0.81

Table 8 Performance of the base-classifier and proposed ensemble model on binary class augmented dataset

Models	Dataset	Accuracy	Precision	Recall	F1-score	MCC
GoogLeNet	Test	0.72	0.73	0.72	0.73	0.44
Xception		0.70	0.72	0.74	0.73	0.36
ResNet-50		0.71	0.73	0.75	0.74	0.43
Ensemble learning		0.73	0.73	0.73	0.73	0.45
Ensemble learning (Weighted Average)		0.75	0.76	0.76	0.76	0.47
GoogLeNet	Validation	0.71	0.72	0.71	0.71	0.44
Xception		0.69	0.68	0.70	0.69	0.34
ResNet-50		0.72	0.71	0.73	0.72	0.44
Ensemble learning		0.74	0.74	0.74	0.74	0.46
Ensemble learning (Weighted Average)		0.76	0.75	0.77	0.76	0.49

5.2 Performance analysis of the proposed classifier

Deep ensemble learning classifier is developed to effectively deal with complex structure of colorectal polyps. Virtual biopsy is a sensitive and complex task which requires accurate classification of polyps. Therefore, for improving polyp classification, the ensemble learning method was developed.

Table 9 Sensitivity and Specificity of the base-classifier and proposed ensemble model dataset binary class augmented dataset

Models	Dataset	Specificity	Sensitivity	TP	TN	FP	FN
GoogLeNet	Test	0.59	0.56	558	256	144	139
Xception		0.43	0.50	540	245	155	180
ResNet-50		0.43	0.50	545	245	145	140
Ensemble learning		0.58	0.62	548	258	142	134
Ensemble learning (Weighted Average)		0.62	0.63	564	263	138	132
GoogLeNet	Validation	0.59	0.59	538	249	141	138
Xception		0.64	0.65	526	210	180	140
ResNet-50		0.64	0.65	547	258	142	141
Ensemble learning		0.65	0.65	557	261	140	132
Ensemble learning (Weighted Average)		0.67	0.69	558	302	138	127

In case of imbalanced dataset, the achieved macro-F1 score for average and weighted-average ensemble models were 0.70 and 0.71 based on validation set, 0.73 and 0.74 based on test set, respectively on PICCOLO dataset. The results show that average ensemble learning does not improve the result in comparison to base-classifiers. However, the quantitative evaluation of average and weighted average ensemble classifier suggests that assigning the suitable combination of weights to the base-classifiers generates promising results and performs better than single base-learner and average ensemble model. In addition, the augmented data has shown better results 0.76 and 0.79 based on validation set, 0.76 and 0.84 based on test set on PICCOLO dataset, respectively. The results are shown in Table 7. Macro and weighted F1-score show that the base-classifiers were able to learn the complex representation of various polyp types. Similarly, for CPDC dataset similar trend is noticed, the accuracy achieved by proposed approach is 0.75 for test set and 0.76 for validation set as shown in Tables. 8 and 9.

The potential of multiple pre-trained CNNs with varying architectural design is evaluated for colorectal polyp classification problem. The performances of these classifiers do not produce exemplary results on colonoscopy images in contrast to the proposed technique (Tables 2, 3, 4, 5, 6, 8 and 9). However, the combined strength of weak learners has shown a considerable improvement in the results. Moreover, assigning the appropriate weights to the base learners significantly improves the classification of images. Figure 3 shows the F1-scorebased comparison based on imbalanced and original datasets.

The proposed method shows 2% increase in the macro F1-score on validation set and 3% on test for original data. However, 4% and 12% increase in macro F1-score on validation and test set was as noticed as compared to the maximum value attained by the individual base-classifiers. Moreover, the proposed weighted ensemble learning significantly improves the macro precision based on both validation set and test set as shown in Fig. 4.

5.3 Precision-recall based analysis

Figure 4 shows macro precision comparison of proposed approach based on both imbalanced and augmented dataset. In addition to sensitivity of model, it is extremely important to analyze the precision of the proposed system. Precision represents the correctly identified positive

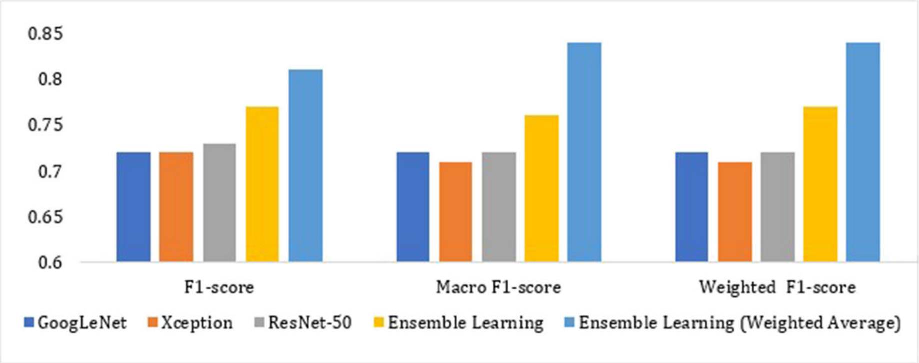

Fig. 3 Performance of base-classifier and ensemble classifier on augmented dataset

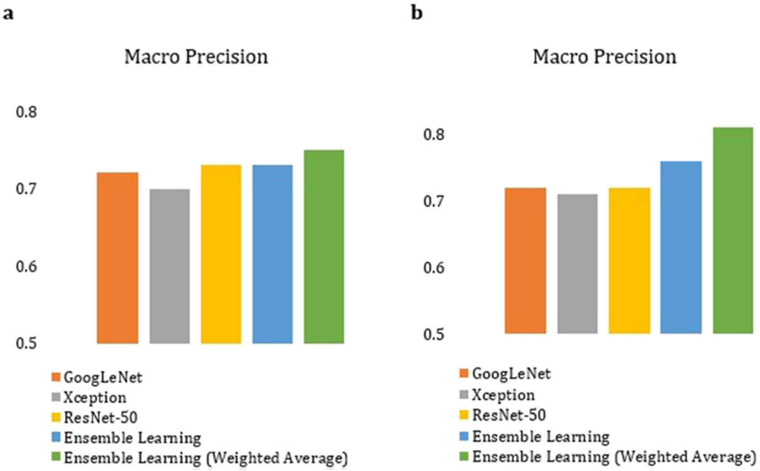

Fig. 4 **a** Precision based comparison of proposed model of imbalanced data and **b** augmented data

cases out of all the positive instance of the data. A small fraction of false positive values as shown in Fig. 5 can considerably affect the precision of the of the CAD system if the data is imbalanced and decreases the F1-score.

Figure 6 presents the performance comparison of CPDC dataset whereas Fig. 7 shows the False positive and Precision comparison of CPDC. In medical domain, where data is usually imbalanced, this misclassification can affect the system classification and have an adverse effect on diagnosis. The precision of the proposed system is 0.81 for augmented data which indicates a good capability of the system to identify positive cases.

Figure 8 shows a comparison of validation and test set with Kappa coefficient and error values. Kappa value for base-classifiers: GoogLeNet, Xception, ResNet-50 are 0.59, 0.55, 0.59.

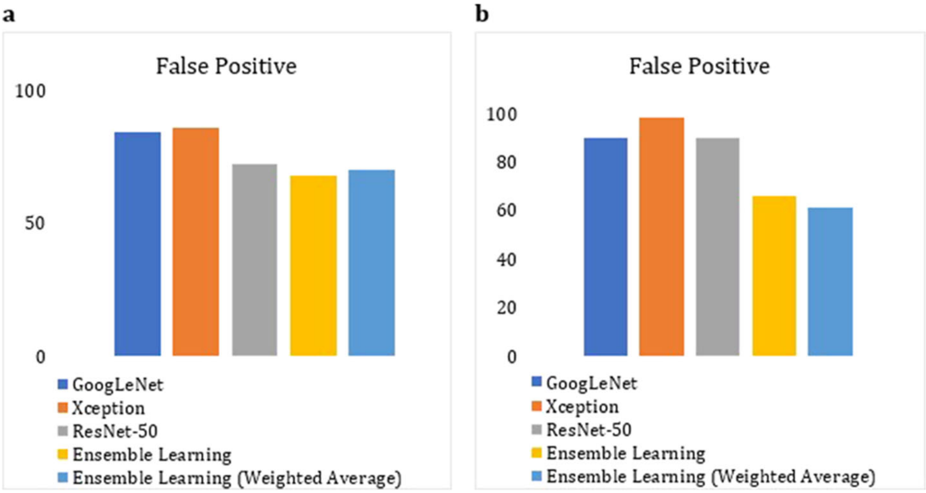

Fig. 5 **a** False positive rate-based comparison of proposed model of imbalanced data and **b** augmented data

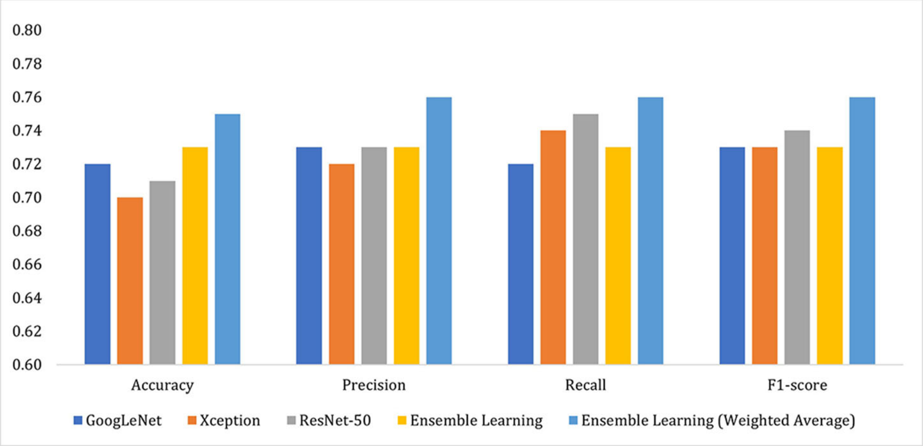

Fig. 6 Performance analysis of CPDC dataset

However, in terms of ensemble classifiers, average ensemble generates 0.61 (Kappa value) and weighted ensemble further improves the result to 0.62. Graph shows that with the increase in Kappa coefficient, error value of the model decreases in both scenarios. This significant increase in Kappa coefficient indicates a that proposed ensemble method has an acceptable degree of reliability. Figure 9 shows the ROC-AUC, 0.94 value that indicates that proposed model has good degree of separability. Fair evaluation metric for imbalanced binary dataset (CPDC dataset) used is MCC. Results shown in Fig. 10 presents an increasing trend in value of MCC that shows improved performance of proposed models compared to base classifiers. Table 10 shows the execution time and error of each architecture on both datasets employed in this study.

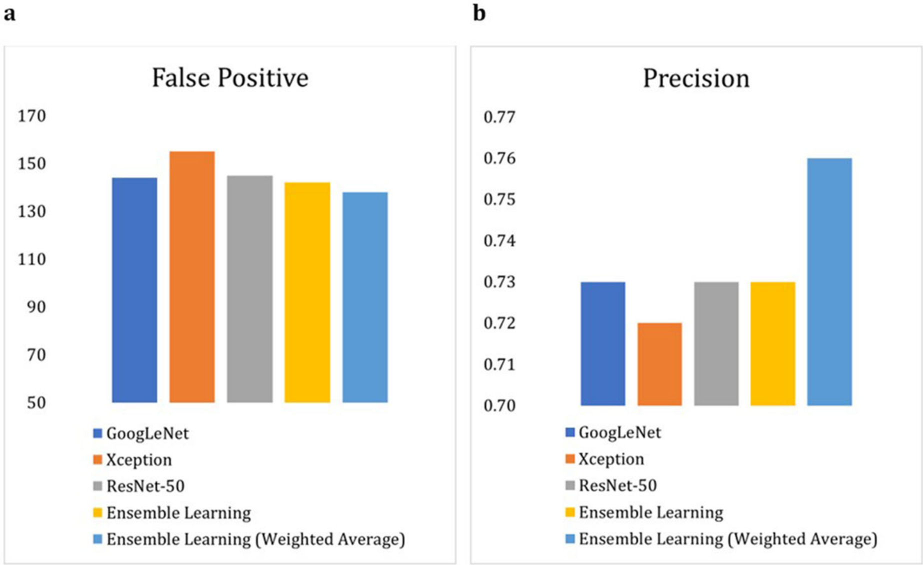

Fig. 7 **a** False positive and **b** Precision comparison of proposed model on CPDC augmented data

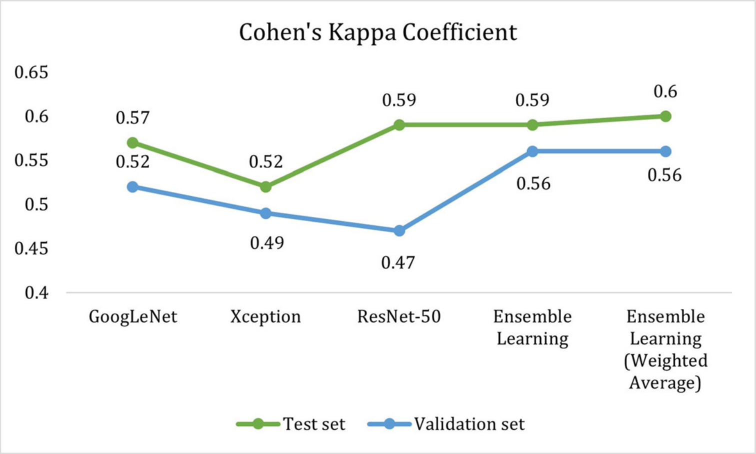

Fig. 8 Reliability comparison of model using Cohen's Kappa Coefficient

The experiments in this paper were conducted with various deep learning models for the classification of colorectal polyps, i.e., GoogleNet, ResNET50, ensemble learning, and weighted average ensemble learning. In the next, the results are compared with the published work with regard to classification of polyps using deep learning models. The highest accuracy we achieved is 82.8% by using CNN model proposed by Chen et al. [2]. Furthermore, AlexNet is employed as a backbone in transfer learning [8], which achieved the highest accuracy of 0.79 with the variations of fully connected networks. However, this method outperforms to the recent published work by achieving the highest accuracy of 96.3% and 90.5% on both balances and imbalanced data using Weighted Average Ensemble Learning. Comparison with existing recent studies is shown in Table. 11.

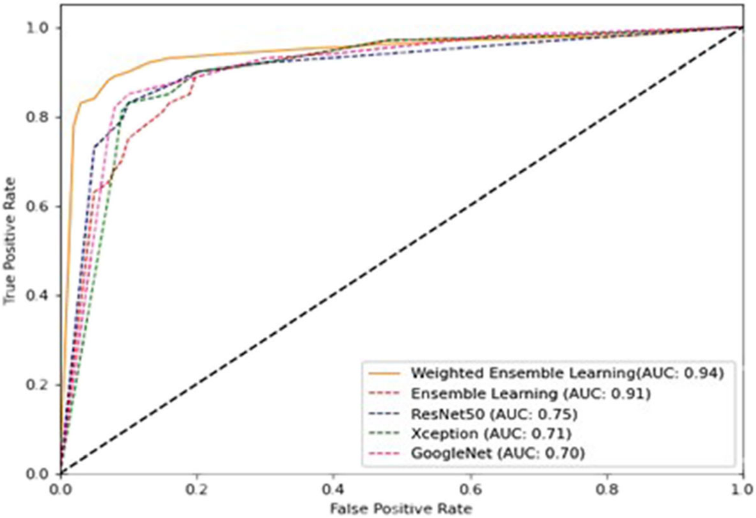

Fig. 9 ROC curves for the proposed classifier on test data

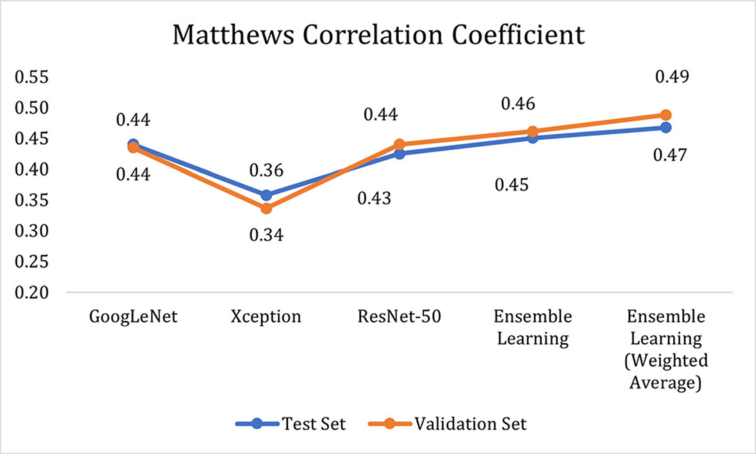

Fig. 10 MCC value comparison of validation set and test set

6 Conclusion

In this paper, we have presented a novel weighted average ensemble classifier to automate the colorectal polyp classification. Additionally, we have investigated how weighted averaging can significantly improve the classification capability of the model. The proposed classifier firstly takes advantage of ImageNet to train three CNN models separately, based on the augmented data to classify CRC images. Afterward, an average ensemble learning model is applied to make the final decision, where a grid search was performed to choose the optimum combination of weights to be assigned to individual base-classifier. The empirical evaluation of the model shows that appropriate weighted aggregation significantly improves the result.

The assessment of results shows that proposed method maintains a reasonable detection rate with a small deviation in macro F1-score. Among the base classifiers, GoogLeNet produced the lowest result of 0.69 macro F1-score. Xception and ResNet-50 gave slightly better results with 0.72 and 0.75 macro F1-scores respectively. The improvement in macro F1-

Table 10 Execution time comparison of datasets

Models	Dataset	PICCOLO Dataset		CPDC Dataset	
		Error	Execution Time	Error	Execution Time
GoogLeNet	Test	0.28	16 min 55 s	0.28	09 min 25 s
Xception		0.28	15 min 04 s	0.30	10 min 14 s
ResNet-50		0.27	16 min 03 s	0.29	12 min 22 s
Ensemble learning		0.23	21 min 55 s	0.27	14 min 53 s
Ensemble learning (Weighted Average)		0.19	22 min 34 s	0.25	15 min 04 s
GoogLeNet	Validation	0.29	22 min 11 s	0.29	10 min 13 s
Xception		0.26	19 min 31 s	0.31	13 min 33 s
ResNet-50		0.24	18 min 11 s	0.28	13 min 10s
Ensemble learning		0.23	23 min 04 s	0.26	14 min 02 s
Ensemble learning (Weighted Average)		0.18	24 min 12 s	0.24	15 min 18 s

Table 11 Comparative Analysis with existing studies

Data Source	Models	Dataset	Accuracy	Precision	Recall	F1-score
PICCOLO Dataset	GoogLeNet	Test	0.73	0.71	0.72	0.72
	Xception		0.71	0.70	0.73	0.70
	ResNet-50		0.72	0.72	0.72	0.73
	Ensemble learning		0.76	0.77	0.74	0.73
	Ensemble learning (Weighted Average)		0.80	0.81	0.82	0.80
	GoogLeNet	Validation	0.69	0.69	0.68	0.68
	Xception		0.73	0.74	0.73	0.73
	ResNet-50		0.74	0.76	0.73	0.75
	Ensemble learning		0.77	0.76	0.76	0.77
	Ensemble learning (Weighted Average)		0.81	0.82	0.80	0.81
CPDC Dataset	GoogLeNet	Test	0.72	0.73	0.72	0.73
	Xception		0.70	0.72	0.74	0.73
	ResNet-50		0.71	0.73	0.75	0.74
	Ensemble learning		0.73	0.73	0.73	0.73
	Ensemble learning (Weighted Average)		0.75	0.76	0.76	0.76
	GoogLeNet	Validation	0.71	0.72	0.71	0.71
	Xception		0.69	0.68	0.70	0.69
	ResNet-50		0.72	0.71	0.73	0.72
	Ensemble learning		0.74	0.74	0.74	0.74
	Ensemble learning (Weighted Average)		0.76	0.75	0.77	0.76
Children's medicine department of Gachon University Gil hospital in South Korea [2]	AlexNet+SOS, No transfer of fc6, fc7		0.706 ± 0.041	0.685 ± 0.077	0.807± 0.140	0.729 ±0.052
	AlexNet+SOS, Transfer of fc6		0.761 ± 0.062	0.770 ± 0.101	0.793 ± 0.139	0.766 ± 0.061
	AlexNet+SOS, Transfer of fc6 and fc7		0.782 ± 0.037	0.737 ± 0.061	0.893 ± 0.052	0.804 ± 0.027
	AlexNet+SOS, Addition of fc9		0.795 ± 0.045	0.766 ± 0.066	0.868 ± 0.069	0.809 ± 0.034
CVC-Clinic for training, CGMH-WL, CGMH-NBI for testing [8]	CNN model-NBI		0.82	0.82	0.95	0.81
	CNN model-WL		0.72	0.75	0.88	0.81

score (0.79) of weighted average ensemble from 0.73 of average ensemble classifier propose that developed method is suitable for multiclass classification task on imbalanced data. The utilization of non-biomedical ImageNet dataset to train the base-classifier also assisted in tackling the training need of data-hungry deep learning architectures. The model also proved to be reliable as evaluated using the Kappa coefficient and Mathew correlation coefficient.

Pre-trained networks were used in this study to provide the model an ability to accomplish the task with less computational effort to train base-networks. Moreover, less effort and knowledge are required to modify the proposed model according to the specific problem as grid search is used to find out the optimum combination of weights. The training phase of the developed method is computationally intensive as it has individual training phases for base classifiers. However, with the availability of GPUs, this is not an issue any longer. The use of sophisticated customized CNNs is sufficient to handle the complex tasks. In future, our goal to develop customized base-classifiers and explore innovative ensemble learning techniques along with Conditional GAN (CGAN) and further exemplify the performance of CAD tool for virtual biopsy.

We have proposed a weighted-average ensemble classifier for accurate classification of colorectal polyps as adenoma, hyperplasia, and adenocarcinoma. The performance of the ensemble-classifier with reasonable macro F1score (0.74) indicates the sufficient accuracy. The classification results of the proposed method presents that it outperforms the pre-trained CNNs with 4% improvement in terms of macro F1-score. Our comparison of averaged ensemble model and weighted ensemble model also shows a significant improvement in classification results, considering the macro F1-scores (0.75 and 0.76). Reliability analysis of the results is accomplished through Cohen's Kappa coefficient. A gradual increase in Kappa values (ranging from 0.55 to 0.60) of test data from pre-trained base classifiers to the proposed weighted ensemble learning classifier indicates there is an improved agreement between the raters and method is reliable. In order to utilize deep learning in colonoscopy, a plethora of processes are needed. The proposed method is to categorize polyps into three classes effectively which shows promising performance. In future, additional work in the CAD colonoscopy in terms of real-time treatment and polyp classification using deep learning is expected to be very beneficial.

Acknowledgments The authors would like to thank Basque Biobank who kindly provided us the access to their PICCOLO RGB/NBI database.

Declarations

Competing interests The authors declare no known potential competing interests with respect to financial interests or the research, authorship, and publication of this article.

References

1. Chan HP, Samala RK, Hadjiiski LM, Zhou C (2020) Deep learning in medical image analysis. In: Lee G, Fujita H (eds) Deep Learning in Medical Image Analysis. Advances in Experimental Medicine and Biology, vol 1213. Springer, Cham. https://doi.org/10.1007/978-3-030-33128-3.1
2. Chen-Ming H, Chien-Chang H, Zhe-Ming H, Feng-Yu S, Meng-Lin C, Tsung-Hsing C (2021) Colorectal polyp image detection and classification through grayscale images and deep learning. Sensors 21(18):5995
3. Corley DA, Jensen CD, Marks AR, Zhao WK, Lee JK, Doubeni CA, Zauber AG, de Boer J, Fireman BH, Schottinger JE, Quinn VP (2014) Adenoma detection rate and risk of colorectal cancer and death. New England J Med 370(14):1298–1306. https://doi.org/10.1056/NEJMoa1309086
4. Geboes K, Geboes K, Jouret-Mourin A (2013) Endoscopy and histopathology. Endoscopy 1:3–32. https://doi.org/10.5772/52739
5. Gomes HM, Barddal JP, Enembreck F, Bifet A (2017) A survey on ensemble learning for data stream classification. ACM Comput Surveys (CSUR) 50(2):1–36
6. Guanghui, W (2021) Replication Data for: Colonoscopy Polyp Detection and Classification: Dataset Creation and Comparative Evaluations, Harvard Dataverse, v1, https://doi.org/10.7910/DVN/FCBUOR
7. Ishaq S, Siau K, Harrison E, Tontini GE, Hoffman A, Gross S, Kiesslich R, Neumann H (2017) Technological advances for improving adenoma detection rates: the changing face of colonoscopy. Dig Liver Dis 49(7):721–727
8. Jae KY, Pyo BJ, Jun-Won C, Dong KP, Kwang GK, Yoon Jae K (2021) New polyp image classification technique using transfer learning of network-in-network structure in endoscopic images. Sci Rep 11(1):1–8
9. Jason WW, Arief AS, Vaickus LJ, Bing R, Xiaoying L, Mikhail L, Naofumi To, Behnaz A, Adam SK, Dale CS (2020) Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. JAMA Netw Open 3(4):e203398–e203398

10. Kaminski MF, Regula J, Kraszewska E, Polkowski M, Wojciechowska U, Didkowska J, Zwierko M, Rupinski M, Nowacki MP, Butruk E (2010) Quality indicators for colonoscopy and the risk of interval cancer. N Engl J Med 362(19):1795–1803. https://doi.org/10.1056/NEJMoa0907667
11. Kaminski MF, Thomas-Gibson S, Bugajski M, Bretthauer M, Rees CJ, Dekker E, Hoff G, Jover R, Suchanek S, Ferlitsch M, Anderson J (2017) Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. Endoscopy 49(04):378–397
12. Kim NH, Jung YS, Jeong WS, Yang HJ, Park SK, Choi K, Park DI (2017) Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. Intestinal Res 15(3):411. https://doi.org/10.5217/ir.2017.15.3.411
13. Kim J, Hong J, Park H (2018) Prospects of deep learning for medical imaging. Precision Future Med 2(2): 37–52
14. Levin B, Lieberman DA, McFarland B, Andrews KS, Brooks D, Bond J, Dash C, Giardiello FM, Glick S, Johnson D, Johnson CD (2008) Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US multi-society task force on colorectal Cancer, and the American College of Radiology. Gastroenterology 134(5):15701595
15. Lyon, F (2018) International Agency for Research on Cancer. Colorectal cancer factsheet, Int Agency Res Cancer
16. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochemia Med 22(3):276–282
17. Min JK, Kwak MS, Cha JM (2019) Overview of deep learning in gastrointestinal endoscopy. Gut Liver 13(4):388
18. Nogueira-Rodríguez A, Domínguez-Carbajales R, López-Fernández H, Iglesias A, Cubiella J, Fdez-Riverola F, Reboiro-Jato M, Glez-Peña D (2021) Deep neural networks approaches for detecting and classifying colorectal polyps. Neurocomputing 423:721–734. https://doi.org/10.1016/j.neucom.2020.02.123
19. Ozawa, T, Ishihara, S, Fujishiro, M, Kumagai, Y, Shichijo, S, Tada, T, (2020) Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks Ther Adv Gastroenterol, 13
20. Poudel S, Kim YJ, Vo DM, Lee SW (2020) Colorectal disease classification using efficiently scaled dilation in convolutional neural network. IEEE Access 8:99227–99238. https://doi.org/10.1109/ACCESS.2020.2996770
21. Rahman, MM, Wadud, MAH, Hasan, MM (2021) Computerized classification of gastrointestinal polyps using stacking ensemble of convolutional neural network. Inf Med Unlocked, p.100603. https://doi.org/10.1016/j.imu.2021.100603.
22. Sànchez-Peralta LF, Pagador JB, Picòn A, Calderòn AJ, Polo F, Andraka N, Bilbao R, Glover B, Saratxaga CL, Sànchez-Margallo FM (2020) PICCOLO white-light and narrowband imaging Colonoscopic dataset: A performance comparative of models and datasets. Appl Sci 10(23):8501. https://doi.org/10.3390/app10238501
23. Sebastian P, Daniel S, Begonya G, Cristian C, Adel E (2020) Kudo’s classification for colon polyps assessment using a deep learning approach. Appl Sci 10(2):501
24. Shen D, Wu G, Suk HI (2017) Deep learning in medical image analysis. Annu Rev Biomed Eng 19:221–248
25. Siegel RL, Miller KD, Fuchs HE, Jemal A (2021) Cancer statistics, 2021. CA Cancer J Clin 71(1):7–33
26. Sohail, A, Khan, A, Nisar, H, Tabassum, S, Zameer, A, (2021) Mitotic nuclei analysis in breast Cancer histopathology images using deep ensemble classifier. Med Image Anal, p.102121. https://doi.org/10.1016/j.media.2021.102121.
27. Suzuki K (2012) A review of computer-aided diagnosis in thoracic and colonic imaging. Quant Imaging Med Surg 2(3):163. https://doi.org/10.3978/j.issn.2223-4292.2012.09.02
28. Yoriaki K, Hisashi H, Tomohiro W, Takanobu N, Misaki K, Toshi- haru S., Ayana O., Tomohiro M., Masashi K., Tadaaki A. (2017) Computeraided diagnosis based on convolutional neural network sys- tem for colorectal polyp classification: preliminary experience. Oncology 93(Suppl. 1):30–34
29. Zachariah R, Samarasena J, Luba D, Duh E, Dao T, Requa J, Ninh A, Karnes W (2020) Prediction of polyp pathology using convolutional neural networks achieves ‘resect and discard’ thresholds. Am J Gastroenterol 115(1):138