Applied Intelligence (2023) 53:2410–2433
https://doi.org/10.1007/s10489-022-03689-9

A deep ensemble learning method for colorectal polyp classification with optimized network parameters

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

Accepted: 28 April 2022 / Published online: 8 May 2022

Abstract

Colorectal Cancer (CRC), a leading cause of cancer-related deaths, can be abated by timely polypectomy. Computer-aided classification of polyps helps endoscopists to resect timely without submitting the sample for histology. Deep learning-based algorithms are promoted for computer-aided colorectal polyp classification. However, the existing methods do not accommodate any information on hyperparametric settings essential for model optimisation. Furthermore, unlike the polyp types, i.e., hyperplastic and adenomatous, the third type, serrated adenoma, is difficult to classify due to its hybrid nature. Moreover, automated assessment of polyps is a challenging task due to the similarities in their patterns; therefore, the strength of individual weak learners is combined to form a weighted ensemble model for an accurate classification model by establishing the optimised hyperparameters. In contrast to existing studies on binary classification, multiclass classification require evaluation through advanced measures. This study compared six existing Convolutional Neural Networks in addition to transfer learning and opted for optimum performing architecture only for ensemble models. The performance evaluation on UCI and PICCOLO dataset of the proposed method in terms of accuracy (96.3%, 81.2%), precision (95.5%, 82.4%), recall (97.2%, 81.1%), F1-score (96.3%, 81.3%) and model reliability using Cohen's Kappa Coefficient (0.94, 0.62) shows the superiority over existing models. The outcomes of experiments by other studies on the same dataset yielded 82.5% accuracy with 72.7% recall by SVM and 85.9% accuracy with 87.6% recall by other deep learning methods. The proposed method demonstrates that a weighted ensemble of optimised networks along with data augmentation significantly boosts the performance of deep learning-based CAD.

Keywords Colorectal Cancer · Deep learning · Ensemble learning · Prediction · Transfer learning · Virtual biopsy

1 Introduction

Cancer has become one of the vital reasons for the growing mortality rate across the world over the past few decades. Colorectal Cancer (CRC), in par- ticular, is a serious form of cancer with high occurrence and mortality rates documented in developed countries [1]. It is ranked second in terms of cancer-related mortality and third in terms of CRC occurrences [2]. In order to prevent colorectal cancer-related mortalities, accurate detection and classification of polyps at a treatable stage are critical for mitigating the risk of cancers. Consid- ering the detection of colonic polyps, colonoscopy is regarded as the standard updated in a number of reports [3–7]. In the initial stage, most of the polyps have not undergone a malignant transformation, i.e., they are not cancerous and, upon their removal, the risk of cancer is reduced. However, these precan- cerous polyps have the tendency to remain unidentified during colonoscopy and may possibly become malignant, i.e., cancerous, becoming a major causality of mortality [5]. Identification of a type of polyps that have malignant transfor- mation is very crucial. Therefore, in addition to the detection, accurate polyp classification is essential to diminish mortality due to colorectal cancer as well. Machine learning along with medical image processing has been employed for cancer detection and classification [8]. Advanced algorithms have been probed to carry out

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] Department of Computer Science and Software Engineering, Auckland University of Technology, 55 Wellesley Street East, Auckland CBD, Auckland 1010, New Zealand

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

the Computer-Aided Diagnosis (CAD) for the accurate and effective diagnosis in the medical domain [1]. In recent times, Artificial Intelligence (AI) and Deep Learning (DL) have made a major contribution in medical image analysis [9–11], and the adenoma detection rate is enhanced significantly through artificially intelligent systems. This interpretation of medical images through CAD has helped physicians to become secondary readers in cancer diagnosis. Therefore, these artificially intelligent models can be used to detect colorectal polyps by interpreting endoscopy images [12].

Previous CAD algorithms relied heavily on feature extraction, which hindered the advancement of visual object detection because of dependency on visual features [13]. In order to overcome the limitations and uplift the efficiency of the CAD algorithms, deep learning superseded feature extraction and trans- formation in traditional machine learning by using convolutional operations in hidden layers [14]. Moreover, deep learning outperformed traditional machine learning and human visual ability. For instance, in healthcare, deep learning has been utilised for automatic disease detection and prediction for early diag- nosis [15].

Convolutional Neural Network (CNN) is a very powerful technique in DL for medical image diagnosis. In contrast to traditionally handcrafted feature extraction, CNN can extract abstract and higher-level features effectively. In Endoscopic Vision Challenge 2015, CNN feature extraction outperformed man- ually extracted features. Therefore, CNN can learn rich features from diverse images automatically and perform classification tasks effectively [12]. Deep CNN is mainly considered as the most suitable option for medical image classification. It has shown a lot of growth in cancer diagnosis using histopathological images [16].

More recently, Deep Learning-based Computer-Aided Diagnosis (DL-CAD) has been popularised as a comprehensive method for cancer diagnosis [17]. However, practical usages for endoscopy detection are still uncertain because these are not reliable systems in practice [18]. Deep learning-assisted colonoscopy is an attractive option to standardise endoscopy practice by elim- inating the missteps of medics and assisting domain experts or specialists in enhancing the accuracy of diagnosis. Surprisingly, the focus of previous work was on polyp detection rather than the precise classification of polyp types [19]. Accurate classification of polyp types, a challenging yet important field, has not shown much growth over the past few years [1].

The successful classification of polyps is resourceful for clinicians in terms of time and effort. Automated classification aims to differentiate gastrointesti- nal polyps which require a biopsy from the ones that need to be resected directly. DL-CAD serves the same purpose visually and virtually by classifying the polyps into its classes. A virtual biopsy is a substitute for taking the samples and submitting them for histopathology, where the polyp type is identified through chromoendoscopy. A deep learning-based virtual biopsy method is beneficial in terms of selecting the polyps which need to be directly removed from the colon, thus avoiding the time-consuming histopathological procedure. Unnecessary biopsies and complicated endoscopic procedures are prevented if polyps are accurately classified by a reliable method. Moreover, a virtual biopsy is also of great value in an actual clinical environment to decide the severity of a patient's colorectal lesions where the patient is suffering from multiple lesions.

Besides the noteworthy benefits of deep learning methods, one of the major requirements of deep learning is the availability of large medical datasets for automated model training. Expensive data acquisition and annotation make the creation of a large and well-annotated training dataset a cumbersome task [2, 20]. In these scenarios, transfer learning along with data augmentation is employed for utilising the power of pre-trained models [1]. Transfer learning is considered appropriate for training the model which does not have enough training data. In transfer learning, deep neural networks have been trained based on a large number of samples; the weights are inherited for new tasks.

Therefore, in order to fulfil this requirement, instead of designing a customised architecture of deep neural networks, pre-trained architectures are utilised in our project. However, there is a good number of pre-trained architectures avail- able for transfer learning; still, there is a need to compare and evaluate their performance to identify the one which consistently provides better results. In this paper, we compared and evaluated the existing transfer learning architec- tures for identifying the best one for colorectal classification.

However, the identification of accurate architecture for deep learning is not a simple task as it requires extensive experimentation by tuning the hyper- parameters of each net. These hyperparameters include optimizers, learning rate, the number of iterations, epochs, batches, and many more aspects. The related work presented by [15, 20] classifies colorectal polyps by using deep learning algorithms, which does not yet provide information on the optimum hyperparameter settings to be chosen so as to produce highly accurate results. Regarding deep learning architecture, it is important to find out optimum set- tings of hyperparameters, so that accurate results can be achieved.

Additionally, we cannot afford any medical risks or mistakes in medication due to automated diagnosis. In the case of colorectal cancer, if polyps are mis- classified, the polypectomy procedure will be delayed, or the physician might decide not to carry on the resection at all, which might be fatal for the patient. Therefore, it is necessary that in addition to accuracy measures, advanced eval- uation metrics such as precision, recall and F1 score are considered to make sure that the misclassification rate is as minimum as possible. It is crucial to have high sensitivity and recall of the models. In other words, tackling the type II classification error is

necessary. A high sensitivity test has a zero false negative, which means that all the negatives will be true negatives. Hence, the high sensitivity of the test is effectively applied to rule out the disease and accord- ingly act as a screening test for a disease with low prevalence. Colonoscopy is a highly sensitive test and therefore has been furnished as a screening test for colon cancer. Thus, we focused on recall measures for the performance eval- uation. If the malignant polyps are misclassified, then false negative is high, which is risky.

Our focus is thus on the computer-aided (CA) method for colorectal polyp classification in discriminating the polyps that should go under histopathol- ogy from those which should be removed directly. The goal of this paper is to classify colon polyps under narrowband imaging endoscopy into three classes: Hyperplasic, adenoma, and serrated adenoma. The first one is considered benign with little or zero ability to transform it into colorectal cancer; the lat- ter two are thought to have malignant transformation potential, where serrated adenomas lack the classic adenoma villous structure and have a mixed nature; therefore, it is difficult to be identified [17, 21, 22]. This work proposes a deep CNN based heterogeneous weighted ensemble classification technique for the analysis of endoscopy images of the colon. The class imbalance problem is handled by data augmentation technique, including rotation, scaling, brightness and flipping of images which are further classified into adenomatous, hyper- plastic and serrated 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. This novel method is based on transfer learning to resolve the classification problems of these three classes and achieve higher diagnostic accuracy with the best hyperparameter setting, which is very beneficial from the clinician's viewpoint to identify the polyps that require polypectomy. The main contributions of this paper are:

- *A novel framework for transfer learning-based virtual biopsy* that classifies colorectal polyps captured under NBI lightning. The framework tackles the problem of insufficient images in the dataset through transfer learning and image augmentation.
- The performance of six architectures in deep learning, namely GoogLeNet,
- ResNet50, Inception-v3, Xception, DenseNet-201, SqueezeNet, was evalu- ated and compared to identify the most suitable deep neural network for colorectal polyp classification.
- *Establishing optimum hyperparameter settings* for optimisation of deep
- neural networks and making the results reproducible and explainable.
- *Classifying polyps into three classes: serrated polyps in addition to hyper- plastic and adenomatous polyps* which lead to CRC through an alternate serrated pathway which are difficult to be identified.
- *Weighted average ensemble model* to deal with complex nature of polyps by improving the generalisation of the classification system.

The structure of this paper is as follows. In Section 2, we introduce the related work which is essential to understand our research background and motivations. In Section 3, we expound the dataset for our experiments. In Section 4, we explicate our proposed framework to classify colorectal polyps into three classes, the architectures are associated with the training process. Experimental results are demonstrated in Section 5. Finally, in Section 6 we present the comparative analysis, followed by the conclusion and future work in Section 7.

2 Related work

There are existing models proposed for automated classification of colon polyps. Komeda et al. [17] have suggested a model that determines polyps in two types, i.e., adenomatous and non-adenomatous. The dataset includes 1,200 adenomatous and 600 non-adenomatous images which were taken out from a digital video of actual medical examinations. The data was collected from the cases of colonoscopy, which was completed between January 2010 and December 2016. Computer vision is a way that classifies objects; hence, it is used to determine colon polyps as well. The work has combined the benefits of both computer vision and convolutional neural networks (CNNs) together for accurate classification. The proposed model, which is CNN based on CAD, generates the results based on real-time endoscopy images, nonetheless, the accuracy is 0.751 which is based on 10-hold cross-validation test. The work does not classify the classes of polyps: Serrated adenomas, adenomatous polyps, and hyperplastic. Moreover, the proposed model might have performed better with a big dataset. Hence, by using transfer learning and image augmentation, the problem of insufficient data is catered and resolved. Lastly, hyperparameter tuning also contributes towards better outcomes with high accuracy. Even though the accuracy is not satisfactory, the CNN-CAD method is still a better choice as it simplifies the operations and classifications.

Another method classifies colon polyps as malignant and non-malignant. Patino-Barrientos et al. put forward a deep learning model based on Kudo's classification schema [23]. The dataset consisting of 600 images was collected from 142 patients, which was further augmented to increase the number of sam- ples. The problem was approached iteratively by firstly implementing a deep neural network, then compared the results by applying VGG-16 net. Further- more, fine-tuning was offered, and the results were comparable. The validation parameters for evaluations are accuracy, precision, recall, and

F1 score. After fine-tuning, the accuracy, precision, recall and F1 score of the model was 83%, 81%, 86% and 83%, respectively. Later, the model was compared with other classifiers such as SVM and KNN. The outcomes of SVM and KNN with 15 neighbours illustrated the same results, nevertheless, the proposed model is proved to be a better way for the classification. However, the amount of data is insufficient, as deep learning model is proposed which requires large dataset for better performance. If the data is augmented, it is able to boost the model and give much satisfactory results.

Furthermore, an AI-based detection and classification of colorectal polyps are developed [24] which takes advantage of deep neural networks. The algorithm is entitled as Single Shot Multibox Detector (SSD), which classifies its classes such as adenoma, hyperplastic polyp, sessile serrated adenoma/polyp, cancer- ous and other polyps. In the work, the data for model training was taken from 12,895 patients who underwent colonoscopies. Moreover, 16,418 images were applied to train the CNN algorithm, among which 3,021 images were of polyp and 4,013 images of normal colorectal. The processing time of CNN was 20ms per frame. The trained CNN model detected 1,246 CP with sensitivity 92% and a positive predictive value (PPV) 86%. The sensitivity and PPV were 90% and 83%, respectively, for the white light images, 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. However, the optimized hyperparameters are not employed for giving better results. Lastly, the results unfold that the accuracy of detection and classification is commendable and has great potential.

The model proposed by [25] is based on deep neural networks to classify colorectal polyps. The neural network architectures in this paper are recom- mended as RestNet which comprise of 5 family members with 18, 34, 50, 101, and 152 layers, respectively. The suggested models classify four major colorectal polyp types: Tubular adenoma, tubulovillous or villous adenoma, hyperplastic polyp, and sessile serrated adenoma. The dataset was split into 3 subsets: 326 slides for training, 157 slides for testing, and 25 for validation. Furthermore, 238 slides were collected from 24 different institutes. The deep learning algorithms were designed and trained for the classification. The slides were segmented into patches by using sliding windows, which were further clas- sified. In addition, the thresholds were classified for each class by using a grid search. The primary purpose was to evaluate the performance of the model in comparison with the results annotated by pathologists. In order to evalu- ate the performance, the metrics include accuracy, sensitivity, and specificity. For the internal dataset, the mean accuracy of the model was 93.5% and that of pathologists was 91.4%. Furthermore, for the external dataset, the model achieved an accuracy 87.0% as compared to the pathologists' accuracy which was 86.6%. A major limitation is the small dataset, which hinders the perfor- mance of the model in practice. Since it is difficult to collect medical data, transfer learning is one way of overcoming the limitations, data augmentation can also be offered. In summary, the difference between the outcomes of the proposed model and that of local pathologists was minor, hence, the model is used to assist doctors so as to improve the diagnosis of colorectal polyps.

The method of [26] is based on deep learning, hence, an automated method for image analysis that classifies colorectal polyps. The model determines five types of colorectal polyps such as hyperplastic, sessile serrated, traditional serrated, tubular, and tubulovillous/villous. The dataset was collected from the patients who were examined for colorectal cancer. In this work, 458 whole- slide images were taken use for training and 239 for testing purposes. There are 2,074 cropped images in total. In order to characterize the polyps, vari- ous deep neural networks were implemented and compared to find the best approach for the problem. The standard architectures such as AlexNet, VGG, GoogleNet, and ResNet were taken, however, ResNet was observed with var- ious numbers of hidden layers. Furthermore, ResNet-A ResNet-B, ResNet-C, and ResNet-D were composed of 50, 101, 152, and 152 layers, respectively. Although ResNet-C and ResNet-D have similar layers, they vary in the mapping such as identity mapping and projection mapping. Among the network architectures, ResNet performed with the highest accuracy. The parameters for validation include accuracy, precision, recall, and F1 score which yielded the results 93.0%, 89.7%, 88.3%, and 88.8%, respectively. Moreover, with hyper- parameter tuning of the proposed model would have more weight.

Mesejo et al. [27] have developed a method that saves clinician's time by performing a virtual biopsy of gastrointestinal lesions. The proposed system combines the algorithms in machine learning and computer vision, classifies lesions, hyperplastic lesions, serrated adenomas, and adenomas. Firstly, the digital images are taken as the input. Next, the color and texture image fea- tures are extracted, respectively. Then, the motion is used to reconstruct 3D lesion, the 3D shape features are extracted. The image is then imported into a classifier for prediction. In this paper, two classifiers were incorporated: Random Forest and Random subspaces. Furthermore, SVM was applied to comparisons with the ensemble leaners. The dataset containing 76 colonoscopy videos was built by the researchers for training. The results were compared with the expert and beginner practitioners. The average accuracy of random forest, random subspaces, and SVM was 0.78, 0.49, and 0.29, respectively. In addition, another computer-aided method [28] detects and classifies hyper- plastic and adenomatous colorectal polyps. In this work, CNNs are employed for the detection and classification. Firstly, a convolutional neural network is employed to detect the polyp. However, the approach to solving the classifi- cation problem differs

from the aforementioned paper [17]. Secondly, another CNN is applied to classify the polyp. The CNN features are learned from two publicly available datasets; ILSVRC and Place205, which contain 1.2 million images and 2.5 million images, respectively. The proposed method has attained an accuracy of 85.9% which was higher in comparison with the result of practi- tioners which was 74.3%. However, the optimized hyperparameters would have contributed towards yielding more substantial results. Lastly, the computer- aided methods assist doctors to make better decisions in the diagnosis of polyps at an early stage. The proposed method is able to diagnose colorectal with the minimum preprocessing procedures compared to other methods.

Dataset used by following two studies [27, 28] is a publicly available dataset which our study has also benefited from. In addition to this data repository another dataset i.e., PICCOLO dataset used in this study is also employed by few recent studies for accomplishing colorectal polyp detection tasks. Work of Pacal et al. [29] has optimized real-time detection architectures YOLOv3 and YOLOv4 architectures for polyp detection. CSPNet network was applied to head and neck structure of YOLOv3 whereas for YOLOv4 it was applied on complete structure. Moreover, to improve the performance SiLU activation function was used that outperformed other activation functions. Results showed the success of proposed method with increasing number of training images. Here, the model trained on combination of SUN, CVC-ClinicDB and PICCOLO dataset gave the best results, and it had the largest number of train- ing images. Authors did not make any modifications in PICCOLO dataset as it is already divided into train, test and validation sets. However, our work has applied augmentation techniques on this dataset to handle data imbalance. Data insufficiency is a major problem in medical domain. Therefore, availabil- ity of publicly accessible dataset is essential for development of detection and classification system and to facilitate fair comparison of the developed systems. Consequently, through the biobank of the Instituto de Investigacio'n Sani- taria Galicia Sur (IISGS) (https://www.iisgaliciasur.es/home/biobank-iisgs) these datasets are currently under the necessary procedure for public access. The publication of dataset by Nogueira-Rodrı'guez et al. [30] will enhance the availability of public datasets which has also been expanded recently by with the addition of the PICCOLO Dataset. Main aim of this study was to develop a deep learning model for real-time polyp detection and the developed model could be integrated into a CAD system in future. Due to a balance between prediction time and performance of YOLOv3, it was employed by this study as the base architecture.

A deep learning model for the classification of polyps, adenomatous polyps, and serrated polyps was put forth by Zachariah et al. [31]. 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 of 5,278 high-quality images was used for training and testing the proposed model. The proposed CNN model consists of two modules, namely the base module and head module. The base module uses the Inception-ResNetv2 algorithm for automated feature extraction. Alternatively, the head module of the algorithm is engaged in transforming the extracted features to a graded scale which can further be used for classification. The colorectal polyps in this project are clas- sified into 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 mod- els based on white light and narrow banded imaging. The negative predictive value for the fresh data was 97%, and overall surveillance concordance was 94%.

In another attempt to classify the polyps [32], five classes were organised: Ade- nocarcinoma, adenoma, Crohn's disease, ulcerative colitis, and normal images. The dataset comprising 3515 images was collected from Gill Hospital. Further- more, the KVASIR dataset consisting of 4000 images was also employed for validation of the proposed model. In the model, the deep layers have their spa- tial information preserved by using diluted convolution for better classification of polyps. Additionally, the architecture ResNet-50 was taken into account so as to avoid overfitting, whereby Drop Block helps in the regularisation of the model. The performance metrics include accuracy, recall, precision, and F1- score for evaluations. The F1 score of 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 in network-based transfer learning model was proposed for the improved classification of polyps. The dataset consists of 1000 instances that 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-in-Network is the stacking of a 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 $\pm$ 0.029, and the AUC was approximately 0.930 $\pm$ 0.020. The performance measures depict the proposed model to be useful to assist doctors in classifying normal and abnormal polyps more accurately. However, other architectures such as ResNet, DenseNet and many such forms 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 by Rahman et al. [33]. 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 is also used to reduce the processing time of the model. Lastly, a stacked ensemble learn- ing was applied. The proposed method consists of three convolutional neural network architectures: Xception, ResNet-101, and VGG-19. The models are fine-tuned and then a softmax classifier was used 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 optimizer in the hidden layers. The performance 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. Table 1 summarises the recent studies done in the area of colorectal cancer diagnosis.

3 Dataset

3.1 UCI dataset

Availability of high-quality large polyp dataset is crucial for developing an efficient deep learning architecture to successfully classify the colonoscopy images through an automated solution. Generally, in order to develop a decent and high performing deep learning model, large datasets such as ImageNet, Microsoft COCO, including millions of hand annotated images with object classification: highlight and labelling are extensively used. However, creating such high-quality large dataset in biomedical domain is a challeng ing and expensive task with regard to finance and expertise needed [20]. In past few years, public and private datasets for colonoscopic polyp detec tion and classification are released but the size of data available is not as large, therefore, several studies have collected their own private dataset for the purpose. Dataset used for polyp classification in this project is Gas- trointestinal Lesions in Regular Colonoscopy Data Set publically avaiable at available at http://www.depeca.uah.es/colonoscopy dataset/ , the dataset has also been used by other CAD researches including Mesejo et al. [27]. The dataset includes 76 images, consisting of 40 adenomatous polyps, 21 hyper- plastic lesions, and 15 serrated adenomas. The dataset was built by 76 short colonoscopy videos recorded by clinicians and varying lightning conditions. Wight Light (WL) and Narrow Band Imaging (NBI) both are included in the data. However, all the experiments were performed based on digital images extracted from colonoscopy videos captured under NBI lightning conditions as it is the advanced optical technique to differentiate lesion types by providing extended details of vascular patterns of WL colonoscopy [28]. The three input images from each class are shown in Fig. 1

3.2 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 and is available at https://www.biobancovasco.org/en/Sample-and-data-catalog/Databases/PD178-PICCOLO-EN.html. The frames during colonoscopy were captured through varying lightning technologies: white light (WL) and narrow band imaging (NBI). Metadata information of acquired data and annotation procedure is described in the subsections below.

- 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 includes 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 performed based on the captured frames to identify the type of lightning condition used to classify them as polyp or non-polyp images. One frame per second was manually annotated (i.e., one out of 25 frames). Frames were collected and revised by a researcher to ensure the completeness of dataset. Colonoscopic video frames were recorded at Hospital Universitario Basurto, Spain between October 2017 and December 2019 using Olympus endoscopes (CF-H190L and CF-HQ190L) [19]. The dataset contains 3,433 WL and narrow band imaging NBI images from clinical colonoscopy procedure videos in human patients. Total 46 patients were examined, and 76 different lesions were included in the dataset. Data was distributed into three sets having 2,203 images in training set, 897 in validation set and 333 in test set. Details of frames in each set is given in Table 2. The dataset contains three types of polyps: Adenoma, Hyperplasia, and Adenocarcinoma. Figure 2 shows multiple samples of each polyp class in dataset.

4 Computer-aided colorectal polyp classification

The proposed framework for colorectal polyp classification is shown in Fig. 3, and the description of each part is presented underneath.

Table 1 Colorectal polyp diagnosis techniques in literature

Author(s)	Year	Focused Area	No of images	Type of Polyps	Data Source	Implemented Network	Evaluation results
Komeda et al. [17]	2017	Classification	1800	• Adenomatous • Non-adenomatous	Kindai University Hospital	CNN	• Accuracy: 0.75
Patino-Barrienos et al. [23]	2020	Classification	600	• Malignant • Non-malignant	Biodonostia Health Research Institute	VGG-16	• Accuracy: 0.83 • Precision: 0.81 • Recall: 0.86 •F1 score: 0.83
Ozawa et al. [24]	2020	Detection and classification	16,418	• Adenoma • Hyperplastic • Sessile serrated	Tada Tomohiro Institute Gastroenterology and Proctology, Japan	CNN	• Sensitivity: 90 • PPV: 83
Wei et al. [25]	2020	Classification	508	• Tubular • Adenoma • Tubulovillous • Hyperplastic • Sessile serrated Adenoma	Dartmouth-Hitchcock Medical Centre, Lebanon, New Hampshire.	Ensemble of ResNets 18,34,50,101,152 layers	• Accuracy: 0.87
Korbar et al. [26]	2017	Classification	2074	• Hyperplastic • Sessile serrated • Traditional serrated • Tubular Tubulovillous	N/A	ResNet50	• Accuracy: 0.93 • Precision: 0.89 • Recall: 0.88 • F1-score: 0.88
Mesejo et al. [27]	2016	Classification	76	• Hyperplastic • Sessile Serrated • Adenoma	UCI Repository	• SVM • Random Forest • Random subspace	Accuracy: • 0.78 • 0.49 • 0.29
Zhang et al. [28]	2016	Detection and Classification	• 1.2 million • 2.5 million	• Hyperplastic • Adenomatous	• ILSVRC • Places205	CNN	• Accuracy: 0.85
Zachariah et al. [31]	2020	Classification	5278	• Adenomatous • Serrated	N/A	Inception-ResNetv2	• Negative predicted value: 0.97
Poudel et al. [32]	2020	Classification	35,154,000	• Adenocarcinoma • Adenoma • Crohn;s Disease • Ulcerative • Colitis Colitis	Gill Hospital KVASIR dataset	ResNet50	F1-score: • 0.93 • 0.88
Rahman et al. [33]	2021	Classification	26,512	• Hyperplastic • Serrated Adenoma • Non-Polyp	University of Alcala	• Xception Ensemble • ResNet-101 • VGG-19	• Accuracy: 0.98 • Recall: 0.96 • Precision: 0.92 • Specificity: 0.98

Fig. 1 The polyps' samples from different classes of UCI dataset

4.1 Data oversampling and augmentation

The input data to the system is the collection of 76 colonoscopy images from UCI Repository and 3433 images from PICCOLO dataset, thus, they are insuf- ficient for model training in deep learning as a large dataset is important for classification; therefore, oversampling is done on the dataset. The presence of imbalance in data was confirmed and label information was extracted from the training dataset. The data was split into 80% training set and 20% testing set. Group indexes associated with the classes were obtained and variable labels at each class were extracted from the training set. The minority classes were oversampled compared to the number of images in the majority class. Because the dataset for this project is small and is not similar to the data of pre-trained model, developing an effective solution is challenging. If we go very deep in the layers, the model easily overfits, the model might not be trained effectively. In order to deal with this problem, data augmentation was conducted for a successful transfer learning.

Table 2 Frames in each of the sets according to clinical metadata

Category	Items	Train Set	Validation Set	Test Set
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	–	–

4.2 Transfer learning and training process

CNNs are usually employed for the development of classification or localisation deep learning models. The classification of objects for digital images is achieved in two different ways, either by implementing an off-the-shelf CNN architecture or by designing a custom architecture where the former approach is the basis of the new architecture. Deep learning architectures are mostly suitable for classification tasks; however, benefiting from off-the-shelf models can significantly simplify the model development because they are able to be modified and adapted according to the new task [2]. Furthermore, this domain benefits from the commonly practised deep learning technique, transfer learn- ing, where a model built for a particular task is used for another custom task. In transfer learning, the model is trained based on public datasets, and the initial weights of this model are used for the task instead of assigning random weights as done in a network designed from scratch. In the next step, the last fully con- nected layer is usually responsible for final classification, i.e., presenting the images of new classification to the network where weights of specific layers are adjusted in the regular training process. In this paper, we consider six CNN architectures: GoogLeNet, ResNet-50, Inception-v3, Xception, DenseNet-201 and SqueezeNet. The information about these architectures is shown in Table 3. Each model has been independently trained with the training data of the three classes.

4.3 Model tuning

Based on the classification task pre-trained model is fine-tuned, fully connected and modified. Each model is

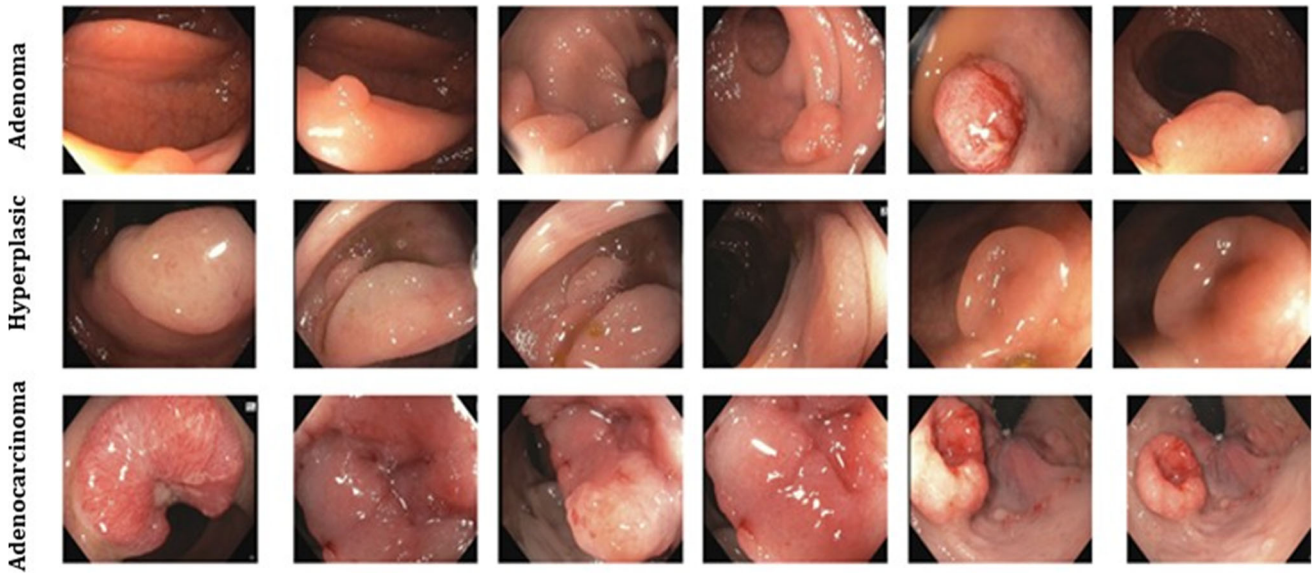

Fig. 2 Sample images of polyps from each class of PICCOLO Dataset

trained separately on oversampled and augmentedimages according to the specified training options to classify the images into the respective categories. Algorithm is given below.

Algorithm of proposed framework

Inputs:

1. Training data $Q = \{q_1, q_2, ..., q_n\}$.
2. Hyperparameter search space $P = \{p_1, p_2, ..., p_n\}$.
3. Number of trials *(X)*
4. Number of classes *(C)*
5. Number of models to be used in the ensemble *(N)*
6. Testing data $Q' = \{q_1', q_2', ..., q_n'\}$.

Training Process:

7. Generate the set of N random hyperparameter combinations from *P*.
8. Create N different networks pertaining to *X* combinations found in step 7.
9. Train each pre-trained network PT_{net} on *X* from step 8.
10. Choose the most efficient PT_{nets} with specified accuracy threshold from step 9.
11. Perform grid search and assign suitable weights $w_{\mathrm{ts}} = argmax\sum_{i=1}^{n} w_1, ..., w_{\mathrm{n}}$ *to PTnet.*
12. Perform the final training for the best-M networks selected from step 11 using entire training dataset

Testing Process:

13. Input testing image *q'* from the *Q'*
14. Generate output *h(q')* predictions from each of the chosen *PTnet* from step 12.
15. Perform the final classification by doing an ensemble of predictions from step 14.

Output:

16. Final classification label *h(q')* for a testing image from *Q'*

4.4 Evaluation metrics

Recall rate is considered as the evaluation metric for the classification model. Colorectal polyp classification is a class imbalance problem. Hence the perfor- mance 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 metric for imbalance dataset. Dataset has a variety of imbalance in data. The evaluation of imbal- anced data results requires advanced metrics. Furthermore, this project aims at three classes classification. Therefore, in addition to accuracy,

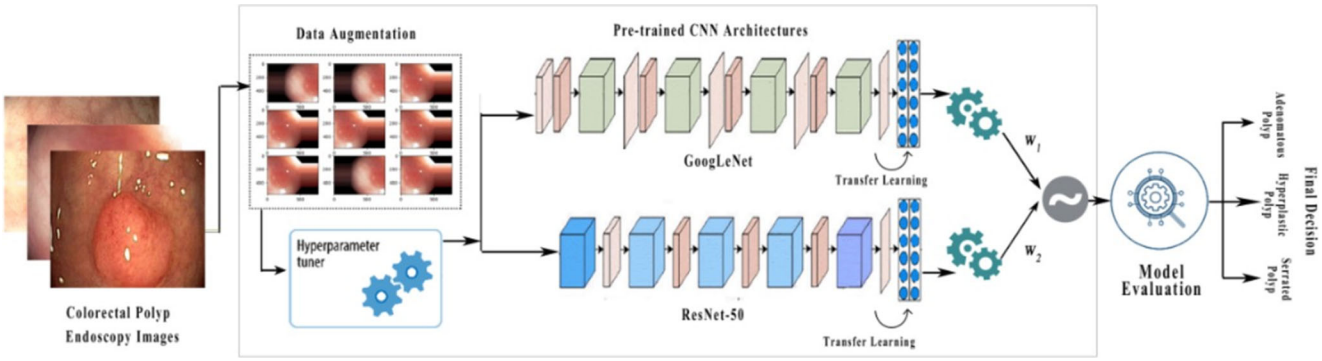

Fig. 3 An overview of the proposed weighted-average ensemble classifier

recall, pre- cision, 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 interpretation in multi-class settings. Confusion matrix of every class $\mathbf{g_i} \in \mathbf{G} = \{\mathbf{1}, ..., \mathbf{K}\}$ such that the i-th matrix takes $\mathbf{g_i}$ class as the positive class and rest of the classes $\mathbf{g_j}$ with $\mathbf{j} = \mathbf{i}$ being the negative classes. Micro average pools the performance over all the samples or in other words, over the smallest possible unit to compute overall performance. Micro-averaged F1-score is computed from micro-averaged recall Rmicro and micro-averaged precision Pmicro. The mathematical equation 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 rep- resents a good performance of individual classes. Mathematical formulas are given in (4), (5), and (6).

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

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

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

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

Cohen's Kappa Coefficient shows the performance evaluation and reliability analysis in imbalanced class problem. In (7), po represents the overall model accuracy and pe represents the model prediction and actual class value by chance agreement. The co-efficient results are interpreted as follows: ≤ No-agreement when values 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 [34].

Table 3 Pre-trained CNN architecture details

Model(s)	Depth	Parameters (Millions)	Image Input size
GoogLeNet	22	1.24	224×224
ResNet-50	50	25.6	224×224
Inception-v3	48	23.9	299×299
Xception	71	22.9	299×299
DenseNet-201	201	20.0	224×224
SqueezeNet	18	1.24	227×227

5 Experimental results

Colon cancer incidence rates are reduced if colorectal lesions are identified at an early stage. These polyps are detected efficiently with the help of high-quality endoscopes having high magnification and improved image capturing capabilities. Since these instruments are highly expensive and are not always available; hence, it is important to develop a computer-aided solution that can perform the classification of the colonoscopy images at a reduced cost, thus making it

affordable for the regions where these devices are neither easily available nor producing reliable results. Two sets of experiments are conducted in this project, aiming to find out which set of training hyperparameters produces the best results. All the variations of the experimental setting implemented in this paper are given in Table 4. The evaluation metrics to measure the performance of models are accuracy, precision, recall, and f1-score.

5.1 Performance on benchmark data

5.1.1 UCI dataset

This set of experiments was conducted on the benchmark data. There are six contemporary CNN architectures for accomplishing the experiments. The pur- pose of these experiments is to find the most suitable hyperparameter settings among the neural networks for the purpose of classifications. At the first step, the dataset was segmented into a training and a test set. Then the training set was passed to the pre-trained network to find out the best possible opti- mizer for the given dataset. Three optimizers selected are ADAM - Adaptive Moment Estimation, SGDM - Stochastic Gradient Descent, RMSprop - Root Mean Square Propagation.

The error rate of the deep neural network model during the training phase can be reduced by optimisation algorithms. Adam optimiser performs well with minimal tuning and has shown its competence in model performance. This method has been utilised in many applications for training neural networks. However, we have aimed to perform a comparative analysis of various popular optimisers to identify the best fit for this study in combination with other hyperparameters. As our study is performed on a balanced dataset as well, the SGDM optimiser is also considered as it performs better on a larger dataset and can outperform ADAM's performance

Table 4 Experimental settings

Networks	Optimizers	Learning Rates	Epochs
GoogLeNet	ADAM	0.001	50
ResNet-50			100
Inception-v3		0.005	50
Xception			100
DenseNet-201			
SqueezeNet	SGDM	0.001	50
			100
		0.005	50
			100
	RMSprop	0.001	50
			100
		0.005	50
			100

[35]. During the first phase of the experiment, six networks were tested with these three optimizers one by one with a learning rate of 0.001 and 50 epochs. In the next step, the networks were tested for the same learning rate but by increasing the number of epochs from 50 to 100. The results obtained are shown in Table 5 and the relevant results are shown in Figs. 4 and 5.

The first experiments were started with a learning rate of 0.001 and the performance is listed in Table 5. Later, the bending ratio was increased to 0.005, and the number of epochs was set to 50 and then 100 for the next phase of the experiment. The results obtained for all the experiments show that ADAM performs the best among the three optimizers by giving consistently higher accuracy and recall of 86.67% and 86.63%, respectively, on all six models.

However, if we examine the results further, we observe that optimizer RMSprop gives the worst results of 53.33% accuracy and 51.39% recall with most of the models and is clearly not a good choice for further experimentation. It is evident from the results that if we must choose the best optimizer among the three tested options, ADAM is the most suitable one as it has high sensitivity.

After establishing the most efficient optimizer, ADAM, further experiments were conducted to select the number of epochs that generate the best result. The numbers of epochs chosen were 50 and 100 for learning rates 0.001 and 0.005 with the ADAM optimizer. The comparison of results shows that 100 epochs produce better results with 86.67% accuracy and 86.63% recall as com- pared to 50 epochs that give 80.0% accuracy and 81.12% recall regardless of the learning rate value. Now, the last parameter to be decided is the learning rate. According to the selected optimizer and number of epochs, results produced by both learning rates were compared. The results presented in Table 6 shows that 0.001 learning rate produces better results in comparison to 0.005 learning rate

5.2 Performance on balanced data

The purpose of this experiment is to examine the performance of the proposed models. One of the major issues that influence the performance of a model is the imbalance between the classes. The datasets utilised for this work were imbalanced, which deteriorated the performance of this model. In order to accomplish this problem, the imbalance was removed from data by making the number of images in all the classes equal, as shown in Table 7. After oversampling the data, each pre-trained network was loaded and modified. In the next step, as the images available in the dataset were limited, the data augmenter was defined for rotating and scaling to perform augmentation before passing the data to the deep learning algorithm. Similar settings were used in this experiment for a fair selection of the best hyperparameters to identify the most

Table 5 Learning rate: 0.001 and 100 epochs

Model(s)	ADAM				SGDM				RMSprop			
	Acc.	Prec	Recall	F1-score	Acc.	Prec	Recall	F1-score	Acc.	Prec	Recall	F1-score
GoogLeNet	**86.7**	**87.6**	**86.6**	**87.1**	73.3	75.4	77.2	76.3	60.0	61.2	66.5	63.8
ResNet-50	73.3	75.4	77.2	76.3	53.3	50.8	51.4	51.1	53.3	50.8	51.4	51.1
Inception-v3	60.0	61.2	66.5	63.8	60.0	61.2	66.5	63.7	60.0	61.2	66.5	63.7
Xception	53.3	50.8	51.3	51.1	53.3	50.8	51.3	51.1	53.3	50.8	51.3	51.1
DenseNet-201	60.0	61.2	66.5	63.8	66.6	77.7	65.2	70.9	60.0	61.2	66.5	63.8
SqueezeNet	60.0	61.2	66.5	63.75	60.0	61.2	66.5	63.7	53.3	50.8	51.3	51.1

efficient deep learning architecture with the chosen settings. In the experiments based on oversampled data, the same experimental settings were applied to balanced dataset in previous experiments; all the possible combinations of hyperparameter settings were tested in this experiment as well. The three chosen optimisers were tested based on augmented and oversampled data to choose the best optimizer among ADAM, SGDM and RMSprop. All the optimizers were tested by changing other parameters, and the results were compared to the best optimizer. After selecting the optimizer, the performance based on various numbers of epochs is compared with the experiment based on benchmark data. In the next step, the chosen optimizer and the number of epochs is kept the same for further experimentation, wherein the best learning rate was chosen between 0.001 and 0.005. Once all the parameters were selected based upon the results, the best architecture of deep learning models is identified. The results are shown in Figs. 6 and 7.

Experimental results show that the SGDM optimizer generates the highest value of all the evaluation metrics with 93.33% accuracy and 95.83% recall on all the settings except 0.001 learning rate and 50 epochs where ADAM performs better than SGDM. The obtained results reflect that the most efficient optimizer, SGDM keeps constituency for the rest of the experimental settings. Further experiments were conducted to select the number of epochs that generate better results. The number of epochs chosen was 50 and 100 for learning rates 0.001 and 0.005 with the SGDM optimizer. The comparison of results in Table 8 indicates that 100 epochs produce better results as compared to 50 epochs regardless of the value of the learning rate.

Another important hyper-parameter to decide is the learning rate. According to the selected SGDM optimiser and number of epochs 100, the results produced by both learning rates are compared. The results are shown in Table 9, which reveal that a learning rate of 0.001 yields a better result. The results of this experiment show that class imbalance affects the performance of the deep learning models; however, if this problem is handled prior to training the model, higher accuracy is achieved. Therefore, handling the imbalance in data is

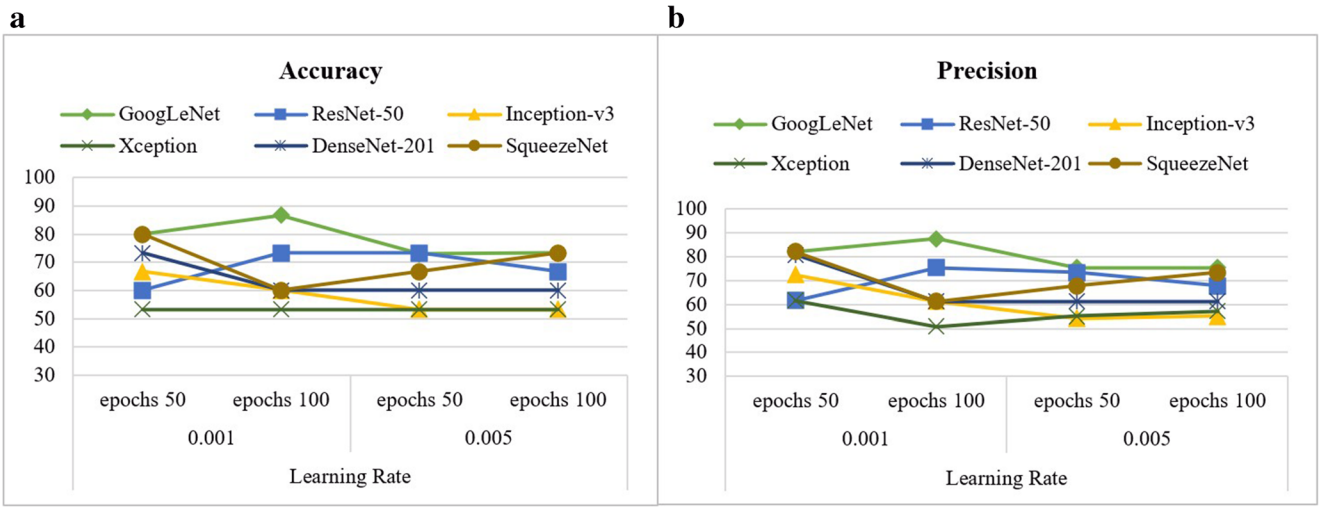

Fig. 4 **a** Accuracy and **b** Precision Results with Optimizer: ADAM, Learning Rate 0.001 And 0.005, number of epochs 50 and 100

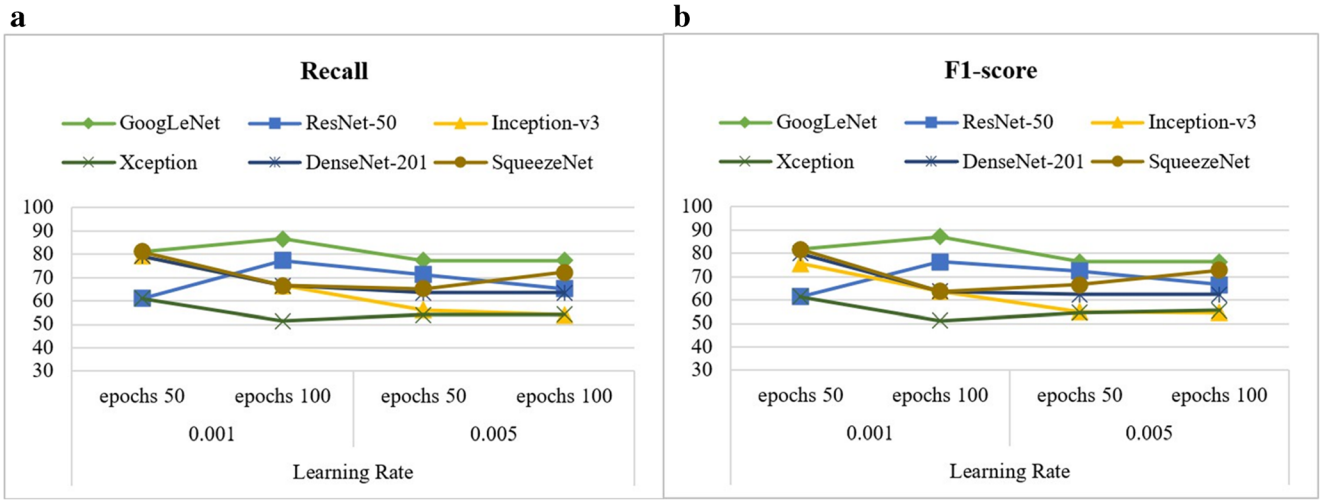

Fig. 5 **a** Recall and **b** F1-score Results with Optimizer: ADAM, Learning Rate 0.001 and 0.005, Number of Epochs 50 and 100

beneficial in obtaining more accurate predictions. The results are generated for two class and three class classification and both experiments with experimental settings as shown in Table 10. GoogLeNet performed the best on benchmark data with 87.5% accuracy, 86.63% precision, 86.63% recall, and 87.12% F1-score with ADAM optimisers, learning rate 0.001 and 100 epochs.

On balanced datasets, ResNet-50 gives the most accurate results with the same learning rate and the number of epochs as of the benchmark data experiment; however, the optimiser that performs better is SGDM yielding 93.33%, 33% precision, 95.83% recall and 94.56% F1-score.

One of the publications [27] has used the same benchmark dataset for the classification of polyps, but the study classified them into merely two classes, namely hyperplasic polyps and adenomatous polyps. For a fair comparison of results, we have also experimented with these two classes. All the experiments were executed with the optimum hyperparameter settings; however, the best results are included to make it easier.

According to the results shown in Table 10, ResNet-50 provided the most accurate results based on benchmark and balanced datasets. ADAM and SGDM optimisers produced similar and highest results with 0.001 learning rate and 100 epochs: 86.67% accuracy, 91.7% precision, 91.76% recall and 91.7% F1 score. However, the worst-performing CNN architectures for two- class classification are DenseNet201 for benchmark data with 40% accuracy, 66.7% precision, 37% recall, 47% F1 score and GoogLeNet for oversampled data yielding 67% accuracy, 72.7% precision, 80% recall and 76.2% F1 score.

5.3 Ensemble learning of optimized networks

5.3.1 UCI dataset

Neural networks have a high level of variance and low bias. In order to reduce the variance of the neural network, a better approach is to train multiple models instead of a single model to combine their predictions; this method is known as ensemble learning. Combining the predictions of multiple models adds a bias to the model, which in turn reduces the variance of a single trained model. In addition to reducing the variance of the model, this approach improves the model performance. The results are predictions that are less sensitive to the specifics of the training data and training scheme.

Ensemble learning can be done with varying training data, varying models, and varying model combinations where an average of model predictions is calculated that can be

Table 6 Results with Optimizer: ADAM, Learning Rate 0.001 and 0.005, Number of Epochs 100

Model(s)	0.001				0.005			
	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
GoogLeNet	**86.7**	**86.6**	**86.6**	**87.1**	75.4	75.4	77.2	76.3
ResNet-50	73.3	77.2	77.2	76.3	67.7	67.8	65.3	66.5
Inception-v3	60.0	66.5	66.5	63.7	55.1	55.2	54.1	54.6
Xception	53.3	51.3	51.4	51.1	57.2	57.2	54.2	55.6
DenseNet-201	60.0	66.5	66.5	63.8	61.2	61.1	63.5	62.3
SqueezeNet	60.0	66.1	66.5	63.5	73.3	73.4	72.2	72.8

Table 7 Number of polyps per category

	Adenomatous Lesions	Hyperplastic Lesions	Serrated Lesions
Benchmark Dataset	40	21	15
Balanced Dataset	40	36	33

enhanced by weighing predictions of each model. The model used in this study is the weighted average ensemble, also known as model blending. [36]. It is difficult to classify colorectal polyps due to their complex mucosal pattern. Therefore, to effectively deal with the problem, we aimed to improve the generalisation of the classification system by benefiting from ensemble learning. The top two optimised pre-trained networks, GoogLeNet and ResNet-50, are selected based on a specified accuracy thresh- old. The strength of deep learning networks with performance more than the specified threshold is combined to improve the overall performance of the classification problem. Base-classifiers are trained individually with the Ima- geNet database. As the individual learners might have a limited capability to capture data distribution, it is a good approach to combine the capabilities of individual networks into an ensemble to generate the outcome of the classifier. An averaging ensemble-based classifier was developed to further enhance the performance of the classifier by assigning carefully chosen weights to the base-classifiers. A grid search was performed to select the weight values in order to maximise the performance of the ensemble model.

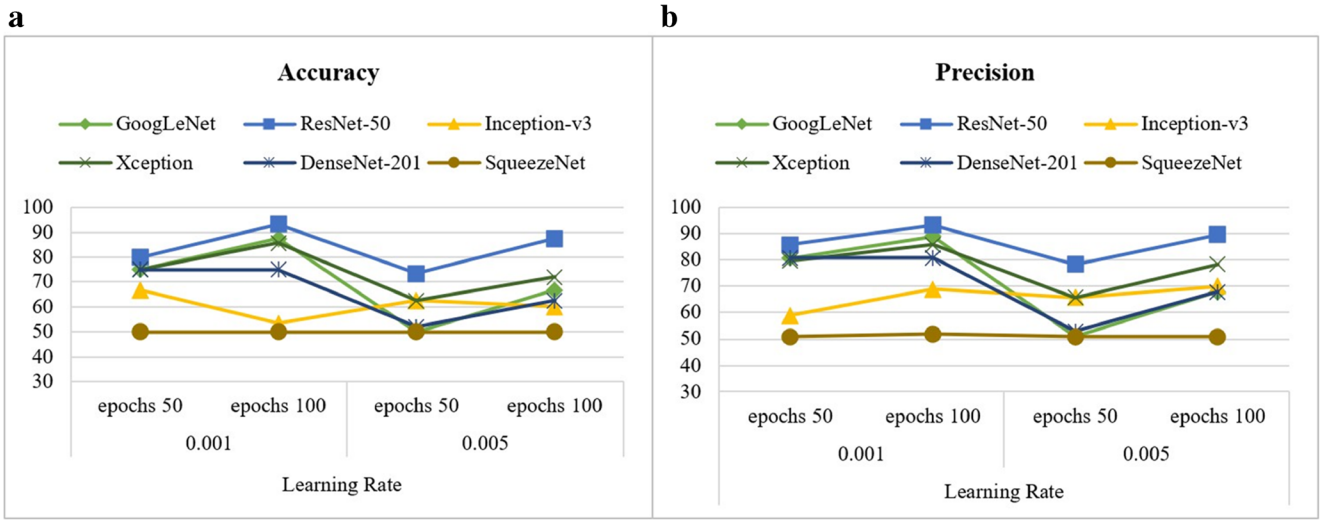

Fig. 6 **a** Accuracy and **b** Precision Results with Optimizers: SGDM, Learning Rate 0.001 And 0.005, number of epochs 50 and 100

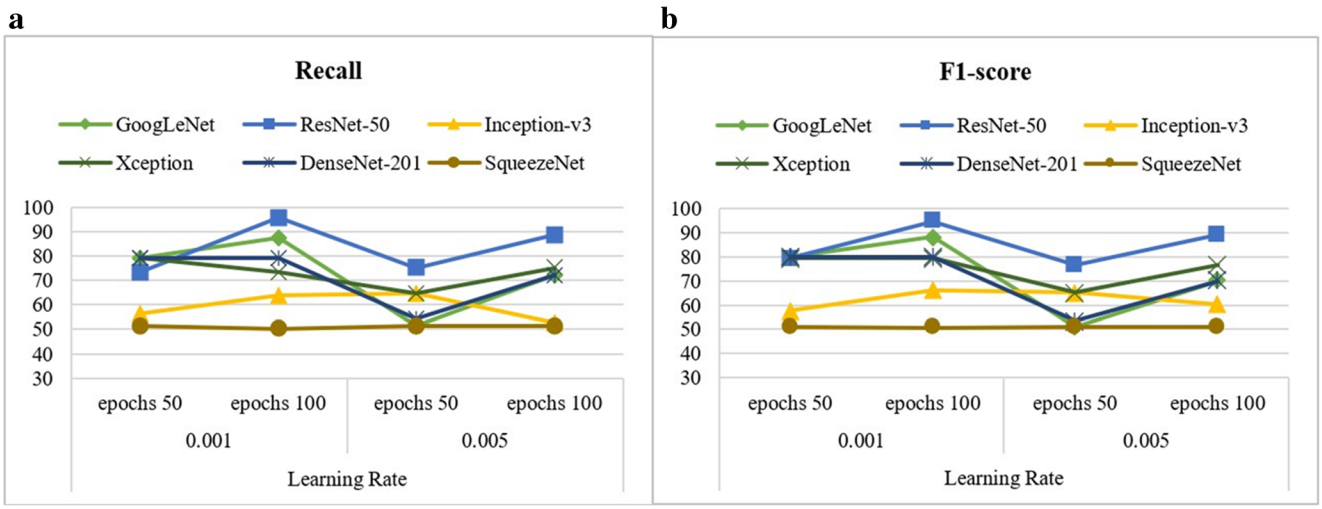

Fig. 7 **a** Recall and **b** F1-score Results with Optimizers: SGDM, Learning Rate 0.001 and 0.005, Number of Epochs 50 and 100

Table 8 Learning rate: 0.001 and 100 epochs

Model(s)	ADAM				SGDM				RMSprop			
	Acc.	Prec	Recall	F1-score	Acc.	Prec	Recall	F1 score	Acc.	Prec	Recall	F1-score
GoogLeNet	25.0	58.0	51.1	54.4	87.5	88.9	87.6	88.2	25	58.01	51.1	54.5
ResNet-50	53.3	48.6	48.6	48.6	**93.3**	**93.3**	**95.8**	**94.6**	60.0	60.0	63.8	61.8
Inception-v3	62.5	66.7	66.7	66.6	53.3	68.9	63.8	66.3	66.7	73.8	68.1	70.8
Xception	75.0	80.8	79.3	80.0	85.7	85.9	73.6	79.3	75.0	80.8	79.3	80.0
DenseNet-201	65.0	61.7	61.1	61.4	75.0	80.8	79.3	80.0	75.0	80.7	79.3	80.0
SqueezeNet	25.0	58.1	51.1	54.4	87.5	88.9	87.6	88.2	25.0	58.0	51.2	54.4

5.3.2 PICCOLO dataset

A deep ensemble learning classifier is developed to effectively deal with the complex structure of colorectal polyps. A virtual biopsy is a sensitive and com- plex task that requires accurate classification of polyps into their respective classes for opportune polypectomy. Therefore, for improved polyp classification, the ensemble learning technique was developed.

In the case of the imbalanced dataset, the achieved macro-F1 scores were just average, and weighted-average ensemble models were 0.73 and 0.74 based on the test set. 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 classifiers suggests that assigning the suitable combination of weights to the base-classifiers generates promising results and performs better than a single base-learner and an aver- age ensemble model. In addition, the balanced data performed a lot better, yielding 0.76 and 0.79 based on the validation set and 0.76 and 0.84 on the test set, respectively. The results are shown in Tables 11 and 12. Macro and weighted F1-score shows that the base-classifiers were able to learn the complex representation of various polyp types.

The potential of multiple pre-trained CNNs with varying architectural designs is evaluated for the colorectal polyp classification problem. The performances of these classifiers do not produce exemplary results on colonoscopy images in contrast to the proposed technique. 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, as shown in Figs. 8 and 9. F1 score-based result comparison on

Table 9 Results with Optimizer: SGDM, Number of Epochs 100

Model(s)	0.001				0.005			
	Acc.	Precision	Recall	F1-score Acc.	Acc.	Precision	Recall	F1-score Acc.
GoogLeNet	87.5	88.9	87.5	88.2	66.7	67.9	72.2	70.0
ResNet-50	**93.3**	**93.3**	**95.8**	**94.6**	87.5	89.6	88.6	89.1
Inception-v3	53.3	68.9	63.8	66.3	60.0	70.0	52.8	60.2
Xception	85.7	85.9	73.6	79.3	72.0	78.3	75.2	76.7
DenseNet-201	75.0	80.8	79.8	80.0	62.5	67.9	72.2	70.0
SqueezeNet	50.0	51.8	50.4	51.1	50.0	50.8	51.4	51.1

Table 10 Best hyperparameter settings

N	Dataset	Model	Optimizer	No of epochs	Learning Rate	Acc.	Precision	Recall	F1-score
2 classes	Benchmark Dataset	ResNet-50	ADAM	100	0.001	86.7	91.7	91.6	91.7
	Balanced Dataset	ResNet-50	SGDM	100	0.001	86.6	91.7	91.6	91.7
3 classes	Benchmark Dataset	GoogLeNet	ADAM	100	0.001	87.5	86.6	86.3	87.1
	Balanced Dataset	ResNet-50	**SGDM**	**100**	**0.001**	**93.3**	**93.3**	**95.8**	**94.6**

Table 11 Performance evaluation of Multi-class Imbalanced and Balanced dataset

Dataset	Model(s)	Data Type	Accuracy	Precision	Recall	F1-score	Specificity	Sensitivity	TP	TN	FP
UCI Dataset	GoogLeNet	Imbalanced Dataset	86.7	87.6	86.6	87.1	0.87	0.84	65	141	10
	ResNet-50		73.3	75.4	77.2	76.3	0.71	0.78	56	132	24
	Ensemble Learning		88.2	88.9	88.3	88.6	0.86	0.90	67	143	11
	Weighted Average Ensemble Learning		**90.5**	**91.2**	**91.5**	**91.3**	**0.9**	**0.91**	**69**	**145**	**7**
	GoogLeNet	Balanced Dataset	87.5	88.9	87.2	87.6	0.87	0.87	95	204	14
	ResNet-50		92.7	91.8	93.5	92.7	0.92	0.93	101	210	9
	Ensemble Learning		93.6	93.6	93.6	93.6	0.93	0.93	102	211	7
	Weighted Average Ensemble Learning		**96.3**	**95.5**	**97.2**	**96.3**	**0.95**	**0.97**	**105**	**214**	**5**
PICCOLO Dataset	GoogLeNet	Imbalanced Dataset	73.1	75.2	72.4	73.2	0.75	0.71	255	593	84
	Xception		72.3	75.2	71.2	73.4	0.75	0.71	265	598	86
	ResNet-50		73.1	78.1	69.2	73.3	0.78	0.69	244	577	72
	Ensemble Learning		73.2	78.4	69.1	73.4	0.78	0.69	235	558	68
	Weighted Average Ensemble Learning		**74.4**	**78.2**	**71.3**	**74.2**	**0.78**	**0.71**	**244**	**572**	**70**
	GoogLeNet	Balanced Dataset	72.1	73.3	71.1	72.4	0.73	0.71	240	573	90
	Xception		72.4	71.1	73.2	72.3	0.71	0.73	239	572	98
	ResNet-50		73.2	73.4	72.3	73.3	0.73	0.72	242	575	90
	Ensemble Learning		77.3	79.1	74.2	77.1	0.79	0.74	253	591	66
	Weighted Average Ensemble Learning		**81.2**	**82.4**	**81.1**	**81.3**	**0.82**	**0.81**	**270**	**603**	**61**

Table 12 Performance of the Base-classifier and Proposed Ensemble Model on Imbalanced and Balanced Dataset

Dataset	Model(s)	Data Type	Macro Precision	Macro Recall	Macro F1-score	Weighted F1-score	Error	Cohen's Kappa Coefficient
UCI Dataset	GoogLeNet	Imbalanced Dataset	83.9	81.6	82.4	82.4	0.15	0.79
	ResNet-50		72.5	73.1	71.1	74.5	0.26	0.58
	Ensemble Learning		86.5	86.8	86.5	88.2	0.12	0.81
	Weighted Average Ensemble Learning		**88.8**	**88.9**	**88.9**	**91.0**	**0.09**	**0.85**
	GoogLeNet	Balanced Dataset	87.3	87.1	87.2	87.3	0.13	0.81
	ResNet-50		93.3	93.3	95.8	94.6	0.07	0.89
	Ensemble Learning		93.6	93.5	93.5	94.3	0.06	0.90
	Weighted Average Ensemble Learning		**92.3**	**96.5**	**96.3**	**96.1**	**0.04**	**0.94**
PICCOLO Dataset	GoogLeNet	Imbalanced Dataset	72.2	72.1	72.3	73.1	0.27	0.59
	Xception		70.1	71.3	71.2	71.1	0.28	0.55
	ResNet-50		73.3	74.2	73.2	69.4	0.27	0.61
	Ensemble Learning		73.1	73.4	73.2	73.3	0.27	0.61
	Weighted Average Ensemble Learning		**75.2**	**75.1**	**74.4**	**75.4**	**0.26**	**0.62**
	GoogLeNet	Balanced Dataset	72.3	72.1	72.3	72.2	0.28	0.59
	Xception		71.1	71.3	71.3	71.1	0.28	0.55
	ResNet-50		72.1	72.4	72.3	72.2	0.27	0.59
	Ensemble Learning		76.2	76.1	76.1	77.1	0.23	0.65
	Weighted Average Ensemble Learning		**81.3**	**81.4**	**84.2**	**84.3**	**0.19**	**0.68**

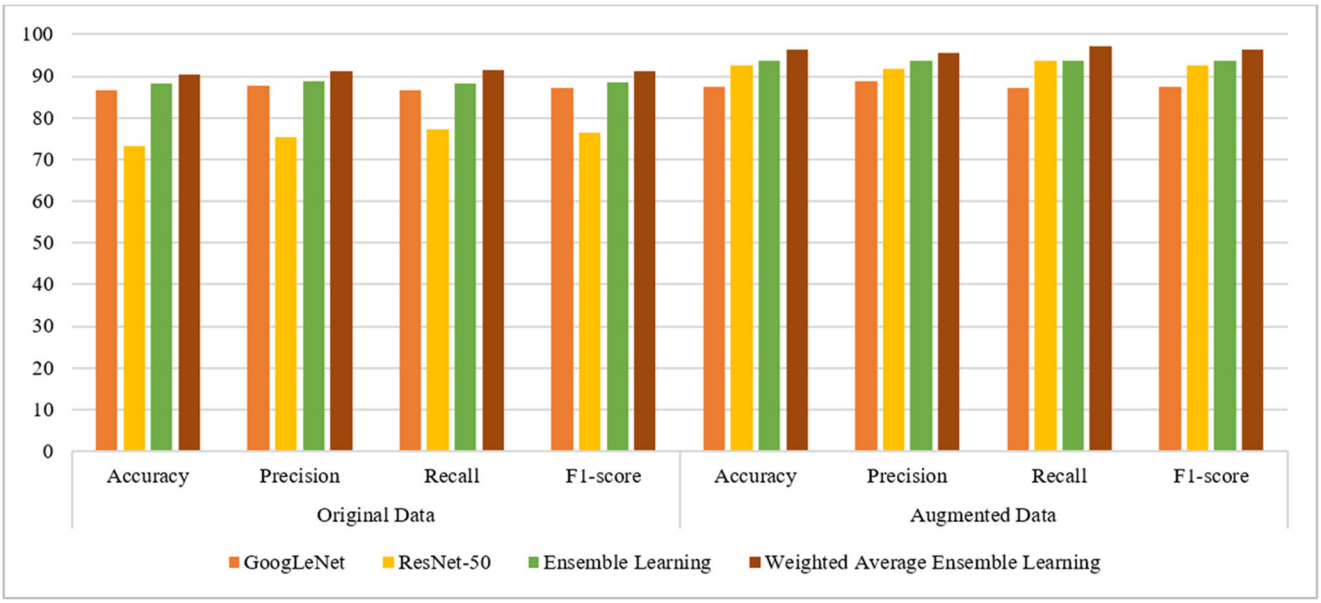

Fig. 8 Comparison of results for Imbalanced and Balanced data using Weighted Ensemble Learning

imbalanced and benchmark data is presented in Fig. 10. The proposed method shows a 3% increase in the macro F1-score on the test set for benchmark data. However, a 12% increase in macro F1-score on the test set was noticed as compared to the maximum value attained by the individual base-classifiers.

5.4 Precision-recall based analysis

In addition to sensitivity of model, it is extremely important to analysis the precision of the proposed system. Precision represents the correctly identified positive cases out of all the positive instance of the data. A small fraction of false positive values can considerably affect the precision of the of the CAD system if the data is imbalanced and decreases the F1-score. In medical domain, where data is usually imbalanced, where mostly cases belong to a larger class and less cases belong to a smaller, yet usually more interesting class. As a result, such systems misclassify the minority instances as majority class, generating a high false negative rate [37]. In such systems, the cost is usually high when a classifier misclassifies the positive class examples and this misclassifica tion can affect the system performance and have an adverse effect on diagnosis. Therefore, the proposed system handled the class imbalance to decrease the false positive and negative predictions. The precision of the proposed system is 95.5 on UCI dataset for balanced data which indicates a good capability of the system

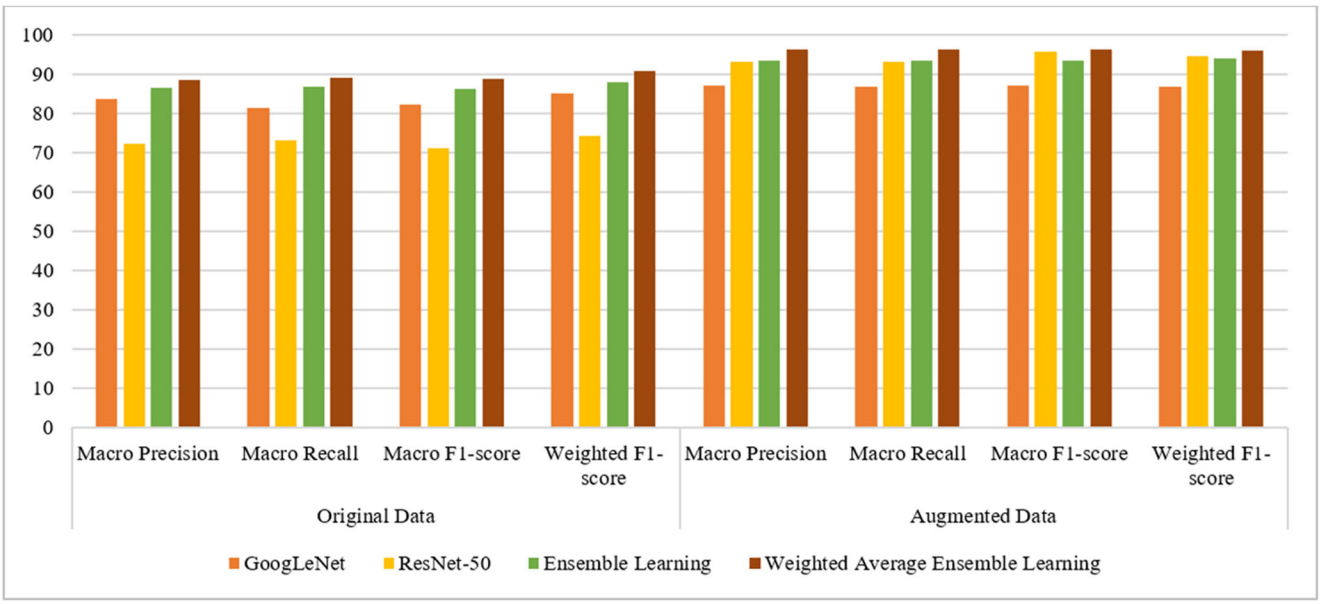

Fig. 9 Comparison of macro results for Imbalanced and Balanced data using Weighted Ensemble Learning

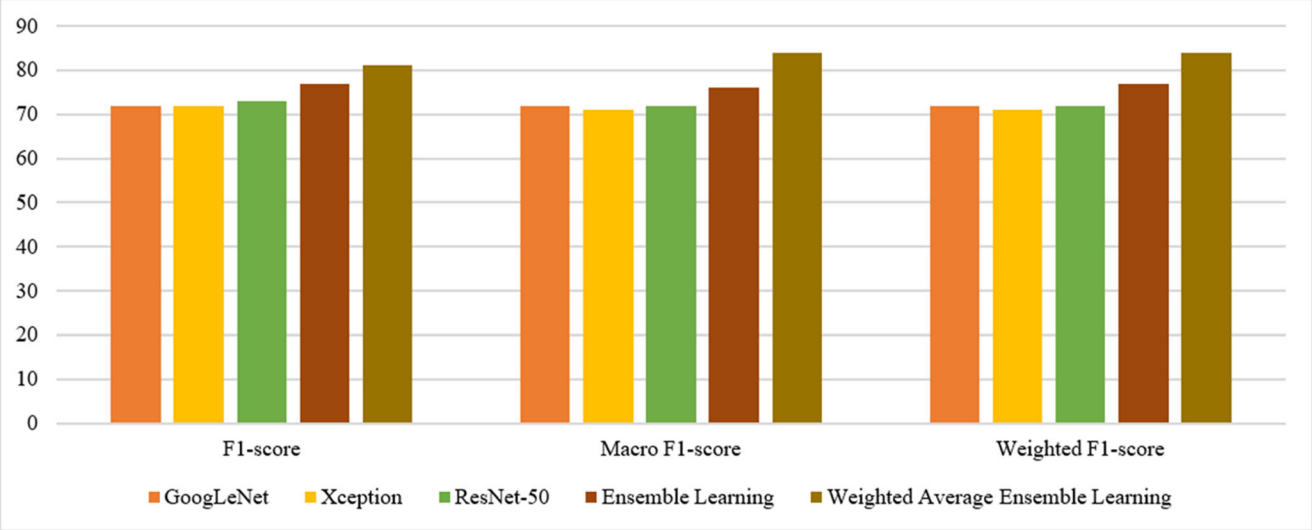

Fig. 10 Performance of base-classifier and ensemble classifier on Balanced dataset

to identify positive cases shown in Fig. 11. Macro precision and false positives comparison of proposed approach based on both imbalanced and balanced PICCOLO dataset is presented in Fig. 12. The precision of the proposed system on this dataset is 0.81 for balanced data which indicates a good capability of the system to identify positive cases.

5.5 Reliability analysis

Figure 13 shows a comparison of imbalanced and balanced data of error rate and kappa coefficient values of UCI dataset. Kappa value for base-classifiers: GoogLeNet and ResNet-50 are 0.79 and 0.58 on benchmark data whereas 0.81 and 0.89 respectively. However, in terms of ensemble classifiers, average ensem- ble generate 0.90 kappa value and weighted ensemble further improves the result to 0.94.

Figure 14 shows a comparison of kappa coefficient and error values of PICCOLO dataset. Kappa value for base-classifiers: GoogLeNet, Xception, ResNet-50 are 0.59, 0.55, 0.59. However, in terms of ensemble classifiers, average ensemble generate 0.61 kappa value and weighted ensemble further improves the result to 0.62. Graph shows thatwith the increase in kappa coefficient, error value of the model decreases in both scenarios. This significant increase in the kappa coefficient indicates a that proposed ensemble method has an acceptable degree of reliability.

Figure 15 shows the ROC-AUC, 0.94 value that indicates that proposed model has good degree of separability for PICCOLO dataset and Fig. 16 shows the ROC-AUC, 0.89

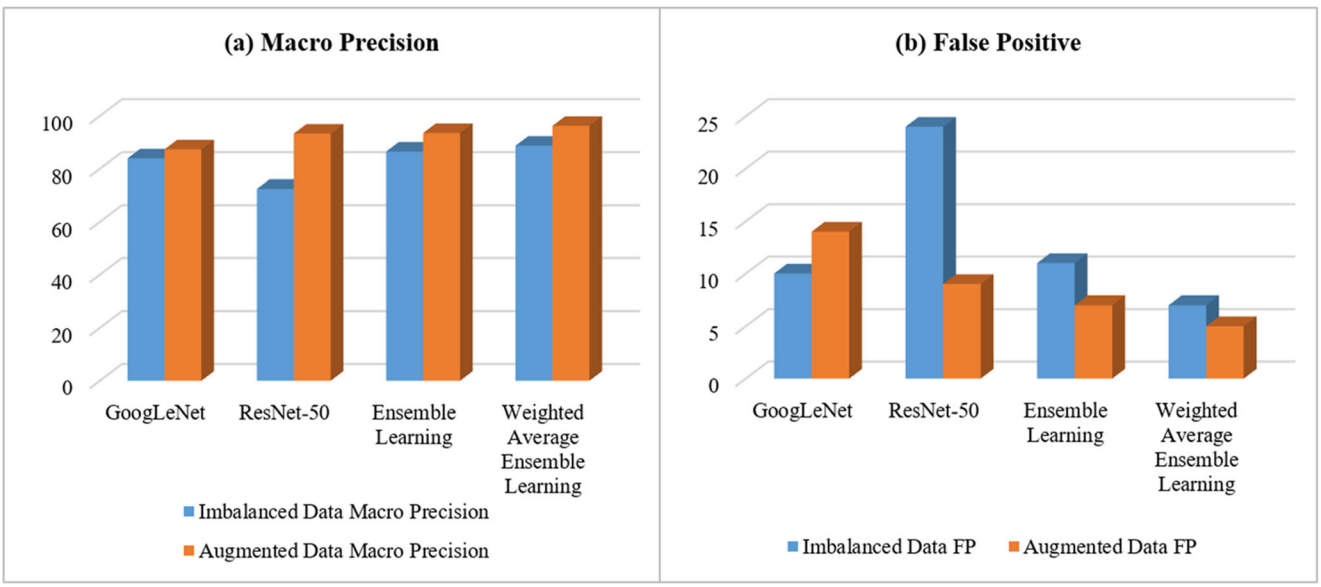

Fig. 11 **a** Macro Precision and **b** False Positive rate comparison of Imbalanced and Balanced UCI dataset

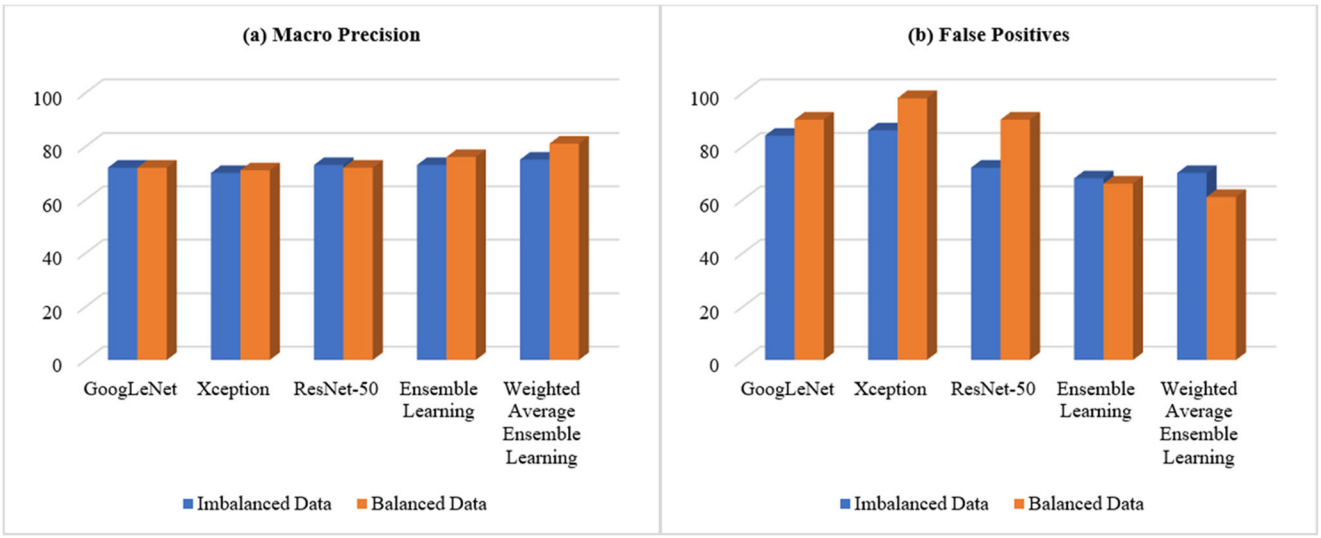

Fig. 12 **a** Macro Precision and **b** False Positive rate comparison of Imbalanced and Balanced PICCOLO dataset

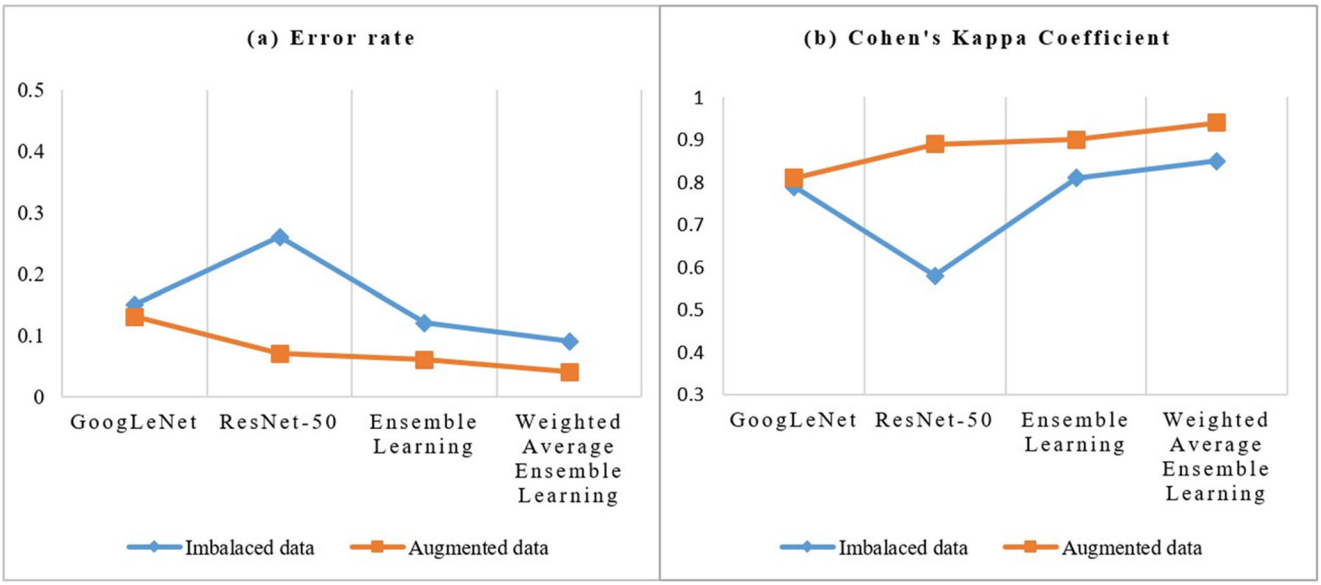

Fig. 13 Reliability comparison of proposed model using Cohen's Kappa Coefficient on UCI Dataset

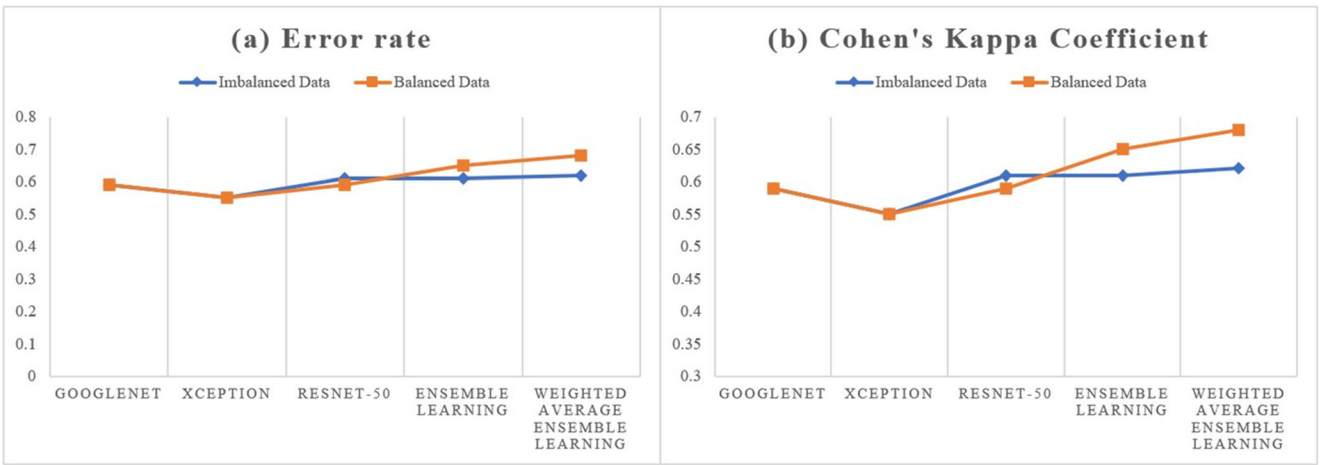

Fig. 14 Reliability comparison of proposed model using Cohen's Kappa Coefficient on PICCOLO Dataset

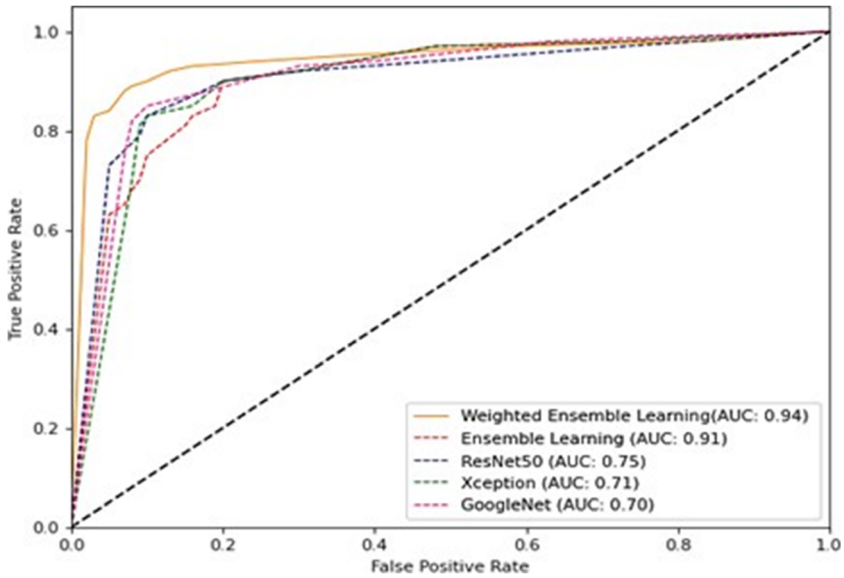

Fig. 15 ROC curves on PICCOLO dataset

for two classes classification and 0.91 for three classes classification on UCI dataset obtained values that indicates that proposed model has good degree of separability.

6 Comparison of models performance

The comparison with other methods has been very difficult as this publicly available dataset is used by a limited number of studies. Our work is com- pared with two studies that have performed colorectal polyp classification on the same benchmark dataset. The comparison of results is shown in Table 13

Zhang et al. [28] proposed a CNN-based transfer learning framework where features learned from the non-medical dataset were utilised. They investi- gated two-class classification; therefore, for comparison of the results with our approach, it is essential to examine our results from this point of view as well. Therefore, with the established optimised hyperparameter configuration obtained by our experiments, colorectal polyps were classified into hyperplas-tic and adenomatous polyps. Two class classification was performed by the work that yielded 85.9% accuracy, 87.3% precision, 87.6% recall and 87.0% F1-scores. Our proposed weighted average ensemble approach improved the performance of the classifier by 2% on imbalanced data and 3% on balanced data. Our framework outperformed their approach by producing 86.6% accu- racy, 91.7% precision, 91.7% recall and 91.7% F1-score with both benchmark and oversampled data.

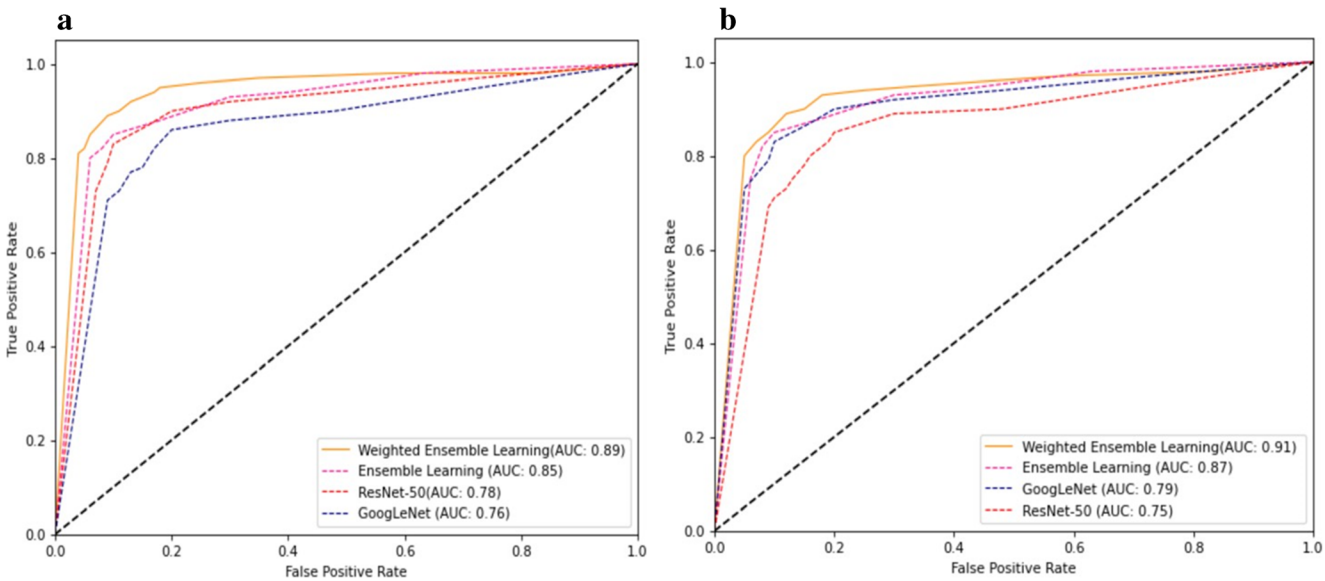

Fig. 16 ROC of (**a**) Two classes classification (**b**) Three classes classification on UCI Dataset

Table 13 Comparative Analysis with existing studies

Dataset	No of classes	Method		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
UCI Dataset	**2 Classes**	**Transfer Learning - Places 205** [27]		85.9	87.3	87.6	86.0
		Imbalanced Data	Optimized Network	86.6	91.7	91.6	91.7
			Ensemble Learning	87.9	91.9	91.7	91.3
			Weighted Ensemble Learning	**88.1**	**92.4**	**91.9**	**91.5**
		Balanced Data	Optimized Network	86.7	91.7	91.6	91.7
			Ensemble Learning	88.3	91.9	92.2	91.8
			Weighted Ensemble Learning	**89.8**	**92.3**	**92.1**	**92.2**
	3 Classes	**SVM** [28]		82.5	N.A.	72.7	N.A.
		Imbalanced Data	Optimized Network	86.7	86.6	86.6	87.1
			Ensemble Learning	88.2	88.9	88.3	88.6
			Weighted Ensemble Learning	**90.1**	**91.2**	**91.5**	**91.3**
		Balanced Data	Optimized Network	92.7	91.8	93.5	92.7
			Ensemble Learning	93.6	93.6	93.6	93.6
			Weighted Ensemble Learning	**96.3**	**95.5**	**97.2**	**96.3**
PICCOLO Dataset	**3 Classes**	Imbalanced Data	Optimized Network	73.1	78.1	69.2	73.3
			Ensemble Learning	73.2	78.4	69.1	73.4
			Weighted Ensemble Learning	**74.4**	**78.2**	**71.3**	**74.2**
		Balanced Data	Optimized Network	73.2	73.4	72.3	73.3
			Ensemble Learning	77.3	79.1	74.2	77.1
			Weighted Ensemble Learning	**81.2**	**82.4**	**81.1**	**81.3**
CVC-Clinic for training, CGMH-WL, CHMH-NBI for testing [38]	**3 Classes**	N.A	CNN model-NBI	82.8	82.5	95.2	81.5
			CNN model-WL	72.2	75.3	88.7	81.5
Children's medicine department of Gachon University Gil hospital in South Korea [39]	**3 Classes**	N.A	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

The other comparative evaluation was accomplished [27] where machine learning and computer vision algorithms were combined together to develop a three-class classification framework for implementing virtual biopsy by classifying colorectal polyps into hyperplastic lesions, serrated adenomas, and adenomas. Machine learning classifiers incorporated by this research work were Random Forest (RF), Random Subspace (RS) and Support Vector Machine (SVM). The results obtained by this approach were 82.46% accuracy, 72.74% sensitivity, 85.88% specificity. On comparing with our results of three-class classification, it was observed that the framework proposed by our study out- performs both traditional machine learning and deep learning approach by producing 90.1% vs 96.3% accuracy and 91.5% vs 97.2% recall on benchmark data and oversampled data, respectively.

This study experiments with various deep learning models for the classification of colorectal polyps, such as GoogleNet, ResNET-50, Ensemble Learning, and Weighted Average Ensemble Learning. Subsequently, the results are compared with the published studies in regard to the classification of polyps using deep learning models. It can be observed that the highest accuracy achieved is 82.8% by using the CNN model proposed by Chen et al. [38]. Further, AlexNet is used as a backbone in the transfer learning model proposed by Kim et al. [39] in which the highest accuracy of 0.79 was achieved with the variations of fully connected networks. However, this study outperforms these recent investigations by achieving the highest accuracy of 96.3 and 90.5 on both balanced and imbalanced data using Weighted Average Ensemble Learning.

7 Conclusion and future work

In this paper, we present a framework designed for the classification of col- orectal polyps with the minimum amount of preprocessing. Early detection and classification of polyps mitigate colorectal cancer-related deaths. Aimed at successful classification, the large dataset is essential, whereas the benchmark dataset in this project was very small. It was observed that if the experiments are performed on the benchmark dataset, the results obtained are not very accurate. However, transfer learning conducted on processed data significantly enhances the performance of pre-trained CNN architectures. A comparative analysis of several pre-trained CNN architectures was conducted to establish the best hyperparameter settings to obtain better results of evaluation metrics. Our results show that the proposed method classifies polyps with 90.1% accuracy and 91.5% recall on benchmark data. In addition, this dataset also has a high degree of imbalance, as one type of polyps is more prevalent than the rare types. Handling this class imbalance has shown a significant improvement in results from 90.1% to 96.3% accuracy. The assessment of results shows that the proposed method maintains a rea- sonable detection rate with a small deviation in macro F1- score. Among the base classifiers, GoogLeNet produced the best results (0.82 macro f1-score) on benchmark data with optimised hyperparameter configuration, whereas ResNet-50 (0.93 macro f1-score) outperformed other networks when tested on balanced data.

The improvement in macro F1-score (0.89) of the weighted average ensemble from 0.86 of average ensemble classifier proposes that the developed method is suitable for multi-class classification tasks on imbalanced data. The utilisation of the 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 after being evaluated using Cohen's Kappa Coefficient. Moreover, the performance of the proposed model shows that it has attained an accurate diagnosis. A higher recall rate indicates that the sensitivity of the classification is high; it can classify all three polyp types correctly, particularly the serrated adenoma and the hybrid polyp, which are difficult to be classified. All the factors are essential for accurate CAD. In addition, it benefits us greatly in completing the virtual biopsy where endoscopists can decide which polyps should be directly resected and which should be sent for biopsy. The proposed architecture with the best hyperparameter settings outperformed the previous methods, which conducted the experiments on the same colonoscopy dataset used in this paper. The promising results generated in our experiments show that this proposed method is beneficial for endoscopists for the identification of different types of polyps.

In future, training on a customised deep network could be designed for accu- rate classification, though it requires a decent number of images in the dataset and creating a labelled large medical dataset is a challenging task. Another approach would be to perform polyp detection prior to classification as well as include white light images in addition to narrowband imaging for efficient classification of diverse images.

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

Data Availability Statement • The UCI dataset analysed during the current study is available at http://www.depeca.uah.es/colonoscopy dataset/.

• The PICCOLO dataset that supports the findings of this study is available from Basque Biobank, but restrictions apply to the availability of this data, which was used under licence for the current study, and so is not publicly available. Data is however available with permission from Basque Biobank at https://www.biobancovasco.org/en/Sample-and-data-catalog/ Databases/PD178-PICCOLO-EN.html.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

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. Pacal I, Karaboga D, Alper B, Akay B, Nalbantoglu U (2020) A comprehensive review of deep learning in colon cancer. Comput Biol Med 126:104003
2. Jha D, Ali S, Hicks S, Thambawita V, Borgli H, Pia HS, de Lange T, Pogorelov K, Wang X, Harzig P et al (2021) A comprehensive analysis of classification methods in gastrointestinal endoscopy imaging. Med Image Anal 70:102007
3. Michael G, Stephan K, Erich L, Bernhard G, Christiane S, Eva B, Hans F, Werner W (2002) High-grade dysplasia and invasive carcinoma in colorectal adenomas: a multivariate analysis of the impact of adenoma and patient characteristics. Eur J Gastroenterol Hepatol 14(2):183–188
4. Tajbakhsh N, Jae YS, Suryakanth RG, Hurst TR, Kendall CB, Gotway MB, Liang J (2016) Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans Med Imaging 35(5):1299–1312
5. Younghak S, Qadir HA, Balasingham I (2018) Abnormal colon polyp image synthesis using conditional adversarial networks for improved detection performance. IEEE Access 6:56007–56017
6. Kopelman Y, Gal O, Jacob H, Siersema PD, Cohen A, Eliakim R, Zaltshendler M, Zur D (2019) Automated polyp detection system in colonoscopy using deep learning and image processing techniques, Journal of Gastroenterology and its Complications 3(1):101
7. Misawa M, Kudo S, Mori Y, Cho T, Kataoka S, Yamauchi A, Ogawa Y, Maeda Y, Takeda K, Ichimasa K, Nakamura H, Yagawa Y, Toyoshima N, Ogata N, Kudo T, Hisayuki T, Hayashi T, Wakamura K, Baba T et al (2018) Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology 154(8):2027–2029
8. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, Jeroen AWM, van DL, Van BG, Sanchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88
9. Kim J, Hong J, Park H (2018) Prospects of deep learning for medical imaging. Precision Future Med 2(2):37–52
10. Shen D, Wu G, Suk H (2017) Deep learning in medical image analysis. Annu Rev Biomed Eng 19:221–248
11. Thany SH, Tricoire-Leignel H, Lapied B (2010) Identification of cholinergic synaptic transmission in the insect nervous system. Adv Exp Med Biol 683:1–10
12. Liew WS, Tang TB, Lin C, Lu C (2021) Automatic colonic polyp detection using integration of modified deep residual convolutional neural network and ensemble learning approaches. Comput Methods Prog Biomed 206:106114
13. Bengio Y, LeCun Y et al (2007) Scaling learning algorithms towards AI. Large- scale kernel machines 34(5):1–41
14. Huang Y, Wu Z, Wang L, Tan T (2013) Feature coding in image classification: a comprehensive study. IEEE Trans Pattern Anal Mach Intell 36(3):493–506
15. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, Summers RM, Giger ML (2019) Deep learning in medical imaging and radiation therapy. Med Phys 46(1):e1–e36
16. Khan A, Sohail A, Zahoora U, Qureshi AS (2020) A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53(8):5455–5516
17. Komeda Y, Handa H, Watanabe T, Nomura T, Kitahashi M, Sakurai T, Okamoto A, Minami T, Kono M, Arizumi T, Takenaka M, Hagiwara S, Matsui S, Nishida N, Kashida H, Kudo M (2017) Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience. Oncology 93(Suppl. 1):30–34
18. Kudo S, Mori Y, Misawa M, Takeda K, Kudo T, Itoh H, Oda M, Mori K (2019) Artificial intelligence and colonoscopy: current status and future perspectives. Dig Endosc 31(4):363–371
19. Sánchez-Peralta LF, Bote-Curiel L, Picón A, Sánchez-Margallo FM, Pagador JB (2020) Deep learning to find colorectal polyps in colonoscopy: A sys- tematic literature review. Artif Intell Med 108: 101923
20. Rodríguez N, Alba RCD, Hugo Fernández L, Iglesias A, Joaquín Cubiella F, Florentino Riverola F, Miguel Jato R, Daniel Pena G (2021) Deep neural networks approaches for detecting and classifying colorectal polyps. Neurocomputing 423:721–734
21. Bauer VP, Papaconstantinou HT (2008) Management of serrated adenomas and hyperplastic polyps. Clin Colon Rectal Surg 21(04): 273–279
22. Butterly LF, Chase MP, Pohl H, Fiarman GS (2006) Prevalence of clin- ically important histology in small adenomas. Clin Gastroenterol Hepatol 4(3):343–348
23. Patino-Barrientos S, Sierra-Sosa D, Garcia-Zapirain B, Castillo-Olea C, Elmaghraby A (2020) Kudo's classification for colon polyps' assessment using a deep learning approach. Appl Sci 10(2):501
24. 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 13:1756284820910659
25. Wei JW, Suriawinata AA, Vaickus LJ, Ren B, Liu X, Lisovsky M, Tomita N, Abdollahi B, Kim AS, Snover DC et al (2020) Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides. JAMA Netw Open 3(4):e203398–e203398
26. Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, Suriawinata AA, Hassanpour S (2017) Deep learning for clas- sification of colorectal polyps on whole-slide images. J Pathol Inf 8:30
27. Mesejo P, Pizarro D, Abergel A, Rouquette O, Beorchia S, Poincloux L, Bartoli A (2016) Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE Trans Med Imaging 35(9):2051–2063
28. Zhang R, Zheng Y, Mak TWC, Yu R, Wong SH, Lau JY, Poon CCY (2016) Automatic detection and classification of colorectal polyps by trans- ferring low-level cnn features from nonmedical domain. IEEE J Biomed Health Inform 21(1):41–47
29. Pacal I, Karaman A, Karaboga D, Akay B, Basturk A, Nalbantoglu U, Coskun S (2022) An efficient real-time colonic polyp detection with yolo algo- rithms trained by using negative samples and large datasets. Comput Biol Med 141:105031

30. Nogueira-Rodríguez A, Domínguez-Carbajales R, Campos-Tato F, Herrero J, Puga M, Remedios D, Rivas L, Sánchez E, Iglesias Á, Cubiella J et al (2021) Real-time polyp detection model using convolutional neural networks. Neural Comput Applic 1–22. https://doi.org/10.1007/s00521-021-06496-4
31. 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):138–144
32. Poudel S, Kim YJ, Vo DM, Lee S (2020) Colorectal disease classification using efficiently scaled dilation in convolutional neural network. IEEE Access 8:99227–99238
33. Rahman MM, Wadud Md AH, Hasan MM (2021) Computerized classifi- cation of gastrointestinal polyps using stacking ensemble of convolutional neural network. Inf Med Unlocked 24:100603
34. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med 22(3):276–282
35. Pacal I, Karaboga D (2021) A robust real-time deep learning based automatic polyp detection system. Comput Biol Med 134:104519
36. Brownlee J (2019) Ensemble learning methods for deep learning neural networks. https://machinelearningmastery.com/ensemble-methods-for-deep-learning-neural-networks/. Deep Learning Performance [Accessed: 2021-12-30]
37. Liu Y, Yu X, Huang JX, An A (2011) Combining integrated sampling with svm ensembles for learning from imbalanced datasets. Inf Process Manag 47(4):617–631
38. Hsu C, Hsu C, Hsu Z, Shih F, Chang M, Chen T (2021) Colorectal polyp image detection and classification through grayscale images and deep learning. Sensors 21(18):5995
39. Kim YJ, Bae JP, Chung J, Park DK, Kim KG, Kim YJ (2021) New polyp image classification technique using transfer learning of network-in-network structure in endoscopic images. Sci Rep 11(1):1–8

Farah Younas is a PhD Candidate at Auckland University of Technology, New Zealand. She has completed MSc in Advanced Computer Science from Manchester Metropolitan University, Manchester, England. She worked as a Lecturer in the Department of Computing at Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad, Pakistan. Her research interest include Deep Learning, Big Data and Healthcare Informatics.