IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 October 2023, accepted 5 November 2023, date of publication 15 November 2023, date of current version 20 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3332826

RESEARCH ARTICLE

XGBoosted Binary CNNs for Multi-Class Classification of Colorectal Polyp Size

PROMIT HALDAR[1], VANSHALI SHARMA[2], YUJI IWAHORI[3], (Member, IEEE), M. K. BHUYAN[1], (Senior Member, IEEE), AILI WANG[4], (Member, IEEE), HAIBIN WU[4], AND KUNIO KASUGAI[5]

[1]Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India
[2]Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India
[3]Department of Computer Science, Chubu University, Kasugai 487-8501, Japan
[4]Higher Educational Key Laboratory for Measuring and Control Technology and Instrumentations of Heilongjiang, Harbin University of Science and Technology, Harbin 150080, China
[5]Department of Gastroenterology, Aichi Medical University, Nagakute, Aichi 480-1195, Japan

Corresponding author: Vanshali Sharma (vanshalisharma@iitg.ac.in)

This work was supported in part by the High End Foreign Experts Introduction Program under Grant G2022012010L, and in part by the Reserved Leaders of Heilongjiang Provincial Leading Talent Echelon 2021. The work of Yuji Iwahori was supported in part by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (C) under Grant 20K11873, and in part by the Chubu University Grant. The work of Vanshali Sharma was supported by the Department of Science and Technology, Government of India, through the Innovation in Science Pursuit for Inspired Research (INSPIRE) Fellowship under Grant IF190362.

ABSTRACT Colorectal cancer (CRC) is marked by the development of tumors/outgrowths known as polyps. AI-assisted endoscopy is inevitable in the modern world for better and more efficient polyp detection and classification. Often, the risk associated with CRC is indicated by the polyp's size. Automated size classification of colorectal polyps from endoscopic images is a boon to endoscopists to monitor and diagnose the polyps. While previous research efforts have predominantly centered around the pathological categorization of polyps, limited attention has been directed towards the classification of polyp size. In this paper, we have proposed a deep learning-based model for the multi-class classification of colorectal polyps into four classes: 0-5 mm, 5-10 mm, 10-14 mm, and >=14 mm. A narrow range in polyp size classification provides more information about the growth of the polyp as opposed to binary classification. We also show that the One vs Rest classification technique using binary classifiers outperforms the usual approach of using a single CNN for multi-class classification. Also, we use XGBoost with the binary classifiers to further increase the performance of the model. The experimental results report the effectiveness of our proposed model in performing multi-class polyp size classification. The approach is expected to assist clinicians in estimating polyp size efficiently.

INDEX TERMS Endoscopy, colorectal polyps, size classification, depth maps, convolutional neural networks, xgboost, one vs. rest classification.

I. INTRODUCTION

Colorectal cancer (CRC) is the third most deadly and fourth most commonly diagnosed cancer in the world [1], [2]. Hence, early detection of colorectal polyps is crucial for the diagnosis and treatment of CRC. Traditionally, colonoscopy has been used for detecting such polyps and examining the digestive tract. This non-surgical procedure involves the insertion of a colonoscope (a light, flexible tube with a light and a camera), which enables the doctor to get views of the colon and look for colorectal polyps. Such polyps may be neoplastic or non-neoplastic. Non-neoplastic polyps are non-malignant tissues, whereas neoplastic polyps are potentially malignant. These polyps possess camouflage properties and appear in large variations in terms of color, size, and shape. Manually reviewing such polyps is a cumbersome task and involves a high miss rate. An automatic, minimally invasive procedure for polyp detection and classification is thus extremely helpful for identifying and characterizing polyps.

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

Artificial Intelligence (AI)-assisted colonoscopy provides an additional source for colonoscopists to reduce missed detections via the naked eye. Such techniques are highly desirable when there are a large number of patients to be examined within a short period of time and the number of available doctors is low. AI-based systems enable the characterization of polyps in terms of the absolute size of polyps and pathological diagnosis, which enables proper screening, diagnosis, and treatment. The polyp sizes provide critical information as larger polyps are associated with more risk of cancer. Smaller polyps are less likely to be cancerous, whereas polyps larger than 10 mm have a greater probability of being cancerous. This makes it crucial to estimate the polyp size for correct decision-making. However, most of the previous works in the area of colonoscopy image analysis have focused on the polyp classification based on pathological diagnosis [1], [3], [4], [5] i.e., Hyperplastic polyp, Sessile serrated lesion, Low-grade adenoma, etc., and ignored the important criteria of polyp size. Also, a substantial amount of research has been dedicated to the detection and segmentation of colorectal polyps from endoscopic images [6], [7], [8], [9]. The classification of polyps based on their absolute sizes has not been much explored in the existing research.

Itoh et al. [10] proposed a binary polyp-size classification method that estimates a polyp's three-dimensional spatial information using a combination of polyp localization and depth estimation (i.e., localized depth maps) with a reported accuracy of 0.88 on 787 polyps of both protruded and flat types into less than 10 mm and greater than equal to 10 mm. Itoh et al. [11] also proposed an approach using RGBD images to classify colorectal polyps based on their absolute size. They achieved binary and trinary polyp-size classification with 79% and 74% accuracy from a single still image of a colonoscopic video. Abdelrahim et al. [12] used two approaches for the binary classification of colorectal polyps according to their absolute size, into less than or equal to 5 mm and greater than 5 mm. They developed a deep learning model based on convolutional neural networks (CNNs) trained on RGB polyp images and found 80% accuracy in 10 videos of human polyps. Chadebecq et al. [13] proposed the Infocus-Breakpoint (IB) technique to estimate an image-wise scale by detecting the blur/unblur break-point in a video sequence. They simultaneously tracked a polyp with a 2D affine transformation and estimated the amount of defocus blur, which led to an area-wise scale estimate. An image-wise estimation of the defocus blur allowed extraction of the IB (sharpest image of the sequence), and the depth of the scene corresponding to the IB is known by calibration. They assumed that the polyp is planar and front parallel to the gastroscope's tip to approximate the size of the polyp. They evaluated their method on three colonoscopic sequences of humans. For the first video sequence, the relative error of estimation of their method is 7%; for the second sequence, the error of estimation is 6%, and 1% for the third sequence. Villard et al. [14] proposed Siamese Networks for binary classification of colorectal polyps based on size less than or above 10 mm. They trained Siamese networks to build a high-dimensional feature embedding extracted for each polyp size. As a second step, they used a k-NN approach to classify polyp sizes based on the distance between the feature embedding of the input image and the whole embedding space learned by the Siamese Network. They tested their model on 2,688 images and obtained 79.2% accuracy in feature classification and 95.7% in polyp size classification.

The existing works focused on polyp size classification mainly considered binary classification and rarely performed ternary classification, thus providing less precise information about the polyp size. Classifying colorectal polyps into narrower ranges enables us to obtain the absolute size of polyps from endoscopic images directly without requiring additional equipment during endoscopy, which otherwise will be cumbersome. Also, endoscopic images related to the identical polyps taken at different time intervals (after two months, six months, etc.) would enable the doctors to determine the growth of polyps in terms of size, which might require significant medical attention. Therefore, in this paper, we have performed multi-class classification of colorectal polyp images according to their absolute size (in mm) using CNNs and XGBoost classifiers. For multi-class classification of images, the traditional approach uses only a single CNN. Unlike this traditional approach, we used four different binary classifiers to perform multi-class classification and further combined each binary CNN with a corresponding XGBoost classifier to improve the binary classification performance of each binary classifier. This approach also improved the test accuracy to 87.07% and the F1-score to 86.95% for multi-class classification. Similar to Itoh et al. [10], localized depth maps of colorectal polyps were used as input to the models. However, our method is different in terms of architecture and the objective of multi-class classification. We have classified colorectal polyps into four classes: polyps sized within 0-5 mm, 5-10 mm, 10-14 mm, and $\geq$ 14 mm. The main contributions of the proposed work are summarized below:

- We have proposed a deep learning based model with multiple binary CNNs and XGBoost classifiers to perform multi-class classification of colorectal polyp size.
- Unlike existing works focusing on the binary classification of polyp size, we performed multi-class classification considering four classes. This helps in getting precise polyp size estimates.
- Our model outperformed the baseline method by 2.2% and 2.48% in terms of accuracy and F1-score, respectively.

II. MATERIALS AND METHODS

A. OVERVIEW

In this paper, we have proposed an approach to perform multi-class classification to determine the colonic polyp size. This is a novel application-based framework in the domain

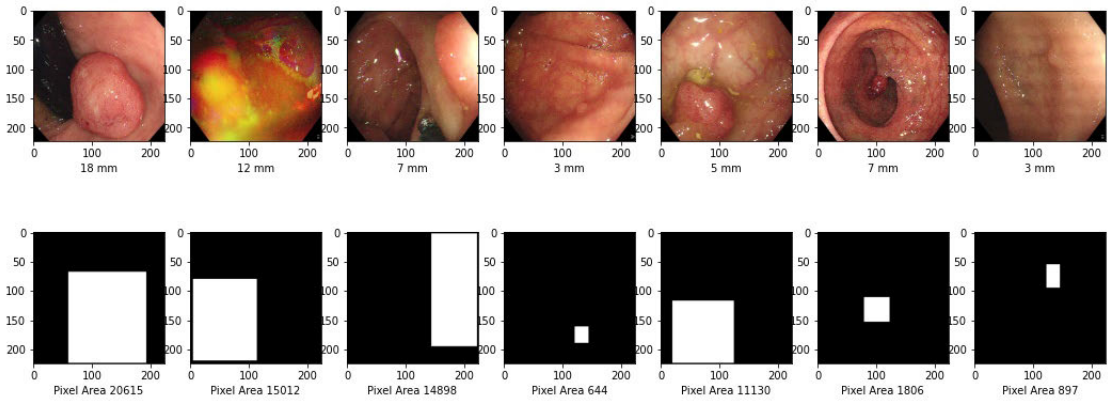


FIGURE 1. Polyp localization maps with polyp size and corresponding pixel area. The first row shows the original RGB images with polyp size, and the second row presents the localization maps with the corresponding pixel area.

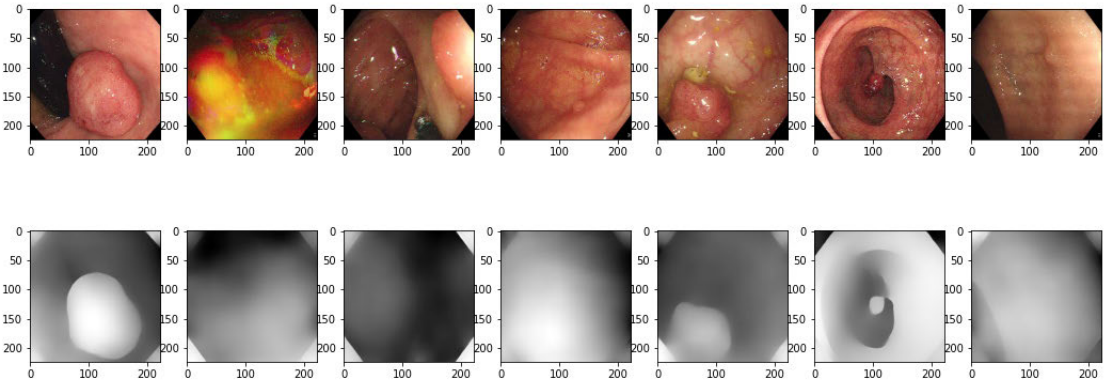


FIGURE 2. The first row presents the original RGB polyp images, and the second row shows the corresponding depth maps.

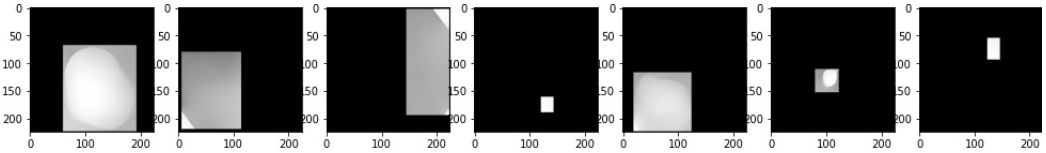


FIGURE 3. Polyp localized depth maps pertaining to only the region of interest are shown.

of Colonoscopy that focuses on polyp size estimation. Our method uses the One vs. Rest classification technique in which N number of binary classifiers (M_0,M_1,…,M_N) are trained where N is the number of classes. The goal of the classifier M_i is to classify between i^{th} class and the rest (all other classes except i^{th} class). During inference, the input image is passed through all of the four models to obtain probabilities of the respective classes. The class corresponding to the maximum probability is chosen as the predicted class. For our objective, we used four binary CNN classifiers with four classes, as shown in Table 1. The CNNs used for training consist of only three convolutional layers and three linear layers. To further increase the performance of the binary classifiers, after each binary CNN M_i is trained, we obtain predictions from the last hidden layer of the CNN (which consists of 512 features here) and train an XGBoost Classifier XG_i on the same. This process is followed for each of the four binary CNNs trained. Finally, the individual binary classifiers (CNN and XGBoost) are combined via the One vs. Rest classification technique, which predicts the class with maximum confidence in probability. The localized depth maps of the corresponding polyp RGB images have been used as the input feature to the model.

B. LOCALIZED DEPTH MAPS

In a localized depth map of an RGB polyp image, the region containing the polyp is substituted with the corresponding depth map pixel values, while all other pixels are set to zero (blacked out). In this way, we have information about both the area associated with the polyp and the relative distance

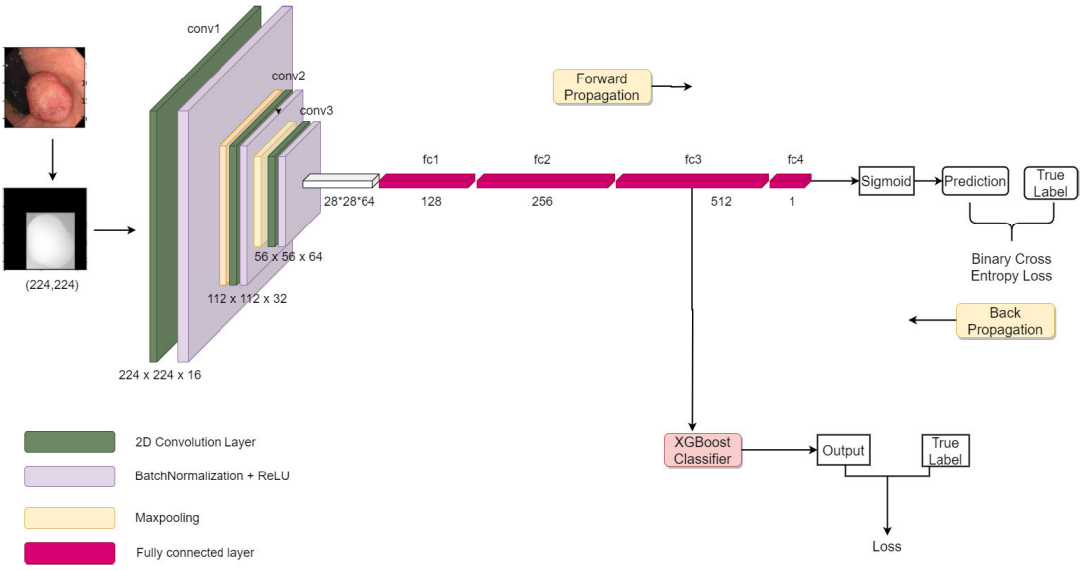


FIGURE 4. Model architecture (Binary CNN with XGBoost). fc1 to fc4 are the fully connected layers. fc4 is dropped after the training of the binary CNN classifier, and instead, an XGBoost classifier is attached.

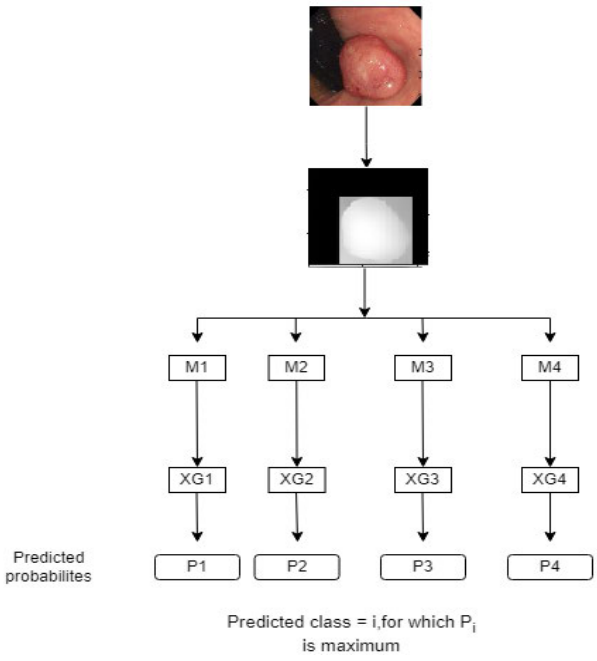


FIGURE 5. Multi-class classification inference. The figure depicts the different binary CNNs along with the attached respective XGBoost classifiers. The architecture of M1 to M4 is the same as the single binary CNN. All these modules are attached to XGBoost classifiers denoted by XG1 to XG4.

of the polyp from the camera (due to the depth map pixel values). This distance information is vital for accurately estimating the actual physical size of the polyp. Without depth information, the model might struggle to differentiate between different-sized polyps, which might appear to be similar (in terms of the area of the polyp) in the 2D image. On a similar basis, the model might differentiate between two polyps of the same size with varying appearances in the 2D image due to different distances from the camera. Including depth information takes care of these limitations. Note that to obtain the localized polyp region, we have used the available ground truth provided in the dataset. There is no absolute relation between the size estimation that we performed and the bounding box annotations. The purpose of using bounding box annotations in the approach is to make the model focus on the region of interest rather than any unnecessary background details. During inference, the localized depth map of the polyp RGB image is obtained by using the DPT model [15]. The model generates the depth map, and localization details of the polyp area are obtained using annotations provided in the dataset. Some sample polyp localization masks and polyp depth maps are shown in Fig. 1 and Fig. 2, respectively. Combining both polyp localization masks and corresponding depth maps, we obtain polyp localized depth maps (see Fig. 3).

C. MODEL ARCHITECTURE

In this subsection, we explain the architecture of the binary classifiers. In total, we have trained 8 different models for this approach. Four binary classifying CNNs (M_0, M_1,M_2, M_3) have been trained for each of the classes, and for each CNN, we have trained a corresponding XGBoost classifier (XG_0, XG_1, XG_2, XG_3) which takes the output of the last hidden layer of the respective CNN as input. At first, we trained all four binary CNN models (M_0, M_1, M_2, M_3). After training, we extracted the output from the last hidden layer of each CNN model and trained an XGBoost classifier for each of the four models. Model M_i classifies the i^{th} class as 1 and all other classes as 0. All CNN models have the same architecture and hyperparameters. Similarly, all XGBoost classifiers have the same hyperparameters.

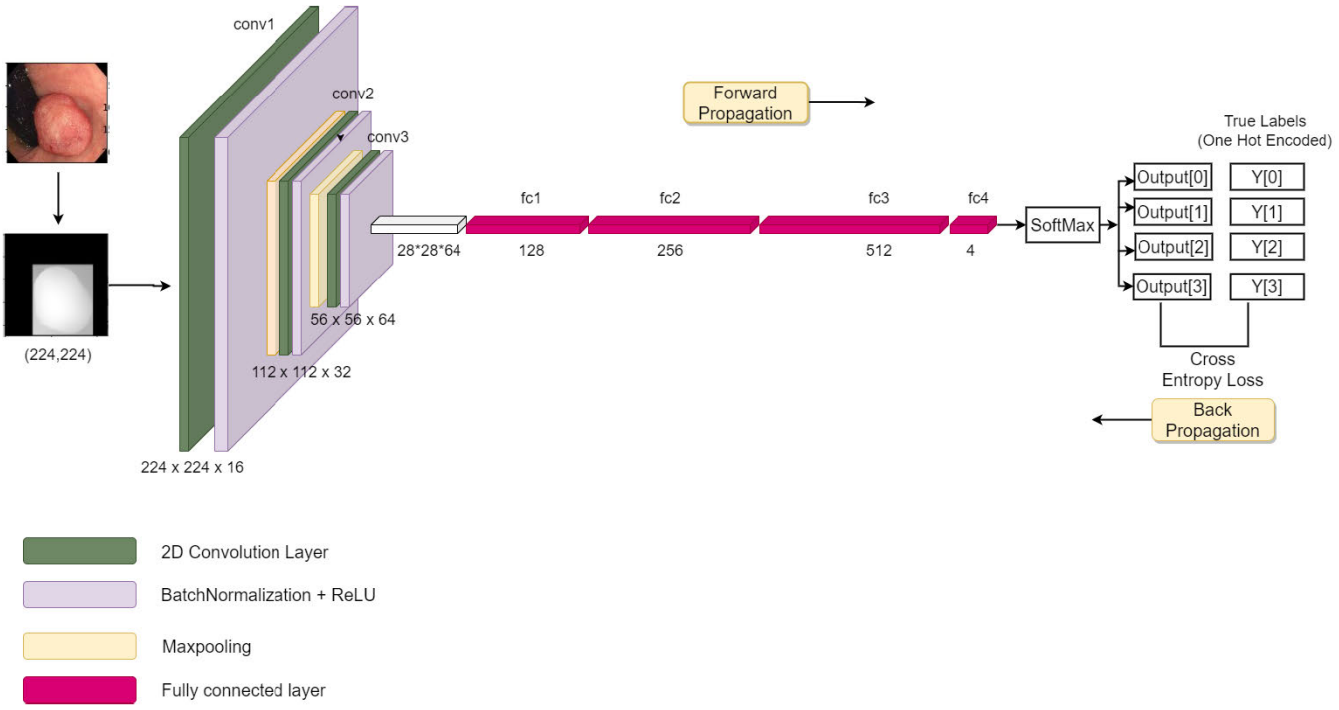


FIGURE 6. Model architecture, using a single CNN. fc1 to fc4 represent the fully connected layers, and output[0] to output[3] and Y[0] to Y[3] denote the predictions and ground truth labels, respectively, for the four classes.

TABLE 1. Binary classifiers and their respective class of focus.

Model	Label=1	Image count	Label=0	Image count	Class
M_0 and XG_0	0-5mm	16139	Rest	32110	0
M_1 and XG_1	5-10mm	17086	Rest	30443	1
M_2 and XG_2	10-14mm	13611	Rest	41745	2
M_3 and XG_3	≥=14mm	7799	Rest	40450	3

The training of the models consists of two parts.

- First, we train the individual binary CNNs against the respective class-wise dataset.

- After the binary CNN classifiers have been trained, we freeze the parameters of the CNN, obtain the output of the last hidden layer of the CNN on the training dataset, and use it to train an XGBoost classifier on the respective class-wise dataset (see Fig. 4). In short, after training the CNN, we drop the very last fully connected layer from the CNN and use the flattened features extracted from the training dataset to train an XGBoost classifier.

This process is repeated for each of the four pairs of models. It is to be noted that for a particular model pair (M_i, XG_i), the goal is to classify the i^{th} class as 1 and all other classes as 0. Here, i=0 implies polyps within 0-5mm (class C0) size, i=1 implies polyps within 5-10mm (Class C1), i=2 implies polyps within 10-14mm (class C2) and i=3 implies ≥14mm (class C3). We have used Binary cross-entropy loss for calculating the difference between network predictions and true labels and Adam optimizer for back-propagating through the CNN. Dropout has been used while training to prevent over-fitting in the CNN [16]. Also, we have used Batch normalization in the convolutional layers.

Hence, we trained each of the CNN (M_i) on the respective class-wise training dataset without using the XGBoost classifier and later used the CNN extracted features to train an XGBoost classifier (XG_i) corresponding to each M_i. This method enables us to combine the powerful feature extraction capabilities of CNN and the classification capabilities of XGBoost classifiers. The output obtained through these models consists of predicted probabilities for each class, the maximum of which is selected to identify the predicted class. This process is depicted in Fig. 5. Using XGBoost classifiers on top of the CNN reported the best accuracy and F1-score among the different experiments performed to evaluate our approach. These experimental setups are explained below:

- **Baseline approach using single CNN model:** Here, we followed the conventional approach of using a single CNN for multi-class classification. The architecture of the model is given in Fig. 6. The number of output nodes is equal to the number of classes, i.e., 4. SoftMax activation was applied on the output nodes to obtain probabilistic values for each class. The target vector was one-hot encoded, and cross-entropy loss was used to compute the difference between predicted values and true values. Adam optimizer was used for weight optimization via back-propagation. Other training hyper-parameters were the same as that in the case of our proposed approach. During inference, a localized depth map of the polyp RGB image was passed to the model and predicted probabilities were obtained, the maximum of which corresponded to the predicted class.

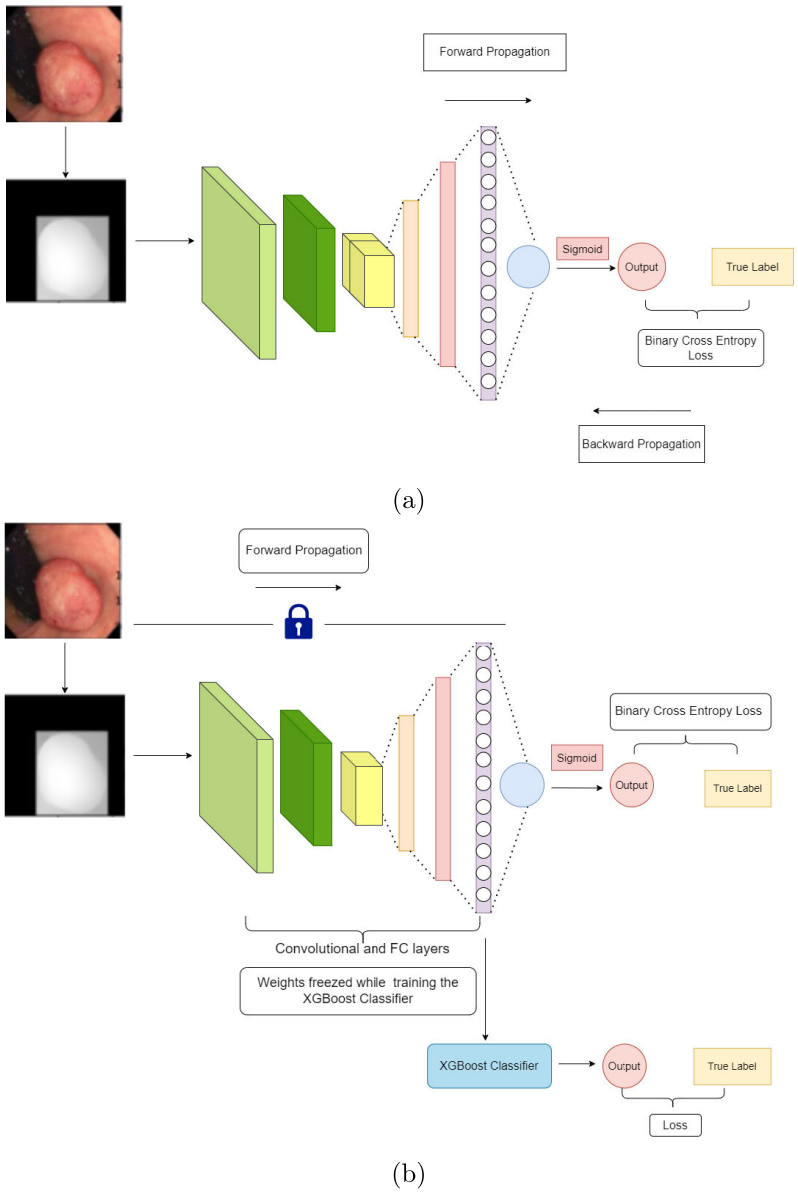


FIGURE 7. Different training phases for ablation study. (a) Without XGBoost classifier, (b) With XGBoost Classifier.

- **Multiple binary CNN model without XGBoost classifier:** We used multiple binary CNN models to validate the significance of XGBoost classifiers in our proposed model. We used the same setup as our proposed model except for the XGBoost classifiers, which are not used in this case. The difference in the two training phases (a) our proposed model (multiple binary CNN models with XGBoost classifiers) and (b) multiple binary CNN models without XGBoost classifiers can be observed in Fig. 7. The figure shows a part of the proposed model, i.e., a single binary classifier M_0 with an XGBoost classifier XG_0. The other three sub-networks of classifiers (M_1 to M_3 and XG_1 to XG_3) have the same architecture.

III. RESULTS

A. DATASET AND TRAINING DETAILS

We used the SUN Database [17], [18] to evaluate our model's performance. It contains 109,554 frames non-polyp frames and 49,136 polyp frames. The training set we used comprises 80% of polyp frames; validation and test sets are 10% each of polyp frames. The distribution of the different classes in the test dataset is shown in Fig. 8. For both training and evaluation purposes, we have used bounding box annotations of polyps already provided in the SUN Database. It is to be noted that for all operations, we have down-sampled the images to (224,224) in order to reduce computational cost. Training hyper-parameters for each CNN have been provided in Table 2. For each XGBoost classifier, we used a maximum

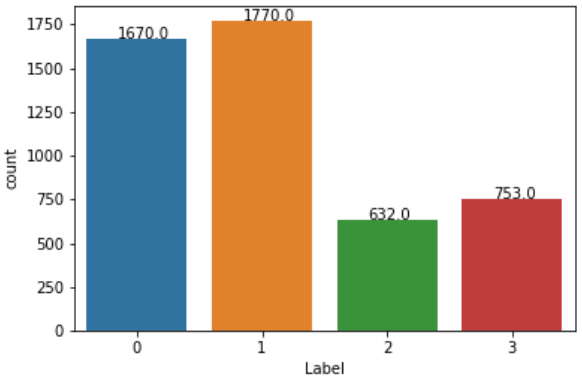


FIGURE 8. Distribution of testing dataset into classes.

TABLE 2. Training hyperparameters for each CNN.

Hyperparameter	Value
Epochs	10
Optimizer	Adam
β_1(for Adam optimizer)	0.9
β_2((for Adam optimizer)	0.99
Learning rate	0.0001
Batch size	64
Loss function (for Binary CNNs)	Binary cross entropy

depth of 6. All model training and inferences have been done in PyTorch. Testing of the models was done on the test dataset comprising randomly selected 4825 images. Accuracy and F1-score were chosen as evaluation metrics.

B. ABLATION STUDY

We performed an ablation study by evaluating our proposed model in two scenarios: (a) with XGBoost classifiers and (b) without XGBoost classifiers. The corresponding training phases for a classifier M_i are shown in Fig. 7. The results obtained in each scenario are reported in Table 3. It can be observed that in almost all cases, the performance of each binary classifier increased in terms of accuracy and F1 score when they were combined with the XGBoost classifiers. XGBoost is a gradient-boosting algorithm that is able to learn non-linear relationships between features and labels applied to tabular data. CNNs, on the other hand, are most efficient in handling two-dimensional data as they effectively learn hierarchical representations of data by applying convolutional filters, which involve extracting local features and patterns from images. CNNs and XGBoost classifiers are combined by feeding the extracted patterns from CNN into the XGBoost classifier. Using CNNs with XGBoost enables the model to effectively capture intricate patterns and relationships in data. This ability is leveraged by integrating CNN as a trainable feature extractor to automatically obtain features from input data and XGBoost as a recognizer in the network's top level to obtain results. These facts and observations illustrate the possible reasons for the increment in polyp-size classification accuracy and F1-score by combining CNNs and XGBoost classifiers as presented in our work.

C. COMPARATIVE ANALYSIS

We compared our model's performance with (a) a baseline multi-class classification approach using the single CNN model, (b) multiple binary CNN models without XGBoost classifiers, and (c) well-known image classification models: ResNet-101, DenseNet-169, and Vision Transformer (VIT). The associated results regarding (a) and (b) are shown in Table 4. For multi-class classification on the test dataset, a single CNN provided an accuracy of 84.87% and an F1-score of 84.47%; binary CNNs combined via One vs. Rest classification provided an accuracy of 86.48% and an F1-score of 86.87%; binary CNNs and XGBoost classifiers combined via One vs. Rest classification provided an accuracy of 87.07% and an f1-score of 86.95%. This shows that our proposed model outperformed the baseline model by a significant margin and achieved better accuracy and F1-score using XGBoost classifiers. The qualitative analyses of these models are shown in Fig. 9, Fig. 10, Fig. 11, and Fig. 12. These figures mention the ground truth class and the predicted class labels for some randomly chosen samples. Fig. 12 presents some polyp images incorrectly classified by the baseline model, whereas our proposed model correctly predicted the labels for these samples. The effectiveness of using One vs. Rest classification using multiple binary classifiers can be attributed to the following facts:

(i) Simplified Decision Boundaries: In a single CNN classifier for multi-class classification, the decision boundaries can be complex and intertwined, making it challenging to separate different classes. By using One vs. Rest classification with binary CNNs, each classifier is trained to differentiate one class from the rest, resulting in simpler decision boundaries for each binary problem. (ii) Targeted Feature Learning: Each binary CNN classifier in the One vs. Rest classification focuses on distinguishing one class, enabling it to learn features specific to that class. This targeted feature learning can enhance the discriminative power of the classifiers. In contrast, a single CNN for multi-class classification learns features that need to be shared across multiple classes, which may result in less specialization for individual classes. (iii) Model Diversity: One vs. Rest classification with binary CNNs creates an ensemble of classifiers, where each classifier specializes in distinguishing a particular class. This ensemble approach harnesses the diversity of the classifiers, allowing them to collectively capture a broader range of class characteristics and improve overall performance, enhancing robustness and generalization capabilities.

In addition to the above results, we have included confusion matrices in Fig.13 for all the mentioned approaches. We also compare our method against three well-known image classification models: ResNet-101 [19], DenseNet-169 [20] and Vision Transformer (VIT) [21]. We used transfer learning in the case of both ResNet-101 and DenseNet-169. The associated results are presented in Table 5. It can be observed that our approach outperforms other well-known image classification models by a significant margin in terms

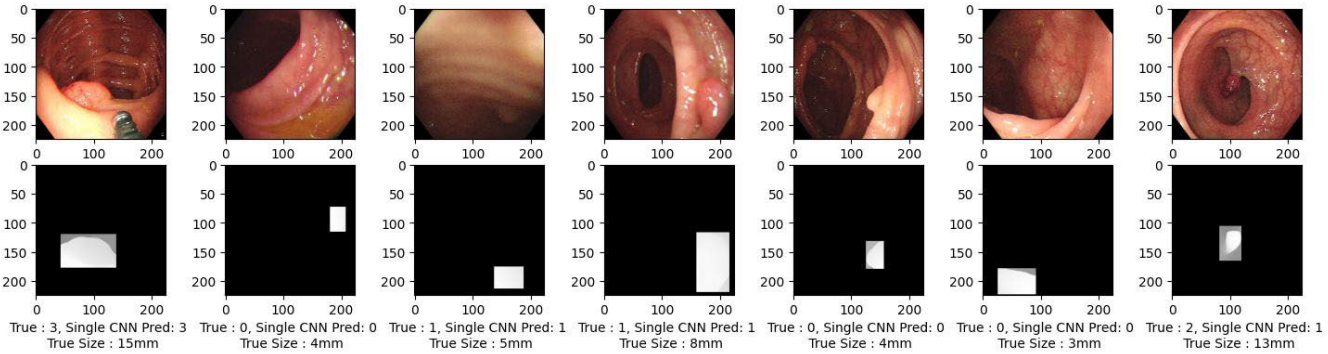


FIGURE 9. Some RGB Polyp images with respective localized depth maps and corresponding predictions on the test dataset using a single CNN, true label, and actual size in mm. Note that class=0 implies (0-5) mm,class=1 implies [5-10) mm,class=2 implies [10-14) mm, and class=3 implies greater than or equal to 14 mm.

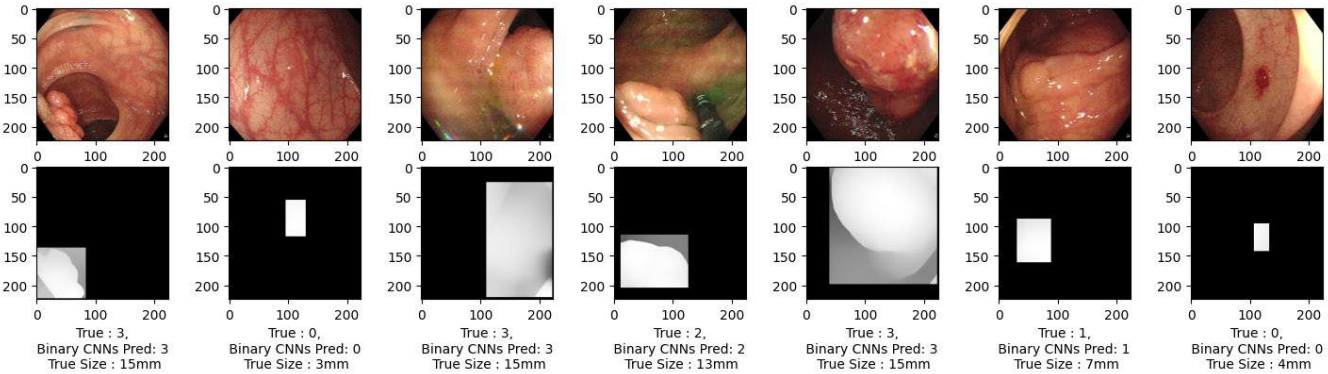


FIGURE 10. Some RGB Polyp images with respective localized depth maps and the corresponding predictions obtained via One vs. Rest classification using binary classifying CNNs without XGBoost classifiers on the test dataset, true label, and actual size in mm. Note that class=0 implies (0-5) mm,class=1 implies [5-10) mm,class=2 implies [10-14) mm, and class=3 implies greater than or equal to 14 mm.

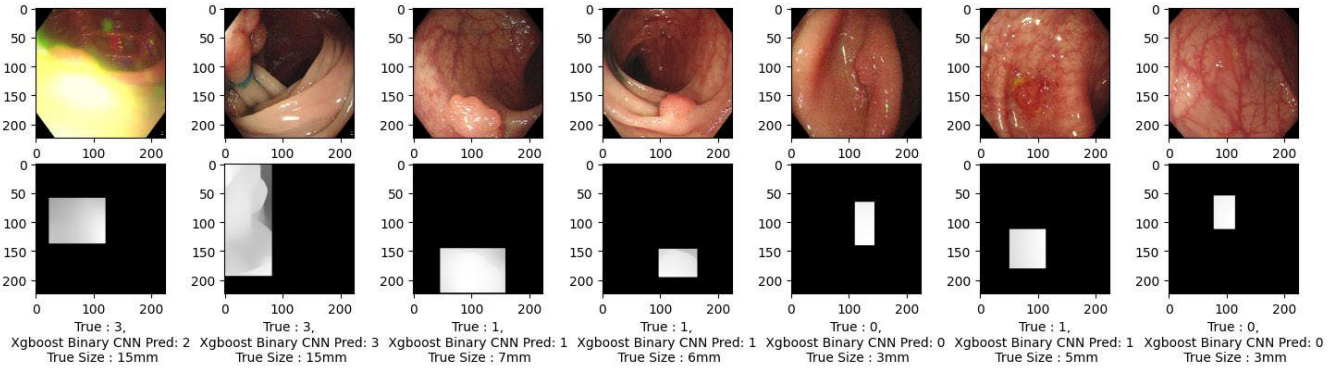


FIGURE 11. Some randomly picked-up images from the test dataset with the true label, predictions from Binary CNNs combined with XGBoost classifiers, with the true size of the corresponding polyp in mm, respectively. Note that class=0 implies (0-5) mm,class=1 implies [5-10) mm,class=2 implies [10-14) mm, and class=3 implies greater than or equal to 14 mm.

TABLE 3. Ablation study on Binary CNNs without and with XGBoost classifiers for Binary Classification. The underlined values represent the best result in the respective comparisons.

	Accuracy		F1-score	
Model	Without XGBoost	With XGBoost	Without XGBoost	With XGBoost
M_0	90.17%	<u>90.88%</u>	<u>88.47%</u>	86.60 %
M_1	88.04%	<u>89.06%</u>	84.40%	<u>84.96 %</u>
M_2	92.39%	<u>94.40%</u>	74.53%	<u>78.22 %</u>
M_3	97.53%	<u>97.7%</u>	92.09%	<u>95.58 %</u>

TABLE 4. Comparative analysis with different approaches. The underlined values represent the best result.

Method	Accuracy	F1-score
Baseline model (Single CNN model)	84.87%	84.47%
Multiple binary CNNs without XGBoost	86.48%	86.87%
Proposed model (Multiple binary CNNs with XGBoost)	87.07%	86.95%

TABLE 5. Comparative analysis with state-of-the-art image classification models. The underlined values represent the best result.

Model	Accuracy	F1-score
DenseNet-169 (Transfer Learning)	69.13%	66.32%
Vision Transformer(VIT)	73.82%	70.78%
ResNet-101(Transfer Learning)	69.49%	65.39%
Single CNN(as described in Figure3)	86.48%	86.87%
Proposed model (Multiple binary CNNs with XGBoost (Figure2))	87.07%	86.95%

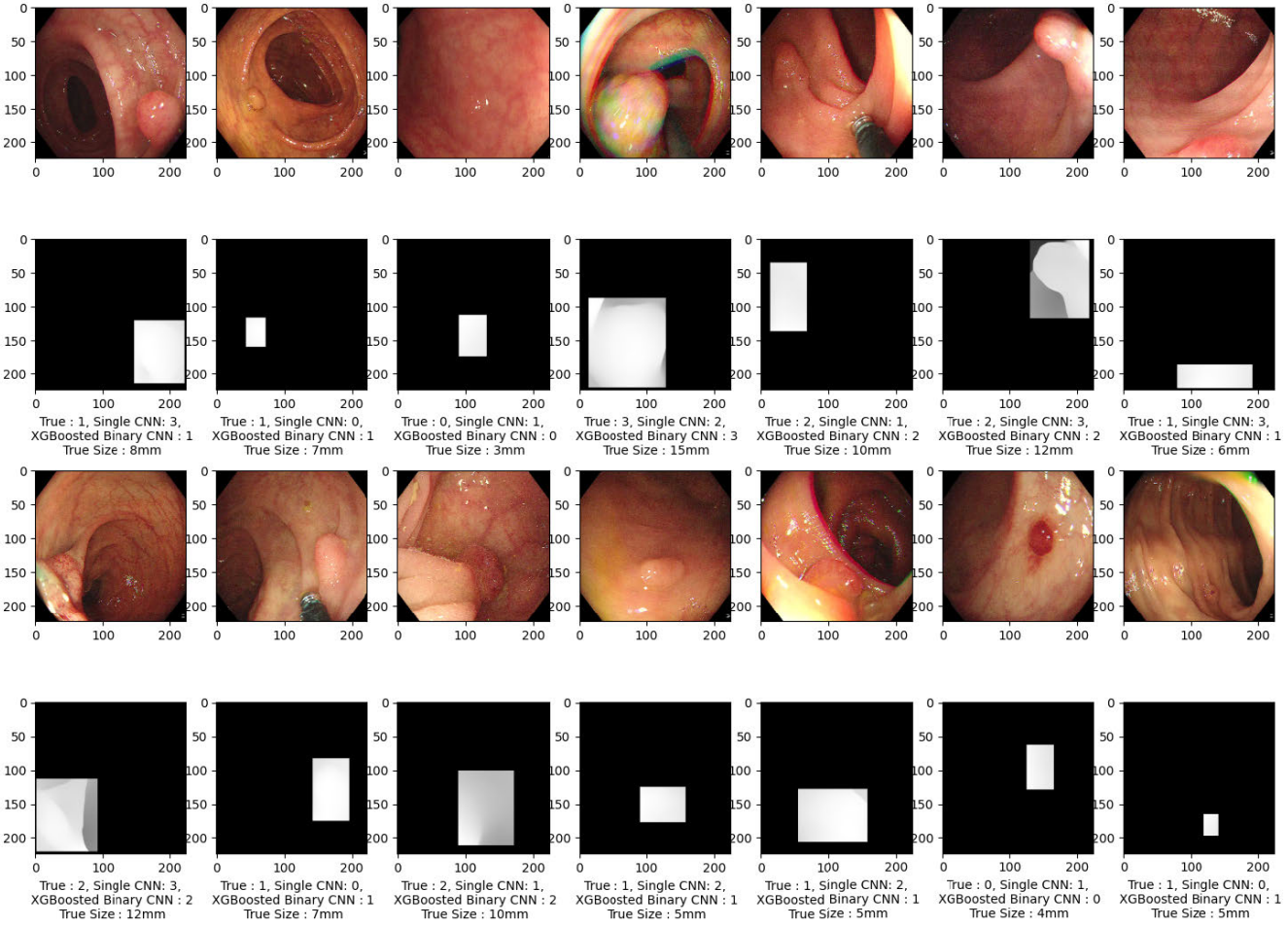
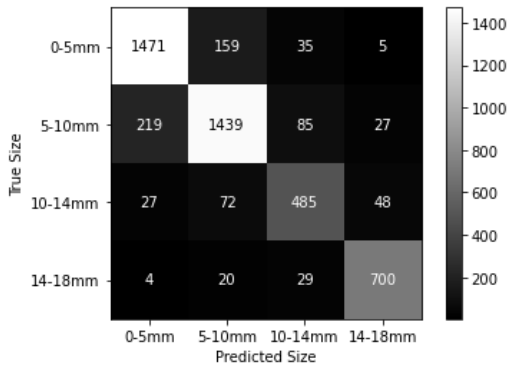
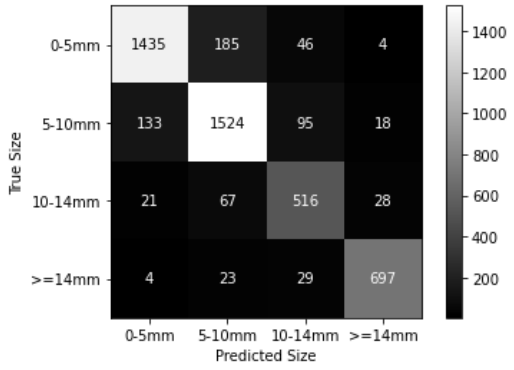


FIGURE 12. Some polyp images whose localized depth maps were incorrectly classified by the single CNN but correctly classified by the combined binary classifiers with XGBoost. The true class, class predicted by single CNN, and class predicted by Binary CNNs with XGBoost, along with the true size of the polyps in mm, have been provided for each image. Note that class=0 implies (0-5)mm, class=1 implies [5-10)mm, class=2 implies [10-14)mm, and class=3 implies greater than or equal to 14mm.

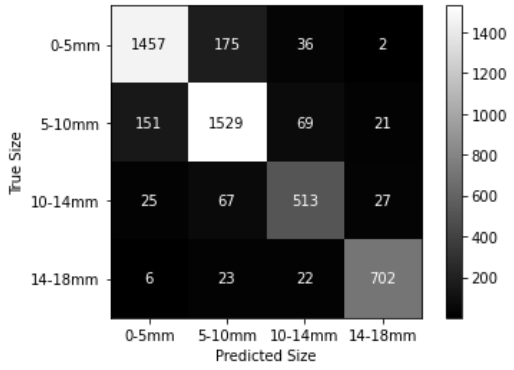
of both F1-score and accuracy for polyp size classification. One possible explanation for these results is the sparse nature of input features (images) that we have used in our work. Localized polyp depth maps are mostly sparse as most part of the image is without polyp area, which is all black or all zeros. Both ResNet-101 and DenseNet-169 are very deep networks with many layers, whereas our proposed model has fewer layers (3 convolutional layers and 3 fully connected layers). From the results tabulated below, it is clear that shallow CNNs might be more effective than very deep CNNs for sparse image classification. Also, the results in Table 5 demonstrate that shallow CNNs might outperform Vision



(a)



(b)



(c)

FIGURE 13. Confusion matrix for (a) Single CNN model (b) Multiple binary CNN model without XGBoost classifier, (b) Proposed approach.

Transformers (VIT) for image classification when the input image is sparse, as in the case of localized depth maps of polyps. Although the specified classification models are well-known for their superior classification performance, for this particular objective, our model, being a shallow model, performs better.

IV. DISCUSSION

We evaluated the models on the test dataset comprising 4825 images (randomly sampled from the dataset) using three approaches: a single CNN, multiple binary CNNs via One vs. Rest classification, and multiple Binary CNNs (total of 4 in number) with XGBoost Classifiers (one for each binary CNN) via the One vs. Rest classification technique. The obtained results report an increase in test accuracy by 1.61% and F1-score by 2.4% using multiple binary classifying CNNs as compared to using a single CNN. Also, combining the binary CNNs with XGBoost classifiers results in an increase of test accuracy by 2.2% (106 images) and an F1-score of 2.48% compared to using a single CNN.

The current approach mainly aims at the size estimation of a given polyp; however, in the future, we might use an object detection model for polyp localization as an initial step. This includes adopting real-time object detection models like YOLO [22], SSD [23] etc., followed by the multi-class classification of polyps based on their absolute size. Nevertheless, our goal of polyp estimation and its clinical significance is evident from the exhaustive experimental analysis.

V. CONCLUSION

In this paper, we have proposed a deep learning based approach for multi-class classification of colorectal polyps according to their absolute size (in mm). The four classes included for classification are (0-5) mm, [5-10) mm, [10-14) mm, and greater than or equal to 14 mm. Our proposed model consists of four binary classifying CNNs along with XGBoost classifiers. The individual binary classifiers are combined via the One vs. Rest classification technique. Our proposed model outperformed the baseline model by 2.2% and 2.48% in terms of accuracy and F1-score, respectively. Instead of a binary classification of polyp size, which has been done in most previous works, a muti-class classification enables a much more accurate range of polyp size, which might be beneficial for surgical procedures and related other medical treatments. Though four different binary classifiers are more computationally expensive, better accuracy is always desired when it comes to AI-assisted endoscopy, as it will have a significant impact on the treatment.

ACKNOWLEDGMENT

The authors would like to thank Iwahori Laboratory members for their useful discussions.

REFERENCES

[1] P. Rawla, T. Sunkara, and A. Barsouk, "Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors," *Gastroenterol. Rev./Przegląd Gastroenterologiczny*, vol. 14, no. 2, pp. 89–103, 2019.

[2] T. Sawicki, M. Ruszkowska, A. Danielewicz, E. Niedźwiedzka, T. Arłukowicz, and K. E. Przybyłowicz, "A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis," *Cancers*, vol. 13, no. 9, p. 2025, Apr. 2021.

[3] S. Tanwar, S. Vijayalakshmi, M. Sabharwal, M. Kaur, A. A. AlZubi, and H.-N. Lee, "Detection and classification of colorectal polyp using deep learning," *BioMed Res. Int.*, vol. 2022, pp. 1–9, Apr. 2022.

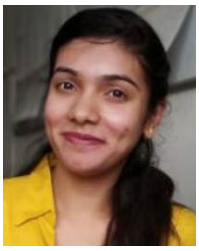
[4] F. Younas, M. Usman, and W. Q. Yan, "A deep ensemble learning method for colorectal polyp classification with optimized network parameters," *Int. J. Speech Technol.*, vol. 53, no. 2, pp. 2410–2433, Jan. 2023.

[5] V. Sharma, P. Sasmal, M. K. Bhuyan, P. K. Das, Y. Iwahori, and K. Kasugai, "A multi-scale attention framework for automated polyp localization and keyframe extraction from colonoscopy videos," *IEEE Trans. Autom. Sci. Eng.*, early access, Oct. 2, 2023, doi: 10.1109/TASE.2023.3315518.

[6] P. Sasmal, M. K. Bhuyan, S. Dutta, and Y. Iwahori, "An unsupervised approach of colonic polyp segmentation using adaptive Markov random fields," *Pattern Recognit. Lett.*, vol. 154, pp. 7–15, Feb. 2022.

[7] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 225–2255.
[8] V. Sharma, P. Sasmal, M. K. Bhuyan, and P. K. Das, "Keyframe selection from colonoscopy videos to enhance visualization for polyp detection," in *Proc. 26th Int. Conf. Inf. Vis. (IV)*, Jul. 2022, pp. 426–431.
[9] F. Mohammad, V. Sharma, and P. K. Das, "Polyp detection in colonoscopy images using improved deformable DETR," in *Proc. IEEE Region Conf. (TENCON)*, Nov. 2022, pp. 1–6.
[10] H. Itoh, M. Oda, K. Jiang, Y. Mori, M. Misawa, S.-E. Kudo, K. Imai, S. Ito, K. Hotta, and K. Mori, "Binary polyp-size classification based on deep-learned spatial information," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 10, pp. 1817–1828, Oct. 2021.
[11] H. Itoh, H. R. Roth, L. Lu, M. Oda, M. Misawa, Y. Mori, S.-E. Kudo, and K. Mori, "Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning," in *Proc. 21st Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Granada, Spain: Springer, Sep. 2018, pp. 611–619.
[12] M. Abdelrahim, H. Saiga, N. Maeda, E. Hossain, H. Ikeda, and P. Bhandari, "Automated sizing of colorectal polyps using computer vision," *Gut*, vol. 71, no. 1, pp. 7–9, Jan. 2022.
[13] F. Chadebecq, C. Tilmant, and A. Bartoli, "Using the infocus-breakpoint to estimate the scale of neoplasia in colonoscopy," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, Apr. 2013, pp. 354–357.
[14] B. Villard, Y. Mori, M. Misawa, S.-E. Kudo, H. Itoh, M. Oda, and K. Mori, "Colorectal polyp size classification using a Siamese network," MIDL Abstract, London, U.K., Tech. Paper 123, 2019.
[15] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.
[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.
[17] M. Misawa, S.-E. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida, H. Itoh, M. Oda, and K. Mori, "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)," *Gastrointestinal Endoscopy*, vol. 93, no. 4, pp. 960.e3–967.e3, Apr. 2021.
[18] H. Itoh, M. Misawa, Y. Mori, M. Oda, S. E. Kudo, and K. Mori, "Sun colonoscopy video database," 2020. [Online]. Available: http://amed8k.sundatabase.org/
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Springer, Oct. 2016, pp. 21–37.

PROMIT HALDAR was born in Berhampore, West Bengal, India. He is currently pursuing the B.Tech. degree with the Department of Electronics and Communication Engineering, IIT Guwahati. He was a Deep Learning and Computer Vision Research Intern with Chubu University, Japan. He was also a Research Intern with Adobe Research, Bengaluru, India. His research interests include deep learning for biomedical imaging, computer vision, large language models, data storytelling, and deep learning.



VANSHALI SHARMA received the B.Tech. degree from Maharshi Dayanand University, Rohtak, India, in 2015, and the M.Tech. degree from IIT (ISM) Dhanbad, India, in 2018. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Indian Institute of Technology (IIT) Guwahati, India. She received the INSPIRE Fellowship from the Department of Science and Technology (DST), Government of India. Her research interests include medical image/video analysis, computer vision, and deep learning.

YUJI IWAHORI (Member, IEEE) received the B.S. degree from the Nagoya Institute of Technology, in 1983, and the M.S. and Ph.D. degrees from the Department of Electrical and Electronics, Tokyo Institute of Technology, in 1985 and 1988, respectively. In 1988, he joined the Educational Centre for Information Processing, Nagoya Institute of Technology, as a Research Associate, and became a Professor of the Centre for Information and Media Studies, in 2002. He joined Chubu University, as a Professor, in 2004, and acted as the Department Head of computer science, the Head of graduate course of computer science, and the Vice-Dean of the College of Engineering. He has been a Visiting Researcher with UBC Computer Science, since 1991. He has also been a Research Collaborator with IIT Guwahati, since 2010, and with the Department of Computer Engineering, Chulalongkorn University, since 2014. He has become an Honorary Faculty with IIT Guwahati, in 2020. His research interests include computer vision, biomedical image processing, deep learning, and the application of artificial intelligence. He received the KES 2008 Best Paper Award and the KES 2013 Best Paper Award from KES International.

M. K. BHUYAN (Senior Member, IEEE) received the Ph.D. degree in electronics and communication engineering from the Indian Institute of Technology (IIT) Guwahati, India. He was with the School of Information Technology and Electrical Engineering, University of Queensland, St. Lucia, QLD, Australia, as a Postdoctoral Researcher. He was an Assistant Professor with the Department of Electrical Engineering, IIT Roorkee, India, and the Jorhat Engineering College, Assam, India, and he also with Indian Engineering Services. In 2014, he was a Visiting Professor with Indiana University and Purdue University, IN, USA. He is currently a Professor with the Department of Electronics and Electrical Engineering, IIT Guwahati, where he is also the Dean of infrastructure, planning, and management. He is also a Visiting Professor with the Department of Computer Science, Chubu University, Japan. His current research interests include machine learning and artificial intelligence, image/video processing, computer vision, human–computer interactions (HCI), virtual reality and augmented reality, and biomedical signal processing. He was a recipient of the National Award for Best Applied Research/Technological Innovation by the President of India, in 2012.

AILI WANG (Member, IEEE) was born in Tianjin, China, in 1979. She received the B.S., M.S., and Ph.D. degrees in information and signal processing from the Harbin Institute of Technology, Harbin, China, in 2002, 2004, and 2008, respectively. She joined the Harbin University of Science and Technology, as an Assistant, in 2004. She was an Associate Professor and the Master's Tutor with the Department of Communication Engineering, in 2010. She has been a Visiting Professor of the research of 3-D polyp reconstruction with the Computer Science Laboratory, Chubu University, Japan, in 2014. She is the author of two books and more than 80 articles, which were published on the IEEE conferences and journals (EI-indexed or SCI-indexed). Her research interests include image super resolution, image fusion, and object tracking. She is a Committee Member of the 11th EAI International on Wireless and Satellites (WISATS) and the Seventh EAI International Conference on Green Energy and Networking (GreeNets).

HAIBIN WU was born in Harbin, China, in 1977. He received the B.S. and M.S. degrees from the Harbin Institute of Technology, Harbin, in 2000 and 2002, respectively, and the Ph.D. degree in measuring and testing technologies and instruments from the Harbin University of Science and Technology, Harbin, in 2008. From 2009 to 2012, he was a Postdoctoral Researcher with the Key Laboratory of Underwater Robot, Harbin Engineering University. From 2014 to 2015, he was a Visiting Scholar with the Robot Perception and Action Laboratory, University of South Florida. Since 2012, he has been a Professor with the Instrument Science and Technology Discipline, Harbin University of Science and Technology. He is the author of three books, more than 40 articles, and more than 20 inventions. His research interests include robotic vision, visual measuring, image processing, medical virtual reality, and photoelectric testing. He was the Director of the Precision Machinery Branch, China Instrumentation Society; and the Visual Inspection Committee, Chinese Graphic and Image Society. He serves as an Editorial Board Member for *Chinese Journal of Liquid Crystals and Displays* and the Associate Editor-in-Chief for *Journal of Harbin University of Science and Technology*.

KUNIO KASUGAI received the M.D. and Ph.D. degrees in bioregulation research from the Medical School, Nagoya City University, Nagoya, Japan. He was a Research Fellow of internal medicine with the University of Michigan. He is currently a Professor of internal medicine with the Division of Gastroenterology and the Vice President with the School of Medicine, Aichi Medical University. He is also the Executive Vice President and the Director of the Endoscopy Center, Aichi Medical University Hospital. He holds the Society Membership of the Japanese Society of Internal Medicine, the Japanese Society of Gastroenterology, the Japan Gastroenterological Endoscopy Society, and the American Gastroenterological Society.