TIPTEKNO'24 MEDICAL TECHNOLOGIES CONGRESS
10-12 October 2024
Yasmin Bodrum Resort, Muğla, Türkiye

Gastrointestinal Image Classification Using Deep Learning Architectures via Transfer Learning

İlker Korkmaz
Department of Computer Engineering
Izmir University of Economics, Izmir, Turkey
ilker.korkmaz@ieu.edu.tr

Fatih Soygazi
Department of Computer Engineering
Adnan Menderes University, Aydin, Turkey
fatih.soygazi@adu.edu.tr

***Abstract*—Computer aided detection of diseases using machine learning mechanisms on medical images has been an interesting applied research topic in both academia and health sector. Practical studies with the aim of improving the process of decision on the diagnosis of the diseases via accurate classification of the medical images would be benefit of the medical doctors. This paper presents an investigation on the classification of gastrointestinal images using deep learning models. The labeled medical images used in the experiments are publicly available within the Kvasir dataset on Kaggle. The deep learning approaches applied through the experiments are based on the following Convolutional Neural Network architectures used with transfer learning: VGG19, ResNet50V2, ResNet152V2, EfficientNetV2B0, EfficientNetV2B3, InceptionV3, DenseNet201, Xception. The performances of these different architectures on learning the training dataset and classifying the test images are evaluated in terms of the following metrics: accuracy, precision, recall, and F1-score. Regarding the results of the experiments conducted using the same dataset on different deep learning models, VGG19 model outperformed the others with the prediction accuracy ratio of 88.6%.**

***Keywords*—machine learning, deep learning architectures, transfer learning, image classification, gastrointestinal disease detection.**

I. Introduction

As practical artificial intelligence (AI)-supported computer programs, particularly machine learning (ML)-based software applications, become increasingly prevalent across various sectors, empirical studies to assess and improve the performance of different ML architectures in diverse application areas are crucial. In this context, particularly for ML-based decision support systems in medical diagnosis, deep learning methodologies have shown considerable promise for classifying medical images.

In the medical diagnosis of diseases affecting either an organ or the entire system of the human body, the relevant medical images of the focused segment must be analyzed by medical professionals. To assist doctors in making accurate diagnoses, the integration of ML-based applications into the diagnostic process is practical, useful, and time-saving. For an effective implementation of an ML-based image classification and diagnostic detection system, though the design of the related processes might differ for the methodologies, the following three essential requirements should be met in general: i. a well-prepared and annotated medical image dataset; ii. appropriate extraction of meaningful features from the images; and iii. a high-performing machine learning model (for both training and testing phases). Regarding the first requirement, the image dataset should either be analyzed and tagged from scratch by experts, specifically medical doctors in the relevant field, or a reliable publicly available image dataset may be utilized. For the second requirement, image processing algorithms and tools should be employed to extract the relevant features. To address the third requirement, various machine learning architectures should be evaluated, either by implementing them independently or by adapting models through transfer learning techniques. Transfer learning is a method aimed at enhancing learning performance by using a model pre-trained on images from a different domain to recognize images in the target domain. For a comprehensive overview of transfer learning techniques, readers may refer to the survey conducted by Zhuang et al. [1].

This study focuses on the classification of the gastrointestinal images to be used for the detection of the corresponding medical diseases. The image dataset, i.e., the second version of Kvasir [2], used in this work is publicly available on [3]. The first version of Kvasir dataset contains 4000 images in 8 different classes showing anatomical landmarks, phatological findinds or endoscopic procedures in the gastrointestinal tract. The updated second version of Kvasir dataset includes 8000 images; 1000 images per each class category. This paper deals with neither dataset preparation nor data labeling since the dataset is available publicly and prepared by trustworthy experts. Instead, the work aims to evaluate and discuss the performance of various Convolutional Neural Network (CNN) architectures tested on the available Kvasir dataset. In this manner, this paper presents an investigation on the classification of gastrointestinal images using deep learning approaches.

The rest of the paper is organized in the following manner: Section II surveys some related works on the classification of medical images; Section III explains the materials as well as the methodology used in the experiments conducted, shows and evaluates the empirical results; Section IV concludes the paper.

979-8-3315-2981-9/24 $31.00 © 2024 IEEE

II. Related Work

This section basically gives a background on the classification of the medical images and particularly surveys the literature on the classification of gastrointestinal images using deep learning approaches.

Kumar et al. [4] presented a survey on the medical images classification using deep learning approaches. The article reviews some of the models (i.e., CNN, transfer learning, long short term memory, generative adversarial networks, autoencoders), summarizes well-known deep learning architectures used in medical image classification, describes the metrics commonly used for classification of the images, and lists some of the medical image datasets.

Malik et al. [5] presented their study using four different multi-classification deep learning models, particularly VGG19 + CNN, ResNet152V2, Gated Recurrent Unit (GRU) + ResNet152V2, and ResNet152V2 + Bidirectional GRU, to classify the three disorders of the gastrointestinal tract (i.e., ulcerative colitis, polyps, and dyed-lifted polyps) from wireless endoscopy images. The authors evaluated the experimental results of the four models in terms of accuracy, loss, Matthew's correlation coefficient, recall, precision, negative predicted value, positive predicted value and F1-score. Based on the comparison of the evaluation results of the four models, the Vgg19 + CNN showed the best accuracy performance, which is 99.45%.

Siddiqui et al. [6] proposed a CNN based framework for classifying gastrointestinal images. The authors showed the accuracy results of the proposed model with different epochs and batch sizes on two versions of Kvasir dataset. The authors also compared their model with existing models after applying feature selection. The accuracy of the model has reported 99.3% on the first version of Kvasir and 97.6% on the second version of Kvasir datasets.

Uçan et al. [7] presented their findings on the classification of gastrointestinal images using EfficientNetB0 model. The authors used the images of the second version of Kvasir dataset, 80% for training, 10% for validation and 10% for testing the model. The batch size and number of epochs used for the training were both chosen as 20. Regarding the classification results, the authors achieved an accuracy of 93.5%.

Ramamurthy et al. [8] proposed a feature fusion network based on EfficientNetB0 to classify the endoscopy images. The authors basically combined EfficientNetB0 with their custom-built feature extractor (i.e., Effimix). Regarding the experiments, the authors aimed to classify 23 classes of images publicly available within HyperKvasir dataset [9]. Based on the analysis of the proposed network mentioned in the paper, the model achieved an accuracy of 97.99%.

Altıntaş and Öziç [10] showed the performance results of their experiments conducted using AlexNet, ResNet50, GoogleNet, InceptionV3, ShuffleNet, and SqueezeNet models through transfer learning approach. To collect a balanced dataset, 300 ischemic, 300 hemorrhage and 300 normal computed tomography images were taken from open databases. Five-fold cross-validation strategy was used for splitting the data for training (80%) and testing (20%). Based on the experiment results, AlexNet has been mentioned as the best with an accuracy of 90.89%.

As another related work on the classification of the endoscopy images, Khan et al. [11] proposed a framework containing the following three components: i. Local-Global Convolutional Neural Network, ii. Endoscopy-Lesion Attention module, and iii. Gastrointestinal-Endoscopy Convolutional Neural Network. Within the last component, ResNet is used as the backbone architecture for the classification of the images. The authors conducted the experiments on the second version of Kvasir and also HyperKvasir datasets. Regarding the evaluation of the experiments, the proposed method achieved an accuracy of 98.50% on Kvasir dataset and 88.72% on HyperKvasir dataset.

III. Methodology

This section explains the methodology employed on the dataset preparation, learning and testing stages for the entire classification process. The experiment results are shown and discussed in this section.

A. Dataset

The Kvasir dataset [2] is a publicly available collection of medical images, involving eight image categories (Table I), specifically designed for the study and development of machine learning and deep learning algorithms in the field of gastrointestinal endoscopy. It is primarily used to advance research in automated image analysis for diagnosing and classifying gastrointestinal disorders. The dataset contains images captured from gastrointestinal endoscopic procedures, including both colonoscopy and gastroscopy. It covers a wide range of conditions and abnormalities such as polyps, ulcers, and other lesions. The images in the Kvasir dataset are typically annotated with labels that indicate various gastrointestinal conditions or abnormalities, as they provide ground truth for supervised learning tasks. The Kvasir dataset is available in two versions, containing 4,000 and 8,000 annotated images, respectively. We have used the extended version of the dataset in this study where the increased size of the dataset yields better results.

TABLE I: Distribution Of Gastrointestinal Image Classes in KvasirV2 dataset

Class No	Category of Images	Number of Images
1	dyed-lifted-polyps	1000
2	dyed-resection-margins	1000
3	esophagitis	1000
4	normal-cecum	1000
5	normal-pylorus	1000
6	normal-z-line	1000
7	polyps	1000
8	ulcerative-colitis	1000

Table I shows the distribution of the image classes. The dataset encompasses three categories of information related to gastrointestinal images: anatomical landmarks, pathological

findings, and polyp removal procedures. Anatomical landmarks include features such as the ”Z-line”, ”pylorus”, and ”cecum”, while pathological findings cover conditions such as ”esophagitis”, ”polyps”, and ”ulcerative colitis”. Additionally, the dataset includes images depicting various stages of lesion removal, such as ”dyed and lifted polyps” and ”dyed resection margins”.

A comprehensive endoscopic report should include both brief descriptions and visual documentation of key anatomical landmarks. Polyps found in the large bowel can be precursors to cancer and are often removed during endoscopy. Accurate computer detection of dyed polyps and resection sites is critical for developing future computer-aided reporting systems.

B. Deep Learning Architectures Used for Transfer Learning

CNN, a well-known deep learning architecture, is the preferred architecture for image classification due to their ability to effectively extract and learn hierarchical features from images. CNNs capture spatial relationships between pixels, enabling them to recognize complex patterns within the images. Pretrained CNN architectures are often utilized across various image classification tasks. Transfer learning approaches typically involve using a pre-trained CNN as a feature extractor, followed by the addition of the custom layers tailored to the specific classification task.

In this paper, we applied a transfer learning approach using VGG, ResNet, EfficientNet, Inception, DenseNet and Xception neural network architectures to achieve optimal performance. Keras has been used as the frontend library, interfacing with Tensorflow. The numerical suffixes in the names of these architectures, such as DenseNet201, indicate the number of layers within the network. For instance, DenseNet201 refers to a CNN with 201 layers. The term ”deep” thus refers to the use of multiple layers within the network.

C. Experimental Setup

We prepared the dataset by dividing it to training, validation and test sets. KvasirV2 dataset includes images with resolutions of 720*576 pixels and 1920*1072 pixels. We downsampled each image to a resolution of 100*100 pixels to focus on the relevant features for the gastrointestinal images. We used metrics commonly used in image classification, including accuracy, precision, recall, and F1-score. The project was executed on Google Colab platform using a T4 GPU.

D. Results

The image classification results were obtained for 8 classes in terms of accuracy, precision, recall and F1-score metrics. The evaluation strategy focused on comparing the accuracy values of each model to identify the model with the highest accuracy. The precision, recall, and F1-scores of the best-performed model were then comprehensively analyzed. Additionally, the confusion matrix for this model was provided to illustrate its performance in classifying each individual class.

The accuracy values are presented in Table II. As highlighted in bold, VGG19 model achieved the highest accuracy (i.e., 0.886). The models in Table II are chosen based on CNN oriented transfer learning approaches listed in [12].

TABLE II: COMPARISON OF CLASSIFICATION RESULTS

Model	Accuracy
VGG19	**0.886**
ResNet50V2	0.841
ResNet152V2	0.861
EfficientNetV2B0	0.863
EfficientNetV2B3	0.872
InceptionV3	0.811
DenseNet201	0.872
Xception	0.842

Fig. 1 illustrates the block diagram of our best performing model where extra layers were integrated into VGG19 architecture. As the backbone, VGG19 was used for feature extraction, whereas the extra layers were used to enhance the efficiency of image classification.

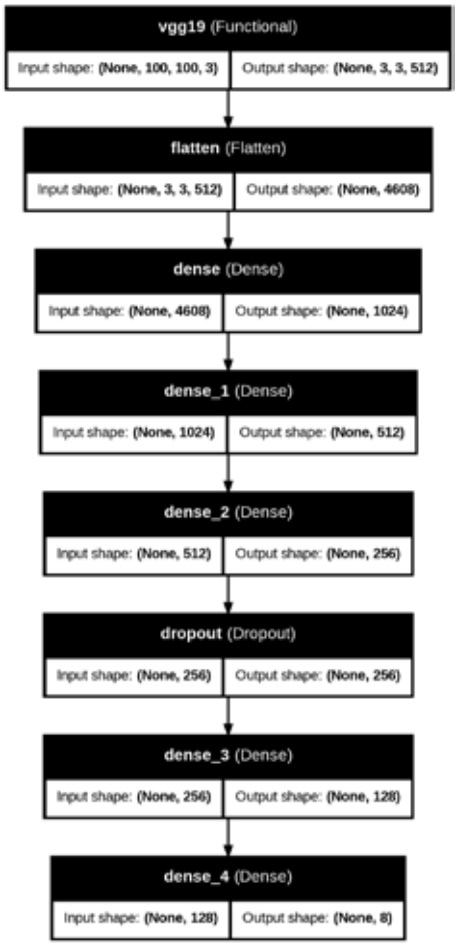


Fig. 1: Architecture of VGG19 model.

The hyperparameters of VGG19 model that performed best in terms of accuracy are shown in Table III. The confusion matrix of VGG19 model expressing the results of the tests using 200 images for each of 8 classes is shown in Fig. 2.

TABLE III: HYPERPARAMETERS FOR VGG19 MODEL

Hyperparameter	Value
Batch size	100
Epoch	20
Initial learning rate	0.001
Minimum learning rate	0.00001
Optimizer	SGD
Weight	Imagenet

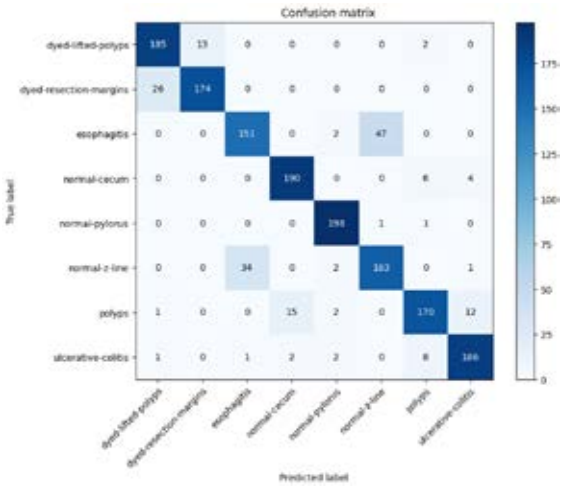


Fig. 2: Confusion matrix of VGG19 model.

It is crucial to evaluate the performance through all classes, not for a single class only. Therefore, the results of the experiments with VGG19 model are listed in Table IV in terms of the performance metrics for each class.

TABLE IV: CLASS-BASED RESULTS OF VGG19 MODEL

Class	Precision	Recall	F1-score	Accuracy
dyed-lifted-polyps	0.87	0.93	0.90	
dyed-resection-margins	0.93	0.87	0.90	
esophagitis	0.81	0.76	0.78	
normal-cecum	0.92	0.95	0.93	**0.886**
normal-pylorus	0.96	0.99	0.98	
normal-z-line	0.77	0.81	0.79	
polyps	0.91	0.85	0.88	
ulcerative-colitis	0.92	0.84	0.84	
weighted average	**0.89**	**0.89**	**0.89**	

IV. CONCLUSION

Various experiments on the following Convolutional Neural Network architectures used via transfer learning have been conducted for gastrointestinal image classification: VGG19, ResNet50V2, ResNet152V2, EfficientNetV2B0, EfficientNetV2B3, InceptionV3, DenseNet201, Xception. Within the experiments, the same initially constructed pre-trained base model was transferred to the learning stage of the publicly available Kvasir dataset using each different architecture. The dataset, including 1000 images per 8 different classes, has been splitted in such a fashion that 20% of the images have been used for testing. In this manner, to test and fairly evaluate the prediction performance of the models for each of 8 classes in a balanced way, the test images were arranged to contain equally 200 images from each class. Based on the experiment results, VGG19 has achieved the best test accuracy with 88.6% amongst the models.

Regarding detailed results of the experiments conducted with VGG19 architecture, the model has performed 89% on average in terms of each of the precision, recall and F1-score metrics.

As the future work, metaheuristic approaches for hyperparameter optimization and feature selection would be considered to improve the performance results of the models. Furthermore, hybrid combinations of the models would be investigated and the corresponding extra experiments could be conducted.

REFERENCES

[1] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y, Zhu, H. Zhu, H. Xiong, Q He, *A Comprehensive Survey on Transfer Learning*, In Proceedings of the IEEE, vol. 109, no. 1, pp. 43-76, 2021, doi: 10.1109/JPROC.2020.3004555.

[2] K. Pogorelov, K. C. Griwodz, S. L. Eskeland, T. Lange, D. Johansen, C. Spampinato, D. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, P. Halvorsen, *Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection*, In MMSys'17 Proceedings of the 8th ACM on Multimedia Systems Conference, pp. 164-169, Taiwan, 2017.

[3] "KvasirV2 Dataset", https://datasets.simula.no/kvasir/, (accessed on 10 July 2024).

[4] R. Kumar, P. Kumbharkar, S. Vanam, S. Sharma, *Medical Images Classification Using Deep Learning: A Survey*, Multimedia Tools and Applications, vol. 83, pp. 19683–19728, 2024, doi: 10.1007/s11042-023-15576-7.

[5] H. Malik, A. Naeem, A. Sadeghi-Niaraki, R. A. Naqvi, S. W. Lee, *Multi-classification Deep Learning Models for Detection of Ulcerative Colitis, Polyps, and Dyed-lifted Polyps Using Wireless Capsule Endoscopy Images*, Complex & Intelligent Systems, vol. 10, pp. 2477–2497, 2024, doi: 10.1007/s40747-023-01271-5.

[6] S. Siddiqui, T. Akram, I. Ashraf, M. Raza, M. A. Khan, R. Damaševičius, *CG-Net: A Novel CNN Framework for Gastrointestinal Tract Diseases classification*, International Journal Of Imaging Systems and Technology, 34(3):e23081, 2024, doi:10.1002/ima.23081.

[7] M. Uçan, B. Kaya, M. Kaya, *Multi-Class Gastrointestinal Images Classification Using EfficientNet-B0 CNN Model*, International Conference on Data Analytics for Business and Industry (ICDABI), pp. 1-5, Bahrain, 2022, doi: 10.1109/ICDABI56818.2022.10041447.

[8] K. Ramamurthy, T. T. George, Y. Shah, P. Sasidhar, *A Novel Multi-Feature Fusion Method for Classification of Gastrointestinal Diseases Using Endoscopy Images*, Diagnostics, 12(10):2316, 2022, doi: 10.3390/diagnostics12102316

[9] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T.D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen, T. de Lange, *HyperKvasir, A Comprehensive Multi-Class Image and Video Dataset for Gastrointestinal Endoscopy*, Sci Data, 7(1):283, 2020, doi: 10.1038/s41597-020-00622-y, PMID: 32859981; PMCID: PMC7455694.

[10] M. Altıntaş and M. Ü. Öziç, *Performance Evaluation of Different Deep Learning Models for Classifying Ischemic, Hemorrhagic, and Normal Computed Tomography Images: Transfer Learning Approaches*, Konya Journal of Engineering Sciences, vol. 12:2, pp. 465–477, 2024, doi: 10.36306/konjes.1346134.

[11] S. D. Khan, S. Basalamah, A. Lbath, *Multi-Module Attention-Guided Deep Learning Framework for Precise Gastrointestinal Disease Identification in Endoscopic Imagery*, Biomedical Signal Processing and Control, vol. 95, part B, 106396, 2024, doi: 10.1016/j.bspc.2024.106396.

[12] "Keras Available Models", https://keras.io/api/applications/, (accessed on 10 July 2024).