








Empirical Study of Attention-Based Models for Automatic Classification of Gastrointestinal Endoscopy Images

Ricardo Espantaleón-Pérez¹ , Isabel Jiménez-Velasco¹ ,
Rafael Muñoz-Salinas^{1,2} , and Manuel J. Marín-Jiménez^{1,2}  

¹ Department of Computing and Numerical Analysis, University of Córdoba, Córdoba, Spain

mjmarin@uco.es

² Maimonides Institute for Biomedical Research of Córdoba (IMIBIC), Córdoba, Spain

Abstract. Automatic and accurate analysis of medical images is a subject of great importance in our current society. In particular, this work focuses on gastrointestinal endoscopy images, as the study of these images helps to detect possible health conditions in those regions. Published works on this topic mainly used traditional classification methods (e.g., Support Vector Machines) or more modern techniques, such as Convolutional Neural Networks. However, little attention has been paid to more recent approaches such as Transformers or, in general, Attention-based Deep Neural Networks. This work aims to evaluate the performance of state-of-the-art attention-based models on the problem of classification of gastrointestinal endoscopy images. The experimental results on the challenging Hyper-Kvasir dataset indicate that attention-based models achieve performance equal to or better than that obtained by previous models, needing fewer parameters. In addition, a new state of the art on Hyper-Kvasir (i.e., 0.636 F1-Macro) is obtained by the fusion of two MobileViT models with only 20M parameters. The source code will be published here: <https://github.com/richardesp/Attention-based-models-for-Hyper-Kvasir/>.

Keywords: Attention · Transformers · Endoscopy · Medical Image

1 Introduction

The analysis of gastrointestinal endoscopy images is of great importance for the accurate diagnosis and treatment of a wide range of gastrointestinal disorders. Endoscopy images provide clinicians with direct visual access to the inner surfaces of the gastrointestinal tract, enabling them to identify abnormalities

Supported by projects TED2021-129151B-I00/AEI/10.13039/501100011033/European Union NextGenerationEU/PRTR and PID2019-103871GB-I00 of the Spanish Ministry of Economy, Industry and Competitiveness.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
N. Tsapatsoulis et al. (Eds.): CAIP 2023, LNCS 14185, pp. 98–108, 2023.
https://doi.org/10.1007/978-3-031-44240-7_10

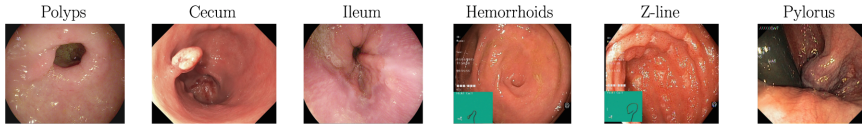


Fig. 1. Hyper-Kvasir dataset of annotated endoscopy images. Sample images belonging to six out of the 23 available classes.

such as ulcers, polyps, and tumors. The timely and accurate diagnosis of these abnormalities is crucial for effectively treating and preventing complications. The development of computer vision systems able to support the diagnosis of medical doctors is currently an important line of research [15].

In recent years, several approaches have addressed the problem of automatic classification of gastrointestinal endoscopy images. There are several works focused on the classification and segmentation of cancerous artifacts in the digestive tract, including polyps based on the quality of the present mucosa [3, 4, 14]. In turn, these types of models can serve as decision support for the timely detection of various gastric cancers [12]. These works use models mostly based on Convolutional Neural Networks with different levels of complexity, such as ResNet [3], DenseNet [3, 16], MobileNet [9, 19] or EfficientNet [10]. Despite these attempts, the problem of automatic classification of endoscopy images is far from solved.

From the computer vision viewpoint, in recent years, attention mechanisms have shown that image classification accuracy can be improved in several tasks [7], as they allow the models to focus on the most relevant parts of the image. In this work, we are interested in, on the one hand, investigating if models incorporating attention mechanisms are able to improve the classification performance on endoscopy images. And on the other hand, if it is possible to find a compact model, in terms of parameters, offering a good trade-off between accuracy and computational cost.

For this purpose, we have selected four families of state-of-the-art attention-based models (MobileViT, CoAtNet, CMT and DaViT) and the largest public dataset of endoscopy images, i.e., Hyper-Kvasir [3]. Then, the main contribution of this paper is an extensive evaluation of four types of attention-based models, with different levels of complexity, on the largest public annotated dataset of gastrointestinal endoscopy images, Hyper-Kvasir. The results of this study include a new state of the art on the classification task using a moderated number of model parameters (i.e., 20M using MobileViT models).

The rest of this paper is organized as follows. Section 2 presents the attention-based models evaluated in this study. Then, the dataset selected to perform the experiments is described in Sect. 3. The experimental results are presented in Sect. 4. And finally, the paper concludes in Sect. 5 including future research lines.

2 Attention-Based Models

Attention-based models initially emerged in the field of Natural Language Processing and were popularized by the success of models such as Transformer [17].

Since then, attention-based models have expanded to other fields, such as Computer Vision and Signal Processing, and have been demonstrated to be highly effective in various tasks. Attention-based models for vision are Deep Learning models that use the attention mechanism to process images and videos more effectively. Unlike traditional neural networks that process the entire image uniformly (treating all parts of the input sequence equally), attention-based models focus on specific parts of the image that are most relevant to the task at hand, focusing on a few specific aspects at a time and ignoring the rest. They do this by assigning different weights to each part of the image according to its importance to that task. This allows larger, more complicated tasks to be reduced to smaller, more manageable areas of attention to understand and process them sequentially. In this section, we will present the attention-based models for vision that have been selected for our experimental study.

2.1 MobileViT Family

MobileViT [13] is a family of computer vision models based on the Transformer (ViT) architecture [7] and characterized by their computational efficiency and their ability to process large-scale images. These models use attention blocks to process images as patches instead of traditional convolution. MobileViT models have been optimized for implementation on mobile devices, making them lighter and more efficient in terms of computational resources and memory consumption. Different versions of this architecture have been designed to suit different performance and size requirements. These models have been demonstrated to be very effective in various computer vision tasks, such as object detection and image classification. For the proposed study, two versions of the MobileViT architecture are used: the XS version of MobileViT, which has 2M of parameters, and a larger version with 18M of parameters.

2.2 CoAtNet

CoAtNet (Convolutional Attention Network) [5] is a deep neural network model that combines convolutions with attention mechanisms to take advantage of both in extracting features from images of different data sizes. The architecture uses cascaded attention blocks combined with convolutional layers to capture contextual and spatial features of images efficiently. Specifically, the architecture comprises parallel contextual attention blocks (CABs) and convolutional blocks. CAB blocks are responsible for extracting meaningful features from images using attention. One of the outstanding features of CoAtNet is its adaptive resizing mechanism, which automatically adjusts the input size to match the size required by the attention layers. This allows the model to process data of any input size without additional preprocessing.

2.3 CMT

CMT (CNNs meet Transformers) [11] is a deep neural network model applied to vision. It is a hybrid architecture incorporating transformer attention blocks

into a standard convolutional network. The attention blocks are inserted into different layers of the CNN, allowing the model to capture both local and global features of the images. Attention is applied in parallel across channels, rows and columns, allowing the model to capture spatial and channel relationships at various scales. The idea behind CMT is to take advantage of the ability of convolutions to capture local patterns and the ability of transformer attention blocks to model long-distance relationships and nonlinear interactions between different features. In addition, CMT introduces an attention modulation technique to adapt attention based on local image features. This allows the model to adjust attention to specific regions of the image that are more relevant to the task at hand.

2.4 DaViT

DaViT (Dual Attention Vision Transformers) [6] is a deep learning model based on Vision Transformers (ViT) that uses two types of attention to improve performance in computer vision tasks: spatial attention and channel attention. Spatial attention refers to the model’s ability to focus on specific regions of the image, which allows the model to pay more attention to important image features and ignore irrelevant features. Channel attention refers to the model’s ability to focus on specific features in different layers of the neural network, which allows the model to learn to distinguish different types of features in the image. Compared to the ViT model, DaViT uses a dual-path structure in its architecture, allowing the model to capture global and local information from the image.

3 Dataset and Metrics

This section presents both the dataset used for performing our experimental study and the metrics used to compare the selected models (Sect. 2).

3.1 Hyper-Kvasir Dataset

The Hyper-Kvasir dataset [3] is currently the largest public dataset of colonoscopies in computer vision. It contains a total of 10,662 labeled images representing 23 different classes. Some example images are shown in Fig. 1. The classes are structured according to the Gastrointestinal (GI) tract’s location and the pathological finding type.

The dataset contains images from four high-level categories: (i) **Anatomical landmarks**, which are used during the endoscopy process to obtain references and confirm that all critical areas have been examined, existing in both the upper and lower GI tracts; (ii) **Mucosal quality**, where complete visualization of the mucosa is crucial for detecting pathological findings, with the Boston Bowel Preparation Scale (BBPS) used as a classification measure in the colon; (iii) **Pathological findings**, which are anomalies that can affect all parts of the gastrointestinal tract depending on the pathology being treated and are

often inferred from the intestinal mucosa walls; and, *(iv)* **Therapeutic interventions**, where intervention is required when a lesion or pathological finding is discovered, such as lifting and resecting a polyp, dilation of a stenosis, or injection of a bleeding ulcer.

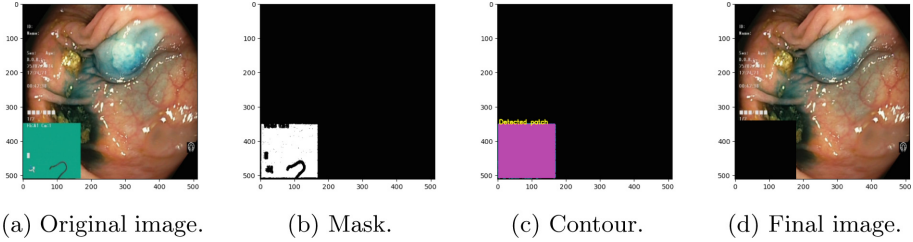


Fig. 2. Green patch removal. Process of removing the green patch broken down into the different steps carried out.

Data Preprocessing. We have observed the presence of green patches in the images for certain critical classes. These patches appear in the same location in the image but vary in position along the y -axis and size. In some cases, these green patches can occupy a significant portion of the image (around 32% of the total area). In our opinion, the presence of these patches may bias the training and resulting model. We have developed a method to remove them automatically. In particular, contour detection algorithms [2] have been used to subsequently perform the most precise polygonal adjustment possible to avoid removing too much information from the original image.

The process involves the following main steps (see Fig. 2): *(i)* *Detection and thresholding of the green pixels* by defining the color interval that includes the ‘green’ hue and masking the values meeting the color condition; and *(ii)* *Contour detection algorithm*, which is a process performed to facilitate the detection of the contour of the rectangular region on the original image. Finally, a bitwise logical operation is applied to the mask to remove the defined rectangular region.

3.2 Performance Metrics

Due to the class imbalance present in the dataset, the model’s performance will be evaluated using micro and macro F1-scores. These performance metrics are preferred over accuracy [14] and other metrics, providing a more accurate conclusion about the model’s performance. The equations of the metrics are shown below:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad F1_{\text{macro}} = \frac{1}{N} \sum_{i=1}^N F1_i \quad (1)$$

$$F1_{\text{micro}} = \frac{2 \cdot \sum_{i=1}^N \text{TP}_i}{2 \cdot \sum_{i=1}^N \text{TP}_i + \sum_{i=1}^N \text{FP}_i + \sum_{i=1}^N \text{FN}_i} \quad (2)$$

where TP, FP and FN indicate the True Positives, False Positives and False Negatives, respectively, and N the number of classes.

A 2-fold evaluation will be carried out using 50% of the stratified data, using the splits proposed by the Hyper-Kvasir article [3], and the models will be evaluated using F1 metrics.

It is worth mentioning that, in this case, improving the macro variant over the micro variant is considered more relevant since the classes related to cancerous artifacts in colonoscopies have a very limited number of samples. Therefore, it is convenient to weigh and average the classes equitably to improve the classifier.

Table 1. Training details. A summary of the model architectures, batch sizes, optimizers, learning rates, and weight decay values used in the experiments.

Model	Batch Size	Optimizer	Learning Rate	Weight Decay
MobileViT	64	AdamW	0.0001	0.00001
CoAtNet	16	AdamW	0.0001	0.00001
CMT	32	AdamW	0.001	0.0001
DaViT	16	AdamW	0.001	0.00001
MobileViT Large	16	Adam	0.0001	0.001

4 Experiments and Results

The experiments have been designed to directly compare with the original Hyper-Kvasir article without applying any additional pre-processing [3], considering only the removal of green patches to favor the generalization of the model.

To determine the improvements compared to the previous state of the art, the training has been carried out using the original splits and keeping the green patches, in order to conclude whether attention-based models and hybrid architectures improve compared to the previous ones.

4.1 Implementation Details

The purpose of these experiments is to determine whether current attention-based architectures are interesting compared to the previous state-of-the-art ones and whether there are significant improvements to be made by opting for these models for this type of problem. For this analysis, as discussed in Sect. 3, the green patches have been removed in order to draw conclusions about possible biases in the model. In turn, training has been conducted both with and without the green patches, using the splits proposed by the Hyper-Kvasir article [3], to determine whether these models outperform the previous state-of-the-art ones.

All models are pre-trained on the ImageNet dataset because they outperform the previous state-of-the-art ones when pre-trained on sufficiently large datasets [7]. However, we do not freeze any weights. In our early experiments, models without pre-training were also tested, but the results were lower than the ones obtained with pre-trained models. The images have been rescaled to a resolution of 224×224 . In our study, we applied several data augmentation techniques, including random image rotation up to a maximum of 40° , random width and height shifts with a maximum range of 20% of the image size, random shearing of up to 20% to mimic perspective or viewing angle, and a random zoom factor of up to 20% to simulate different distances between the object and the camera. Additionally, we allowed horizontal flipping of images, effectively doubling the amount of available training data and helping models learn symmetries. We have used third-party libraries for the attention-based models, available online¹. All input and output layers have been adapted to the respective format required by each of the models.

Table 2. Architecture comparison. Results comparing the pre-processed dataset removing green patches and using the default dataset with green patches. The results are obtained using data augmentation and class weighting.

Model	Parameters	FLOPs	Non-green patches		Green patches	
			F1-Macro	F1-Micro	F1-Macro	F1-Micro
MobileViT	2M	1.6×10^{10}	<i>0.618</i>	<i>0.890</i>	<i>0.630</i>	0.892
MobileViT Large	18M	11.2×10^{10}	0.604	0.892	0.634	0.900
CMT	9M	2.6×10^{10}	0.579	0.873	0.605	0.854
CoAtNet	23M	8.3×10^{10}	0.619	0.889	0.626	<i>0.895</i>
DaViT	87M	31.1×10^{10}	0.598	0.876	0.609	0.878

In order to improve the macro variant, given the class imbalance, we apply class weighting using the method *compute class weight* indicated by the Scikit Learn library [1].

As a preliminary starting point, a Bayesian hyperparameter optimization has been performed with the purpose of optimizing resources regarding grid-based search methods [18]. The possible values of the learning rate and weight decay are in the set $\{0.00001, 0.0001, 0.001, 0.01\}$. The set of values for the batch size is as follows $\{16, 32, 64, 128, 256\}$. Adam, AdamW, and AdaDelta have been used as optimizers. All hyperparameters used for each architecture are provided in Table 1 to facilitate the training of the models for the proposed problem.

4.2 Comparison of Architectures

After the previously performed hyperparameter optimization, the training of the models presented in Sect. 2 will be carried out using the preprocessed dataset indicated in Sect. 3.

¹ Base models: https://github.com/leongarse/keras_cv_attention_models.

As previously commented, given the data imbalance, it has been considered appropriate to prioritize macro metrics through weighting in those classes with fewer samples. Therefore, in certain cases, micro metrics worsen slightly in exchange for very favorable improvements in macro F1 compared to the results obtained in previous articles [3, 14]. The obtained results are summarized in Table 2. We observe that the MobileViT architectures offer the best average results, followed by CoAtNet.

4.3 Influence of the Green Patches

As previously mentioned, the original dataset presents green patches on specific images (see Fig. 2). The results proposed by the Hyper-Kvasir article [3] used the default dataset without any preprocessing of the green patches. Therefore, we decided to train the models both with and without the green patches, using the same hyperparameters and architectures, to compare their influence on the model results. Note that if we wanted to obtain a classification model able to deal with endoscopy images from a third-party dataset, we could not assume the existence of these green landmarks. The obtained results are included in Table 2.

Table 3. State of the art. Our best models (w/o removing green patches) are compared to other published models. The best directly comparable results (same splits and image resolution; top five rows) are marked in bold.

Model	Split	Resolution	Parameters	F1-Macro	F1-Micro
MobileViT Large (ours)	2-fold	224 × 224	18M	0.634	0.900
Late fusion (ours)*	2-fold	224 × 224	20M	0.636	0.905
DenseNet-161 [3]	2-fold	224 × 224	28M	0.619	0.907
ResNet-152 [3]	2-fold	224 × 224	60M	0.606	0.906
ResNet-152 & DenseNet-161 [3]	2-fold	224 × 224	88M	0.617	0.910
Teacher-Student(EfficientNetB6) [10]	2-fold	336 × 336	43.3M	-	0.886
DenseNet-161 [16]	2-fold	512 × 512	28M	0.635	-
MobileNet-V2 [19]	5-fold	512 × 512	3.4M	0.651	-
ResNet-50x1/BiT-M[8]	5-fold	640 × 512	-	-	0.918
MobileNet-V2 [9]	5-fold	Unknown	3.4M	0.641	-

(*): Averaged MobileViT Large & MobileViT with $\alpha=0.6$

We observe that, in general, the results obtained using images without the green patches are lower than the ones obtained with the unmodified images. This suggests that those green regions contain information used by the models (see Fig. 3). This fact may imply that, in those cases, the actual features of the digestive tract are not properly exploited.

Note that we have also experimented with a training set combining the original images and the ones where the green patches were removed. As a reference,

for the MobileViT-Large the obtained performance values are 0.625 and 0.889, for F1-macro and F1-micro, respectively.

4.4 Comparison to the State of the Art

Inspired by [3], we perform a late fusion of the two best models obtained in our previous experiments. Late fusion involves averaging the predictions of the last softmax layer between models, setting a weighting parameter α that determines the weight to assign to each model, in this case, to the first model indicated. The final results are included in Table 3. All model comparisons have been made with those that used the original 23 classes without any grouping, in order to have a common starting point despite the differences in experimental conditions that certain articles present (i.e., different splits and image resolution).

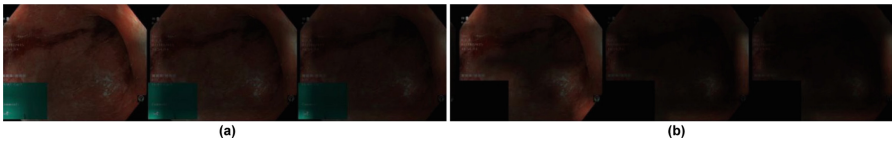


Fig. 3. Extraction of attention on trained models (a) with and (b) without green patches. Example of bias in the presence of the green patch at classification time. Brighter areas indicate higher attention. (Color figure online)

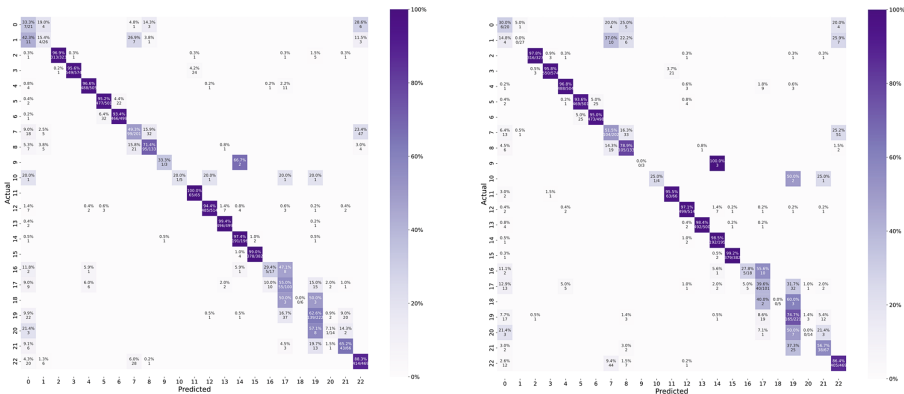


Fig. 4. Confusion matrices on Hyper-Kvasir. Obtained from our best model on the corresponding test splits. Zoom in for details. (View in digital format).

We observe improvement by performing the ensemble of models compared to their results individually. Our best fusion model, obtained through the ensemble of MobileViT architectures, has a total of 20M parameters (2M+18M), achieving an F1-macro value of 0.636. As a reference, the ensemble model described in [3] has 88M parameters, achieving 0.617 F1-macro performance.

In the resulting confusion matrices (see Fig. 4) on the original splits in tests proposed by the Hyper-Kvasir article [3], it is possible to observe the same classification problem that existed in previously proposed models, where the classes representing the different degrees of ‘ulcerative colitis’ pose a difficulty in classifying these instances.

For the sake of completeness, we have included in Table 3 other methods existing in the literature that present classification results on Hyper-Kvasir. However, their experimental setup is not directly comparable to ours. For example, a higher F1-macro value is reported by the recent work of [19], where the image resolution is twice ours and the number of training samples is larger (5-fold cross-validation) than ours. Using the same 2-fold split we use, the work in [16] presents the results obtained using a DenseNet-161 (28M parameters) at a resolution of 512×512 pix. This higher resolution model achieves a similar F1-Macro compared to our lower resolution MobileViT with only 18M parameters. These results support the benefit of using attention-based models for this task.

5 Conclusions and Future Work

This work presents an empirical study of attention-based models for classifying gastrointestinal endoscopy images. The selected models were evaluated on the challenging Hyper-Kvasir dataset [3], achieving state-of-the-art results on the classification task. The results show that the lightweight models from the MobileViT family offer very favorable results considering their reduced number of parameters, and compared to previous non-attention-based models. In addition, a late fusion approach on two selected models shows a boost in the classification performance.

This work has also studied the influence on the classification accuracy of some existing green landmarks in the images of the Hyper-Kvasir dataset. The experimental results suggest that the presence of those image artifacts may compromise the generalization capability of the trained models on this dataset, for direct use on other endoscopy datasets.

As future work, we plan to study in-depth the image regions attended by the different models and validate their possible medical relevance in collaboration with medical doctors. In addition, we may consider performing pre-training with the rest of the unlabeled Hyper-Kvasir dataset using unsupervised learning techniques (e.g., AutoEncoders or clustering).

References

1. Compute class weight in sklearn utils. https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html
2. Structural analysis and shape descriptors. https://docs.opencv.org/3.4/d3/dc0/group_imgproc_shape.html#ga17ed9f5d79ae97bd4c7cf18403e1689a
3. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci. Data* **7**(1), 283 (2020)

4. Chen, P.J., Lin, M.C., Lai, M.J., Lin, J.C., et al.: Accurate classification of diminutive colorectal polyps using computer-aided analysis. *Gastroenterology* **154**(3), 568–575 (2018)
5. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: marrying convolution and attention for all data sizes. In: *NEURIPS*, vol. 34 (2021)
6. Ding, M., Xiao, B., Codella, N., Luo, P., et al.: DaViT: dual attention vision transformers. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) *ECCV 2022*. LNCS, vol. 13684, pp. 74–92. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20053-3_5
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: *ICLR* (2021)
8. Galdran, A., Carneiro, G., Ballester, M.A.G.: A hierarchical multi-task approach to gastrointestinal image analysis. In: *ICPR Workshops and Challenges* (2021)
9. Galdran, A., Carneiro, G., González Ballester, M.A.: Balanced-MixUp for highly imbalanced medical image classification. In: de Bruijne, M., et al. (eds.) *MICCAI 2021*. LNCS, vol. 12905, pp. 323–333. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_31
10. Gjestang, H.L., Hicks, S.A., Thambawita, V., Halvorsen, P., Riegler, M.A.: A self-learning teacher-student framework for gastrointestinal image classification. In: *2021 IEEE 34th International Symposium on CBMS* (2021)
11. Guo, J., Han, K., Wu, H., Tang, Y., et al.: CMT: Convolutional neural networks meet vision transformers. In: *IEEE/CVF CVPR* (2022)
12. Hirasawa, T., Aoyama, K., Tanimoto, T., Ishihara, S., et al.: Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer* **21**, 653–660 (2018)
13. Mehta, S., Rastegari, M.: MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. In: *ICLR* (2022)
14. Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., et al.: Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: *Proceedings of ACM on Multimedia Systems Conference*, pp. 164–169 (2017)
15. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. *Ann. Rev. Biomed. Eng.* **19**, 221–248 (2017)
16. Thambawita, V., Strümke, I., Hicks, S.A., Halvorsen, P., et al.: Impact of image resolution on deep learning performance in endoscopy image classification: an experimental study using a large dataset of endoscopic images. *Diagnostics* **11**(12), 2183 (2021)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al.: Attention is all you need. In: *NeurIPS*, vol. 30 (2017)
18. Wu, J., Chen, X.Y., Zhang, H., Xiong, L.D., et al.: Hyperparameter optimization for machine learning models based on Bayesian optimization. *J. Electron. Sci. Technol.* **17**(1), 26–40 (2019)
19. Yue, G., Wei, P., Liu, Y., Luo, Y., et al.: Automated endoscopic image classification via deep neural network with class imbalance loss. *IEEE Trans. Instrum. Meas.* **72**, 1–11 (2023)