Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 14(3), 988-999, 2024
Journal of the Institute of Science and Technology, 14(3), 988-999, 2024
ISSN: 2146-0574, eISSN: 2536-4618
Bilgisayar Mühendisliği / *Computer Engineering* **DOI: 10.21597/jist.1501787**
Araştırma Makalesi / *Research Article*
Geliş tarihi / *Received*: 15.06.2024 **Kabul tarihi / *Accepted*: 21.07.2024**

To Cite: Ayan, E. (2024). Classification of Gastrointestinal Diseases in Endoscopic Images: Comparative Analysis of Convolutional Neural Networks and Vision Transformers . *Journal of the Institute of Science and Technology*, 14(3), 988-999.

Classification of Gastrointestinal Diseases in Endoscopic Images: Comparative Analysis of Convolutional Neural Networks and Vision Transformers

Enes AYAN[1*]

Highlights:

- Transfer Learning
- Fine Tuning
- Endoscopic Image Classification
- DenseNets

Keywords:

- Medical Image Classification
- Convolutional Neural Networks
- Vision Transformers
- Fine Tuning
- Transfer Learning
- Gastrointestinal Diseases

ABSTRACT:

Gastrointestinal (GI) diseases are a major issue in the human digestive system. Therefore, many studies have explored the automatic classification of GI diseases to reduce the burden on clinicians and improve patient outcomes for both diagnosis and treatment purposes. Convolutional neural networks (CNNs) and Vision Transformers (ViTs) in deep learning approaches have become a popular research area for the automatic detection of diseases from medical images. This study evaluated the classification performance of thirteen different CNN models and two different ViT architectures on endoscopic images. The impact of transfer learning parameters on classification performance was also observed. The tests revealed that the classification accuracies of the ViT models were 91.25% and 90.50%, respectively. In contrast, the DenseNet201 architecture, with optimized transfer learning parameters, achieved an accuracy of 93.13%, recall of 93.17%, precision of 93.13%, and an F1 score of 93.11%, making it the most successful model among all the others. Considering the results, it is evident that a well-optimized CNN model achieved better classification performance than the ViT models

[1]Enes AYAN (**Orcid ID:** 0000-0002-5463-8064), Kırıkkale University, Faculty of Engineering and Natural Sciences, Department of Computer Engineering, Kırıkkale, Türkiye

***Sorumlu Yazar/*Corresponding Author:** Enes AYAN, e-mail: enesayan@kku.edu.tr

INTRODUCTION

The gastrointestinal (GI) system is a tubular system that performs a range of digestive functions, including chewing, swallowing, digestion, absorption, and excretion. The GI system includes several organs, starting from the mouth and extending to the anus. The sequence of organs in the GI system comprises the mouth, pharynx, esophagus, stomach, small intestines, large intestines, rectum, and anal canal (Sivari et al., 2023). The human GI system may be affected by a number of diseases, the most prevalent of which are esophageal, stomach and colorectal cancer. For example, stomach cancer is one of the GI disorders and is the fourth most common type of cancer in women and the seventh most common in men (Siddiqui et al., 2024). According to research, the success rate of treating cancer diagnosed at the second stage is 91.5%, whereas at the fourth stage, this rate drops to 16.4% (Katai et al., 2018). Therefore, early diagnosis of stomach cancer is of great importance for the success of the treatment process. Endoscopy is a commonly employed imaging method in the diagnosis of GI cancer types in GI system. It is a diagnostic procedure that employs the use of a lighted camera at the tip of a flexible tube to image the internal organs of the digestive system. This allows for the detection of potential issues. There are different types of endoscopies, including gastroscopy, colonoscopy, magnifying endoscopy, and capsule endoscopy (Sivari et al., 2023). The manual evaluation of endoscopic images is a time-consuming and labor-intensive process. Additionally, subjective evaluations can result in a high rate of misdiagnoses, leading to delays in the application of effective treatment. Statistics reveal that about 22% to 28% of polyps and 20% to 24% of adenomas are either missed or misdiagnosed (Leufkens et al., 2012). It is probable that missed polyps will develop into cancer. Therefore, there is a pressing need for the development of reliable computer-aided diagnostic systems that are capable of automatically analyzing endoscopic images and providing a secondary opinion to experts. In recent years, deep learning methods have been employed to address a variety of computer vision problems, including image classification, segmentation, and object detection (Chai et al., 2021; Pacal, 2024; Sermet and Pacal, 2024), have also been preferred by researchers for the analysis of endoscopic images. A summary of the studies in the literature is as follows:

Agrawal et al. employed transfer learning along with Convolutional Neural Networks (CNNs) for classification of endoscopic images (Agrawal et al., 2019). They developed a metric to determine the model to be used for transfer learning. The proposed method achieved a classification accuracy of 83.8%, as reported in the study. Zhang et al. improved the architecture of the single-shot multi-box detector in the detection of polyps (Zhang et al., 2019). The method proposed in the study was successful in detecting polyps with a mean average precision (mAP) rate of 90.4%. Gjestang et al. proposed a semi-supervised teacher-student learning method aimed at enhancing the classification performance of endoscopic images, achieving an accuracy of 89.3% (Gjestang et al., 2021). Meanwhile, Losenko et al. introduced a deep convolutional neural network (CNN)-based spatial attention mechanism for the classification of gastrointestinal (GI) diseases, utilizing encoder-decoder layers in their implementation (Lonseko et al., 2021).With their proposed method, 93.19% classification success was achieved in classifying endoscopic images. Yogapriya et al. trained VGG-16, ResNet-18 and GoogLeNet architectures for the classification of endoscopic images using the transfer learning method (Yogapriya et al., 2021). In the study, it was reported that the VGG-16 model achieved 96.33% accuracy, which was more successful than other models. Karaman et al. utilized the artificial bee colony algorithm to determine the training hyperparameters of various You Only Look Once (YOLO) models (Karaman et al., 2023). Optimized YOLO models were used for the detection of polyps from endoscopic images in the study. It was reported that the YOLOv4 model outperformed other models with a 78% mean average

precision (mAP) value in the study. Mukhtorov et al. suggested an interpretable deep learning method based on Grad-CAM combined with ResNet152 for the classification of endoscopy images (Mukhtorov et al., 2023). In the study, the proposed method achieved a classification accuracy of 93.46%. Demirbaş et al. developed an architecture called Spatial Attention ConvMixer for classifying endoscopic images (Demirbaş et al., 2024). They compared the classification performance of their developed model with models such as Vanilla Vision Transformer (ViT), Swin Transformer, ConvMixer, MLPMixer, ResNet50 and SqueezeNet. The study reported that the developed model achieved a classification accuracy of 93.37%. Huo et al. introduced Self-Peripheral-Attention (SPA), an novel methodology that incorporates peripheral vision modeling into the attention mechanism (Huo et al., 2024). This approach aims to enhance the accuracy and efficiency of classification and segmentation tasks in endoscopic imaging, achieving a classification accuracy of 92.7%.

The literature shows that there has been research into the segmentation, detection and classification of polyps from endoscopic images. In these studies, CNN-based models are mostly preferred. The results show the clear success of CNNs and learning in classifying endoscopic images. However, improving the classification performance of endoscopic images remains an open area of research. Our contributions to improving the classification performance of endoscopic images in this study are as follows. A wide range of pre-trained CNN models (thirteen) have been evaluated for the classification performance of endoscopic images using transfer learning strategy. The impact of transfer learning parameters on the classification performance of the CNN model has been observed. A comparison was made between the performance of CNN models and ViTs in the classification of endoscopic images.

MATERIALS AND METHODS

In this study, the classification performance of thirteen different pre-trained CNN models and two different ViT models on endoscopic images was compared. The effect of changing the tra1nsfer learning hyperparameters of the most successful CNN model on classifying performance was observed. Figure 1 provides a summary of the study.

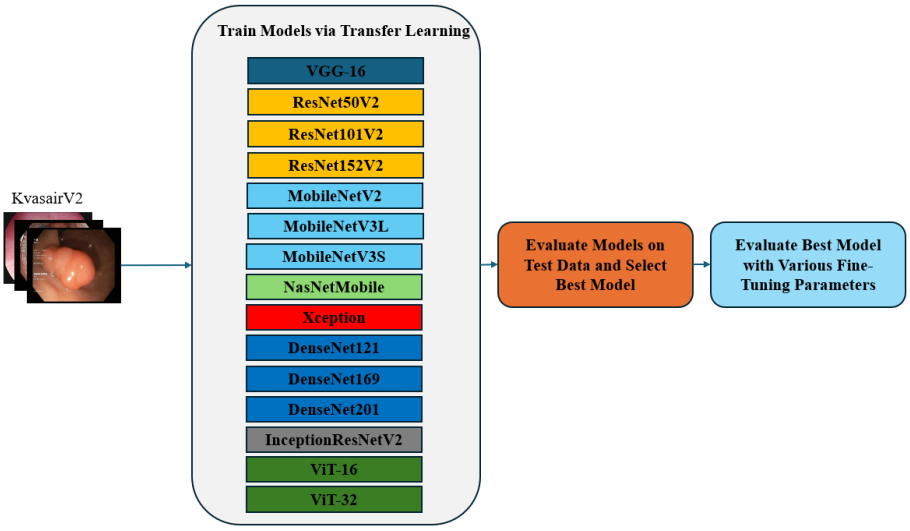

Figure 1. A visual representation of the study

Dataset

The dataset used in the study consists of endoscopic images in the field of gastroenterology, created to support research in medical image analysis. This dataset was collected by the health organization in Norway and is called Kvasir-V2 (Pogorelov et al., 2017). The Kvasir-V2 dataset was

shared as open source in 2017 as part of the Mediaeval Medical Multimedia Challenge. There are eight classes in total in the dataset and each class contains 1000 images. Images are in 1920x1072 resolution from 720x576. The dataset was organized based on three key anatomical landmarks and three clinically significant findings. It also features two categories of images related to endoscopic polyp removal. The anatomical landmarks include the z-line, pylorus, and cecum, while the pathological findings encompass esophagitis, polyps, and ulcerative colitis. Additionally, various images related to lesion removal are provided in the dataset; for example, dyed and lifted polyps, and dyed resection margins. Figure 2 shows some example images from the dataset. In the study, the dataset was randomly divided into training, validation, and test sets with a ratio of 80:10:10. Detailed information on the class distribution is given in Table 1. Also, to prevent the models from overfitting, various data augmentation techniques were applied to the training dataset during training. These techniques include width shift range and height shift range with a 0.2 ratio, shear range with a 0.2 ratio, zoom range with a 0.2 ratio, and vertical and horizontal flips.

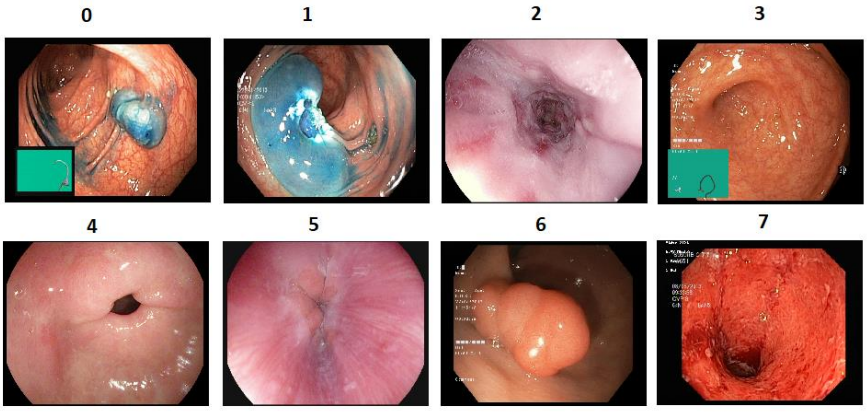

Figure 2. Image samples and their id values from Kvasir-V2 dataset

Table 1. The class names and dataset distribution in train, validation and test groups

Class Name-(Id)	Train	Validation	Test
Dyed-lifted-polyps (0)	800	100	100
Dyed-resection-margins (1)	800	100	100
Esophagitis (2)	800	100	100
Normal cecum (3)	800	100	100
Normal pylorus (4)	800	100	100
Normal-z-line (5)	800	100	100
Polyps (6)	800	100	100
Ulcerative colitis (7)	800	100	100
Total	6400	800	800

Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are deep learning algorithms widely used in image processing and computer vision. CNNs automatically learn problem-related features from raw images and exhibit high performance in tasks such as classification, object detection, segmentation and recognition (Li et al., 2022). A traditional basic CNN architecture consists of convolutional layers, pooling layers and fully connected layers. Sequential convolutional and pooling layers are used to automatically extract features. The extracted features are used in fully connected layers to perform classification or regression, depending on the type of problem to be solved (Li et al., 2022). Various CNN architectures have been developed in the literature. In this study, thirteen pre-trained CNN architectures are used to classify endoscopic images in eight class.

DenseNet

DenseNet (Dense Convolutional Network) is proposed by (Huang et al., 2018). This architecture features a unique connectivity pattern where each layer receives feature maps from all preceding layers and passes its output to all subsequent layers. This connectivity helps DenseNet mitigate the vanishing gradient problem commonly seen in deep networks and enables more efficient parameter usage. The fundamental building blocks of DenseNet are called "dense blocks." Within each dense block, every layer takes the output of all previous layers as its input and includes its own output as part of this set. This enhances the flow of information and the propagation of gradients, allowing the network to continue learning effectively even as it becomes deeper. Moreover, this architecture increases parameter efficiency, achieving better performance with fewer parameters. These characteristics of DenseNet make it particularly effective when working with limited datasets or performing complex tasks that require deeper networks. Models with different depths, such as DenseNet-121, DenseNet-169, and DenseNet-201, have been evaluated within this study.

VGG

VGG (Visual Geometry Group) is a deep learning architecture developed by (Simonyan & Zisserman, 2015). The key novelty of VGG is the use of small 3x3 convolution filters applied sequentially instead of large kernel filters. This approach allows for the creation of deeper and wider networks, resulting in better generalization and higher accuracy rates. Due to its simple and clear design, VGG is widely used in deep learning research and applications. There are different versions of VGG, such as VGG16 and VGG19, depending on the number of layers. In this study VGG16 architecture was used.

ResNetV2

ResNet architecture, first introduced by (He et al., 2015) and ResNetV2 is an improved version of ResNet architecture (He et al., 2016). The ResNet architecture uses residual connections to address the vanishing gradient problem encountered in training deep neural networks. These connections enable the network to reach deeper layers and perform better. ResNetV2 introduces improvements to this structure. One significant novelty is the application of batch normalization and activation functions (ReLU) before and after each residual block. Additionally, ResNetV2 employs a full pre-activation approach, allowing better gradient propagation and easier network training. These enhancements support the development of deeper and more effective neural networks, achieving superior performance in image recognition and other deep learning tasks. There are different versions of ResNet architecture. In this study ResNet50V2, ResNet101V2 and ResNet152V2 were evaluated.

MobileNetV2

MobileNetV2, developed by Google (Sandler et al., 2019). It is a deep learning architecture optimized for use on mobile and embedded devices that aims to deliver high performance with low latency and light computational requirements. It builds on MobileNetV1 and includes key enhancements such as inverted residual and linear bottleneck layers. These layers enhance network efficiency by minimizing information loss, with inverted residual structures expanding and then compressing feature maps to reduce computational costs, and linear bottleneck layers preserving the non-linear properties of activation functions. In addition, deep separable convolutions reduce the number of parameters and computational load, enabling efficient and accurate image classification, object detection and segmentation on mobile devices and embedded systems.

MobileNetV3

MobileNetV3, introduced by Google in 2019 (Howard et al., 2019), is a deep learning architecture optimized for mobile and embedded devices. The architecture is built on MobileNetV1 and MobileNetV2 to further enhance performance. Key innovations include the integration of SE (Squeeze-and-Excitation) blocks, the "hard-swish" activation function, and more efficient depth wise separable convolutions. MobileNetV3 has two main versions, MobileNetV3-Large and MobileNetV3-Small and both models were used in the study.

NasNet

NasNet (Neural Architecture Search Network) is an architecture developed by Google Brain (Zoph et al., 2018), designed to automate the creation of deep learning models. NasNet utilizes the NAS (Neural Architecture Search) algorithm to minimize human intervention in configuring deep neural networks. This algorithm explores and optimizes numerous potential network configurations to discover the best-performing architecture. The modular design of NasNet enhances computational efficiency and optimizes the number of parameters, making it both high-performing and flexible. NasNetMobile version was used in this study.

Xception

Xception (Extreme Inception) is a deep learning model developed by (Chollet, 2017), inspired by the Inception architecture. Xception primarily uses depth-separable convolutional layers to improve computational efficiency and model performance. Each convolution layer is divided into two steps: depth-wise convolution, which filters each input channel independently, followed by pointwise convolution, which combines all channels linearly. This approach significantly reduces the number of parameters and computational cost, while maintaining flexibility and learning capacity.

InceptionResNetV2

InceptionResNetV2 is a hybrid deep learning model that combines the Inception architecture with residual connections, introduced by (Szegedy et al., 2016). This architecture integrates the strengths of both Inception modules, which efficiently handle multi-scale features, and ResNet's residual connections, which mitigate the vanishing gradient problem in deep networks.

Vision Transformers

Transformers models are particularly known for their successes in natural language processing (NLP). However, in recent years, researchers have developed vision transformers (ViTs) to extend this success to the field of computer vision (Dosovitskiy et al., 2021). ViTs have managed to become an alternative to traditional CNN architectures with their success in computer vision problems. ViTs use Transformer architecture for image classification and other vision tasks. They divide images into small patches and process these patches as sequences to learn the necessary features for classification. ViTs consist of four main components: Patch Embedding, Positional Encoding, Transformer Blocks, and Classification Head. Image Patching (Patch Embedding): The input image is divided into fixed-sized patches. For example, a 224x224 image can be divided into 16x16 patches, resulting in 14x14 = 196 patches. Each patch is then flattened and transformed into a vector of a certain dimension using a linear layer. Positional Encoding: Positional information of the elements in the sequence is required. Therefore, positional encoding is added to each patch vector. Transformer Blocks: ViTs utilize classic Transformer blocks. Each block comprises multi-head self-attention mechanisms and feed-forward neural networks. These blocks learn relationships between image patches and extract important features. Multi-layer attention mechanisms effectively capture both global and local features of the image. Classification

Head: The output of the Transformer blocks is passed through a classification layer (usually an MLP - Multi-Layer Perceptron) for final classification. Two ViT models (ViT-16, ViT-32) pre-trained on the ImageNet dataset were used in the study. The classifier layers of the models were removed, and an eight-dimensional output layer with softmax activation function was used. Input images with a resolution of 224 x 224 x 3 were utilized. For the ViT-16 model, the patch size was set to 16 x 16, while for the ViT-32 model, the patch size was set to 32 x 32. The optimizer employed was Stochastic Gradient Descent, with a learning rate of 0.001. The batch size was configured to 32, the epoch size to 50, and the loss function used was categorical cross-entropy. A visual representation of ViTs processing is given in Figure 3.

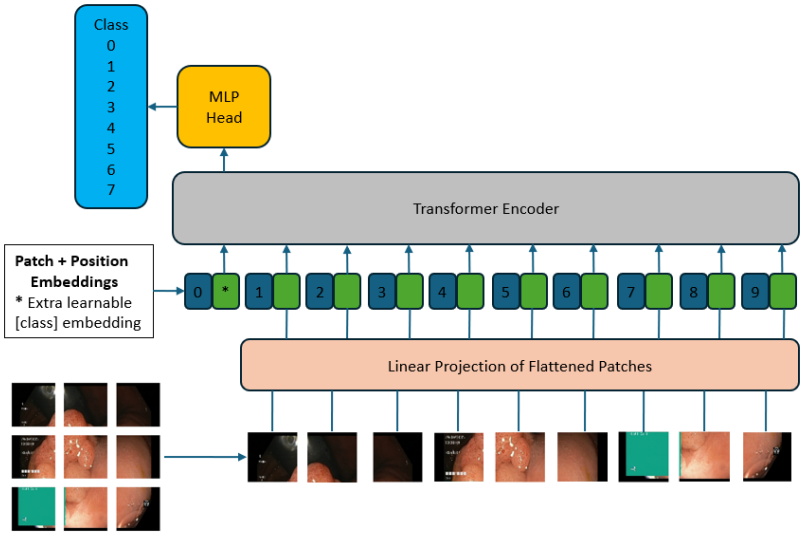

Figure 3. ViTs architecture for endoscopic image classification

Transfer Learning and Fine Tuning

Transfer learning is the process of using the features learned by a model in one task to solve a similar task. This method is particularly useful for problems with insufficient training data (Ribani & Marengoni, 2019). In deep learning models, transfer learning often involves adapting pre-trained models on large and diverse datasets to specific and smaller datasets. In CNNs, the layers closer to the input tend to capture more general features such as edges, corners, shapes, and colors, while layers closer to the output learn more task-specific features (Krizhevsky et al., 2012). Using CNNs for transfer learning in other problem domains is a widely preferred methodology. However, determining the depth, width, and the number of layers to be fine-tuned in the fully connected layers at the output requires expertise. In this study, all models used were trained with the ImageNet dataset. Fully connected layers were not added to any models except for the output layer, utilizing the weights of the pre-trained models. Additionally, no freezing operation was performed on the convolutional layers of the models. Among these models trained in this manner, the most successful model was selected, and fully connected layers were added, with various freezing rates applied to the convolutional layers to analyze classification performance. The depth, width, and freezing rates of the convolutional layers were determined using a random search algorithm. Fully connected layers were added to the DenseNet-201 model, which yielded the most successful results among pre-trained CNN models, with configurations of (1920-8), (1920-256-8), and (1920-256-256-8). After adding these layers, classification performance was evaluated by applying freezing rates of 0%, 25%, 50%, 75%, and 100% to the convolutional layers of the model.

Experimental Environment and CNN Model Hyperparameters

The study's experiments conducted on a system running the Ubuntu operating system, featuring 32 GB of RAM, and a 1080Ti graphics card. Training of the CNN models was conducted using the Keras deep learning library. Table 2 outlines the hyperparameters applied throughout the model training process.

Table 2. Hyperparameters of CNN models

Hyperparameter	Value
Input Size	(224x224x3) - (299x299x3)
Epochs	50
Loss Function	Categorical Cross Entropy
Batch Size	32
Learning Rate	0.001
Output Activation	Softmax
Optimizer	Stochastic Gradient Descent

Evaluation Criteria

The classification performance of the models was assessed using accuracy, sensitivity, precision, and F1 score metrics. These metrics were calculated on a per-class basis, and the final results were reported as the averages of these values. The calculations were carried out by using confusion matrix and formulas in Figure 4.

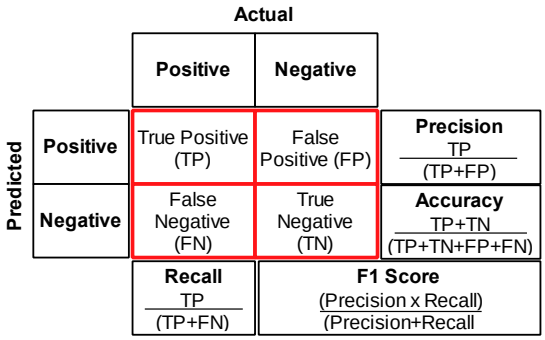

Figure 4. Confusion matrix and metric formulas

RESULTS AND DISCUSSION

All models trained in the study were evaluated using an external test dataset. No data augmentation was applied to the test dataset. The evaluation results are presented in Table 3.

Table 3. Average classification performances of models

Model	Accuracy	Precision	Recall	F1-Score
VGG-16	90.75	90.89	90.75	90.76
ResNet50V2	91	91.18	91	90.96
ResNet101V2	91.25	91.38	91.25	91.22
ResNet152V2	90.87	90.93	90.87	90.88
InceptionResNetV2	91.75	91.98	91.75	91.71
MobileNetV2	90.62	91.57	90.62	90.48
MobileNetV3Large	88.62	89.45	88.62	88.53
MobileNetV3Small	87.62	89.11	87.62	87.38
NasNetMobile	90.12	90.95	90.12	89.94
Xception	90.12	90.54	90.13	90.03
DenseNet121	91.87	92.02	91.87	91.87
DenseNet169	92.25	92.40	92.25	92.23
DenseNet201	92.75	92.75	92.75	92.74
ViT-16	91.25	91.34	91.25	91.24
ViT-32	90.50	91.01	90.50	90.44

According to Table 3, the DenseNet201 architecture achieved the highest accuracy rate of 92.75%. Therefore, the impact of transfer learning and fine-tuning parameters on the classification performance of the DenseNet201 model is presented in Table 4. Additionally, the confusion matrix for the DensNet201 and ViT-16 model are shared in Figure 5. Table 5 shows the best fine-tuned CNN model DenseNet201's class-based classification performance.

Table 4. Acc-1, Pre-1, Rec-1, F1-1 Indicates That after global average pooling two fully connected layers with 256,256 neurons, acc-2, pre-2, rec-2, f1-2 indicates that after global average pooling one fully connected layers with 256 neurons, acc-3, pre-3, rec-3, f1-3 indicates that after global average pooling only classification layer with 8 neurons

Froze Rate	Acc-1	Acc-2	Acc-3	Pre-1	Pre-2	Pre-3	Rec-1	Rec-2	Rec-3	F1-1	F1-2	F1-3
0%	92.63	91.87	92.75	92.66	91.90	92.75	92.63	91.87	92.75	92.63	91.87	92.74
25%	92.13	91.63	92	92.22	91.77	91.97	92.12	91.63	92	92.10	91.60	91.97
50%	92.13	**93.13**	92.25	92.42	**93.17**	92.29	92.12	**93.13**	92.25	92.09	**93.11**	92.25
75%	90.87	90.25	90	90.97	90.34	89.99	90.87	90.25	90	90.87	90.25	89.98
100%	87.75	91.87	87.75	88.84	86.94	87.84	87.75	86.75	87.75	87.71	86.75	87.71

Acc: Accuracy, Precision, Recall: Rec, F1 Score: F1

Table 5. Fine tuned DenseNet201 model class based average classification performance

Class Name-(Id)	Precision	Recall	F1-Score
Dyed-lifted-polyps (0)	93.14	95	94.06
Dyed-resection-margins (1)	94.90	93	93.94
Esophagitis (2)	88.27	82	84.97
Normal cecum (3)	96.04	97	96.52
Normal pylorus (4)	98.04	1	99.01
Normal-z-line (5)	83.18	89	85.99
Polyps (6)	95.88	93	94.42
Ulcerative colitis (7)	96	96	96
Average	**93.17**	**93.13**	**93.11**

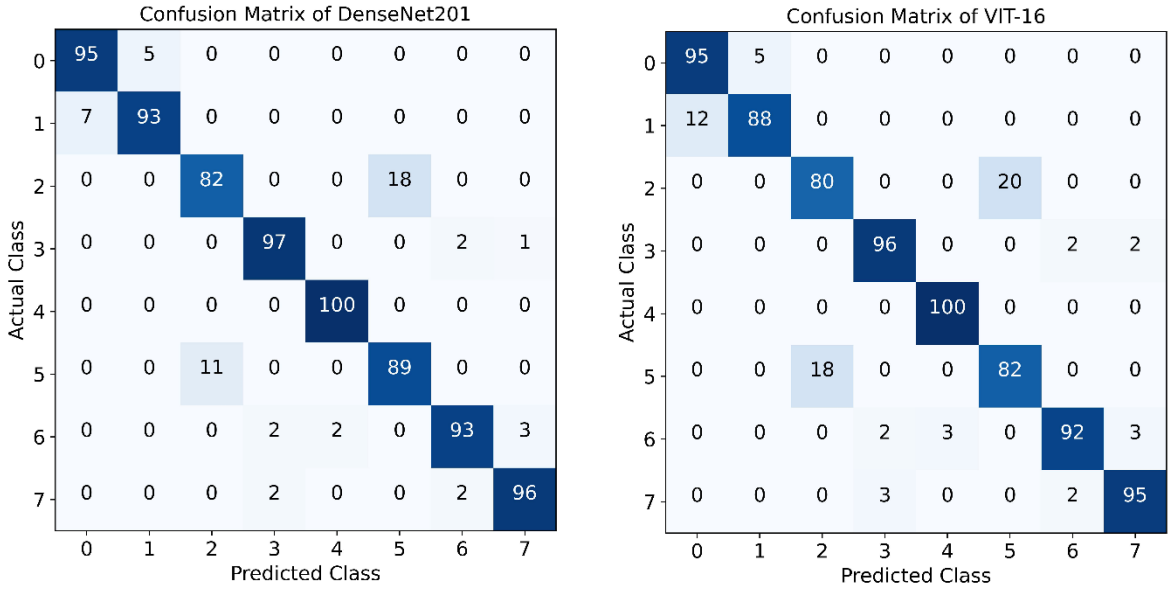

Figure 5. Confusion matrices of DenseNet201 and ViT-16

In this study, thirteen different CNN models (VGG-16, ResNet50V2, ResNet101V2, ResNet152V2, InceptionResNetV2, MobileNetV2, MobileNetV3Large, MobileNetV3Small, NasNetMobile, Xception, DenseNet121, DenseNet169, DenseNet201) and two ViT models (ViT-16, ViT-32) were trained using transfer learning to classify endoscopic images. Based on the evaluation of the test results, the fine-tuned DenseNet201 architecture achieved the highest performance among the

models with 93.13% accuracy, 93.17% precision, 93.13% recall, and a 93.11% F1 score. On the other hand, the MobileNetV3Small model was observed to have the lowest classification performance. It is believed that the small number of parameters of this model negatively affected its classification performance. The impact of freezing layers in the convolutional layers and the number of fully connected layers on the classification performance of the most successful DenseNet201 model is shown in Table 4. For this dataset, it was observed that the model trained with a 50% freezing rate and a single fully connected layer consisting of 256 neurons performed better in all metrics compared to the model trained without freezing and without a fully connected layer. Although the ViT models did not surpass the CNN models in classification, they achieved a similar classification performance. Table 5 shows that the Esophagitis (2) and Normal-z-line (5) classes are the most challenging to detect. The confusion matrices in Figure 5 indicate that the high number of false negatives for these two classes is due to their similarity. Additionally, the confusion matrices show that the model also confused the Dyed-lifted-polyps (0) and Dyed-resection-margins (1) classes. An image of a visual incorrectly predicted by the model and belonging to the predicted class is shared in Figure 6. Although ViTs are powerful models that could potentially replace CNNs, they require a large number of examples to be well-trained. In this study, it was observed that they were not as effective as CNNs on small datasets. Table 6 provides a comparison of the results obtained with studies in the literature using the same dataset.

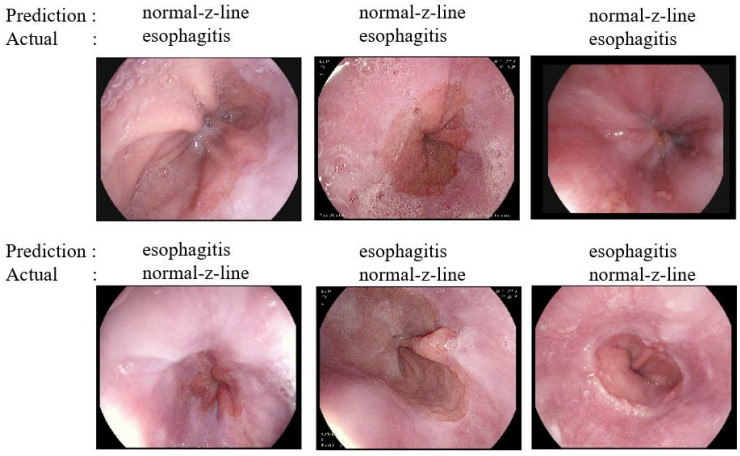

Figure 6. Misclassified images by the DenseNet201

Table 6. A comparision of the results obtained in this study with those reported in previous studies that have used the kvasir dataset

Number	Study	Accuracy	Precision	Recall	F1-Score
1	Yogapriya et al (2021)	96.33	96.50	96.37	96.50
2	Losenko et al. (2021)	93.19	92.8	92.7	92.8
3	Gjestang et al. (2021)	89.3	89	89.3	88.6
4	Mukhtorov et al. (2023)	93.46	-	-	-
5	Huo et al. (2024)	92.87	93.01	92.87	92.88
6	Demirbaş et al. (2024)	93.37	93.66	93.37	93.42
7	Here	**93.13**	**93.17**	**93.13**	**93.11**

As indicated in Table 6, the proposed method yielded superior outcomes in terms of accuracy compared to studies 3 and 5. In terms of recall, the study performed better than studies 2, 3, and 5, but lagged behind studies 1 and 6. Regarding precision, the proposed method outperformed studies 2, 3, and 5, but was inferior to studies 1 and 6. The proposed method also achieved better F1 scores than studies 2, 3, and 5. In this context, the proposed method demonstrated promising performance. The test results indicate that simple transfer learning methods are still effective compared to complex and difficult-to-

train architectures. Only study 1 appeared to be more successful than the others in Table 6. However, in this study, the data was augmented before training, and the augmented data was divided into training, validation, and test sets. This indicates a data leakage problem. One of the limitations of the study is the small number of data samples. Although online data augmentation methods were used, the quantity and quality of the data have a positive impact on classification performance.

CONCLUSION

In this study, thirteen different CNN models and two ViT models were trained to classify endoscopic images into eight different classes. Among the models, the one with the best classification performance was fine-tuned, and the results were analyzed. The fine-tuned DenseNet201 model achieved 93.13% accuracy, 93.17% precision, 93.13% recall, and a 93.11% F1 score. According to the obtained results, the fine-tuned model outperformed the other models. Although the two ViT models trained in the study achieved classification performance close to that of the CNN models, they did not yield better results. Future work plans to explore methods to enhance the effectiveness of ViTs on small datasets.

Conflict of Interest

The article authors declare that there is no conflict of interest between them.

REFERENCES

Agrawal, T., Gupta, R., & Narayanan, S. (2019). On evaluating CNN representations for low resource medical image classification. *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 1363–1367.

Chai, J., Zeng, H., Li, A., & Ngai, E. W. T. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6, 100134.

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *arXiv preprint arXiv:1610.02357*.

Demirbaş, A. A., Üzen, H., & Fırat, H. (2024). Spatial-attention ConvMixer architecture for classification and detection of gastrointestinal diseases using the Kvasir dataset. *Health Information Science and Systems*, 12(1), 32.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale *arXiv preprint arXiv:2010.11929.*

Gjestang, H. L., Hicks, S. A., Thambawita, V., Halvorsen, P., & Riegler, M. A. (2021). A self-learning teacher-student framework for gastrointestinal image classification. *IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 539–544.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385.*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity Mappings in Deep Residual Networks. *arXiv preprint arXiv:1603.05027.*

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3 (arXiv:1905.02244).

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). Densely Connected Convolutional Networks *(arXiv:1608.06993).*

Huo, X., Tian, S., Yang, Y., Yu, L., Zhang, W., & Li, A. (2024). SPA: Self-Peripheral-Attention for central–peripheral interactions in endoscopic image classification and segmentation. *Expert Systems with Applications*, 245, 123053.

Karaman, A., Karaboga, D., Pacal, I., Akay, B., Basturk, A., Nalbantoglu, U., Coskun, S., & Sahin, O. (2023). Hyper-parameter optimization of deep learning architectures using artificial bee colony (ABC) algorithm for high performance real-time automatic colorectal cancer (CRC) polyp detection. *Applied Intelligence*, 53(12), 15603–15620.

Katai, H., Ishikawa, T., Akazawa, K., Isobe, Y., Miyashiro, I., Oda, I., Tsujitani, S., Ono, H., Tanabe, S., Fukagawa, T., Nunobe, S., Kakeji, Y., & Nashimoto, A. (2018). Five-year survival analysis of surgically resected gastric cancer cases in Japan: A retrospective analysis of more than 100,000 patients from the nationwide registry of the Japanese Gastric Cancer Association (2001–2007). *Gastric Cancer*, 21(1), 144–154.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25.

Leufkens, A., Van Oijen, M., Vleggaar, F., & Siersema, P. (2012). Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy*, 44(05), 470–475.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12), 6999–7019.

Lonseko, Z. M., Adjei, P. E., Du, W., Luo, C., Hu, D., Zhu, L., Gan, T., & Rao, N. (2021). Gastrointestinal disease classification in endoscopic images using attention-guided convolutional neural networks. *Applied Sciences*, 11(23), 11136.

Mukhtorov, D., Rakhmonova, M., Muksimova, S., & Cho, Y.-I. (2023). Endoscopic image classification based on explainable deep learning. *Sensors*, 23(6), 3176.

Pacal, I. (2024). Improved Vision Transformer with Lion Optimizer for Lung Diseases Detection. *International Journal of Engineering Research and Development*, 16(2), 760-776.

Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., De Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.-T., Lux, M., Schmidt, P. T., Riegler, M., & Halvorsen, P. (2017). KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. *Proceedings of the 8th ACM on Multimedia Systems Conference*, 164–169.

Ribani, R., & Marengoni, M. (2019). A survey of transfer learning for convolutional neural networks. *SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, 47–57.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2019). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv preprint arXiv:1801.04381*.

Sermet, F., & Pacal, I. (2024). Deep learning approaches for autonomous crack detection in concrete wall, brick deck and pavement. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, 15(2), 503-513.

Siddiqui, S., Akram, T., Ashraf, I., Raza, M., Khan, M. A., & Damaševičius, R. (2024). CG-Net: A novel CNN framework for gastrointestinal tract diseases classification. *International Journal of Imaging Systems and Technology*, 34(3), e23081.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

Sivari, E., Bostanci, E., Guzel, M. S., Acici, K., Asuroglu, T., & Ercelebi Ayyildiz, T. (2023). A new approach for gastrointestinal tract findings detection and classification: Deep learning-based hybrid stacking ensemble models. *Diagnostics*, 13(4), 720.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv preprint arXiv:1602.07261; Version 2*.

Yogapriya, J., Chandran, V., Sumithra, M. G., Anitha, P., Jenopaul, P., & Suresh Gnana Dhas, C. (2021). Gastrointestinal Tract Disease Classification from Wireless Endoscopy Images Using Pretrained Deep Learning Model. *Computational and Mathematical Methods in Medicine*, 2021, 1–12.

Zhang, X., Chen, F., Yu, T., An, J., Huang, Z., Liu, J., Hu, W., Wang, L., Duan, H., & Si, J. (2019). Real-time gastric polyp detection using convolutional neural networks. *PloS One*, 14(3), e0214133.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning Transferable Architectures for Scalable Image Recognition. *arXiv preprint arXiv:1707.07012*.