Accurate Deep Learning-based Gastrointestinal Disease Classification via Transfer Learning Strategy

Jessica Escobar*, Karen Sanchez†, Carlos Hinojosa*, Henry Arguello*‡, Sergio Castillo*
*Department of Computer Science. Universidad Industrial de Santander, Bucaramanga, Colombia.
†Department of Electrical Engineering. Universidad Industrial de Santander, Bucaramanga, Colombia.

2021 XXIII Symposium on Image, Signal Processing and Artificial Vision (STSIVA) | 978-1-6654-1668-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/STSIVA53688.2021.9591995

***Abstract*—The automatic detection of diseases and gastrointestinal tract anomalies is a challenge for medical experts, affecting patient treatment decisions. Deep learning emerges as a new tool in interpreting medical images for the diagnosis, disease prediction, and clinical treatment analysis. Several works proposed in the literature rely on convolutional neural networks to classify medical images accurately. However, the state-of-the-art methods for gastrointestinal anomalies classification have complex architectures, which require multiple parameters to be trained. Then, there is a scope for developing a light and easily replicable deep-learning-based method that maintains the high precision of more complex models. This work proposes a workflow to classify diseases and anomalies of the gastrointestinal tract using image processing and a transfer learning strategy. Our proposed method is tested on the Kvasir-V2 dataset, containing 8000 endoscopic images divided into eight classes. Our proposed approach achieves more than 98% accuracy during testing by only using the fifth part of trainable parameters compared to the state-of-the-art methods we compare our approach in the experiments.**

***Index Terms*—Classification, Convolutional neural network, Data augmentation, Gastrointestinal Disease, Transfer learning.**

I. Introduction

The gastrointestinal tract is responsible of decompose and absorb the food that a person eats. This tract is composed of the stomach, small and large intestine. Usually, gastrointestinal diseases are studied using an endoscopic camera that captures real-time images of the internal gastrointestinal cavities allowing the medical experts to make a diagnosis. According to the World Health Organization, diseases of the gastrointestinal tract are among the top ten causes of death; therefore, the research concerning them is crucial to reduce their mortality [1]. Specifically, diarrheal diseases caused by various bacterial, viral, or parasitic organisms are ranked eight. In turn, diarrheal diseases are the second leading cause of death for children under the age of five, killing 525,000 children each year [2]. Gastrointestinal diseases include gallstones, polyps, ulcerative colitis, celiac disease, diverticulitis, and cancer, which may or may not be treatable depending on the severity. An incorrect diagnosis of the disease and its level of severity could cost the patient his life. For this reason, it is essential to implement computer-aided diagnosis (CAD) systems [3] that support the detection of these diseases in endoscopic images. This would bring multiple advantages, such as improving the quality, precision, and time of the diagnosis. In addition, this can be a tool that provides an opinion to the health professional in interpreting the images. It is common for the diagnosis to vary from one medical expert to another, especially when their training comes from different geographical locations since the profile of these diseases can vary due to demographic conditions. Additionally, in some cases, it is difficult to accurately identify the level of severity of the disease. Therefore, it is crucial to make use of new computational technologies [4]–[7] that can aid diagnosis, minimize human error, optimize the time and productivity of health professionals.

Related Works. In recent years, different works relying on machine learning [8], deep learning [9]–[13], and more particular, convolutional neural networks (CNNs) [14], [15], have been proposed. However, although many of these methods exhibit high performance in the analysis of medical images, this success depends mainly on the amount of data. For this reason, the transfer learning technique [16] in the medical field is beneficial since the model can be adjusted with previously trained models from natural images such as the ImageNet dataset [17] to correctly transfer the classification task to a medical domain. For instance, in [18], the authors propose the classification of the Kvasir V1 dataset using three approaches, using global features, deep learning in CNNs, and transfer learning in deep learning. The best result was obtained training from scratch a three-layer CNN, with an accuracy of 95.9%. In [19], they performed a pre-processing of edge removal, contrast enhancement, filtering, color mapping, and scaling to each image in the Kvasir-V2 dataset; besides, they used the data augmentation technique. These images were used to train and test three CNNs: Inception-v4, Inception-ResNet-v2, and NASNet, obtaining the best result with the CNN of Inception-ResNet-v2 with an accuracy of 98.48% accuracy.

On the other hand, taking advantage of the transfer learning technique, in [20], the authors implemented a transfer learning technique with fine-tuning on two deep CNNs: ResNet50 and DenseNet121, pre-trained with the ImageNet dataset. Then, they classified the Kvasir V1 dataset, resulting in an accuracy of 87.8% in the residual network and 86.9% in the dense model. An approach of combining features extracted by several CNNs was addressed in [21], they proposed to classify the Kvasir-V2 dataset using a set of six CNNs: DenseNet-201, ResNet-18, VGG-16, InceptionV3, Xception, InceptionResnetV2, with a global average pooling layer to obtain feature

978-1-6654-1668-9/21/$31.00 ©2021 IEEE

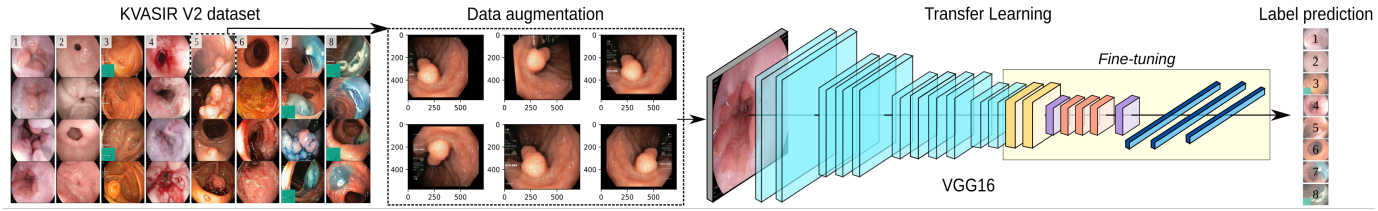

Fig. 1. The workflow of the proposed approach to detecting gastrointestinal anomalies and diseases from endoscopic images using CNN and transfer learning.

vectors. Then, they obtained the final feature vector for the classification task by adding the vectors generated in each CNN. Finally, they feed a single layer of decision that allows them to obtain an accuracy of 97.38%.

Paper Contribution. This work addresses the Kvasir-V2 dataset classification problem using pre-processing, transfer learning, and hyperparameter fitting. Due to the lack of labeled medical images, we first use data augmentation with geometric transformations to generate new images. Then, we use the VGG-16 convolutional neural network with previously trained weights for the Imagenet dataset to extract features. Instead of retraining all the architecture from scratch, we investigate different fine-tuning settings that lead us to achieve the best accuracy with significantly fewer parameters. We found that performing a fine-tuning of the weights from the convolutional layer 3 in block 4 of the VGG-16 architecture to the fully connected layer lead to the best results in terms of accuracy. Our proposed framework provides high classification precision from light and a computationally inexpensive neural model. In this way, we can show that it is unnecessary to readjust all the pre-trained weights of a neural model for a long retraining period to achieve very high performance in classifying endoscopic images in eight classes of diseases or anomalies of the gastrointestinal tract. It should be noted that this work aims to contribute to the study of tools that support, but do not replace, the medical evaluation of endoscopic images in terms of their classification and severity.

II. Proposed Method

In this section, we described our proposed approach for diseases and anomalies classification of the GI tract. We separate the steps and explain them in the following subsections.

A. Image Preprocessing

In the training of CNNs, it is essential to have a large amount of data available to obtain a good classification performance and avoid over-fitting. Therefore, in this work, data augmentation is applied to increase the size of the training set by performing different geometric transformations. In our experiments, we first divided the dataset randomly into three subsets: 80 %, 15 %, and 5 % for training, testing, and validation, respectively. Second, we perform data augmentation to the training subset through six transformations: flip horizontal, vertical, displacement width, height, rotation, and zoom. Finally, we resize each training image to 400×400 pixels before feeding the images to the CNN.

B. Transfer Learning Strategy

One of the main challenges of employing deep learning models in the medical area is the lack of training data [22]. Furthermore, computationally expensive models are more challenging to scale. For these reasons, transfer learning is advantageous in the medical field. From a CNN model previously trained with natural image data sets, the classification task can be efficiently transferred to a domain with medical images avoiding the expensive training from the scratch process.

Consequently, we propose to use the transfer learning technique by comparing five different CNN models to choose the best-performing network. Also, we use fine-tuning to re-train only some layers of the neural networks using the preprocessed training data in II.A. For comparison purposes, we retrain the layers of the last convolutional block for each of the five models and keep all the other weights from their ImageNet pretraining unaltered. We replace the previous layer of the five networks with a new layer of eight units representing the number of classes in the Kvasir dataset. Figure 1 shows the pipeline of the proposed approach to detecting gastrointestinal anomalies and diseases from endoscopic images using CNNs via transfer learning strategy.

III. Results

This section presents the numerical results of the proposed framework for gastrointestinal diseases classification through endoscopic images using the Kvasir-V2 dataset. We trained the proposed model using the *Adam* optimizer with a batch size of 32 and an initial learning rate of 1×10^{-4}. We applied an inverse time-decay learning rate schedule with a decay factor of 1×10^{-5}/epochs that was triggered every epoch. We trained the network for 15 epochs, which took about 45 minutes on an Nvidia Tesla T4 GPU.

A. Dataset

The Kvasir-V2 dataset [18] is composed of endoscopic images from inside the gastrointestinal tract. This version of the dataset consists of 8,000 images grouped into 8 different classes (i.e, 1000 per class) which have been annotated and verified by experienced endoscopists. The classes are based on three anatomical landmarks (z-line, pylorus, cecum), three pathological findings (esophagitis, polyps, ulcerative colitis), and two other classes related to the removal of polyps (dyed and lifted polyps, dyed resection margins) as shown in Fig. 2.

TABLE I
EVALUATION METRICS.

Metric	Accuracy	Precision	Recall	Specificity	F1-Score	Matthews Correlation Coefficient
Formula	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{TN}{TN+FP}$	$\frac{2TP}{2TP+FP+FN}$	$\frac{TPxTN-FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

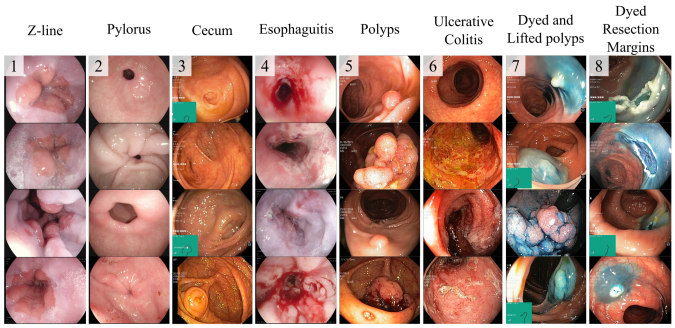

Fig. 2. Four random samples from each class in the Kvasir-V2 dataset.

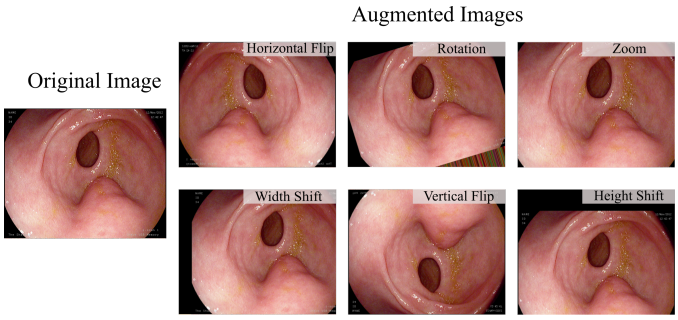

Fig. 3. Geometric transformations in data augmentation.

B. Preprocessing

We perform data augmentation on the Kvasir-V2 dataset using the geometrical transformations exposed in Section II.A. (see Fig. 3). Then, the training images were doubled with the data augmentation, from 800 to 1600 training images for each class, obtaining a total of 12800 training images, considering that there are 8 classes and 1600 training images in each class as seen in Table II. Note that the number of images for testing and validation was not changed.

C. Evaluation Metrics

To evaluate the performance of our classification approach, we use the following metrics: accuracy (ACC), precision (PRE), recall (REC), specificity (SPEC), F1 score (F1), Matthew correlation coefficient (MCC), and area under the ROC curve (AUC). The formulas to compute these values are shown in Table I; their variables are extracted from the confusion matrix results where true positive (TP) and true negative (TN) denote the number of positive and negative samples correctly classified. Meanwhile, false positive (FP) and false negative (FN) denote the number of wrongly classified positive and negative samples. Regarding the selected metrics, *accuracy* measures the ratio of correct predictions over the total number of samples evaluated. *Precision* measures samples correctly identified as positive between the total identified positive samples. *Recall* is the ratio of samples correctly identified as positive among all existing positive samples. *Specificity* is the ratio of correctly identified negative samples. *F1 score* represents the harmonic mean of the precision and recall. *Matthews correlation coefficient* shows a high result to the extent that the prediction performed well in all four variables. Finally, *AUC* assesses the ability of the model to distinguish between classes.

TABLE II
IMAGES DISTRIBUTION BEFORE AND AFTER AUGMENTATIONS.

	Before		After	
	Images per class	Total images	Images per class	Total images
Train	800	6400	1600	12800
Test	150	1200	150	1200
Validation	50	400	50	400
Total	1000	**8000**	1800	**14400**

TABLE III
ACCURACY FOR DIFFERENT TESTED CNN ARCHITECTURES.

Pretrained CNNs	Accuracy
DenseNet201	78.55
ResNet50	90.42
Xception	78.26
VGG19	97.64
VGG16	**98.20**

D. Ablation Studies

We empirically analyze the performance the proposed pipeline by comparing our method with the baselines based on the same experimental setting. First, we tested five CNNs pretrained with the ImageNet dataset: DenseNet201, ResNet50, Xception, VGG19, and VGG16. The results obtained are shown in Table III. We reported the *accuracy* score for each experiment and found the best result using the VGG16 network. Therefore, we selected this network to continue with our study.

Once the network was defined, we tested the fine-tuning in the different convolutional blocks of the VGG16 network. As shown in Table IV, the best performance was obtained by training from "block4_conv1". Next, we performed the tests described in Table V, where we apply the fine-tuning from each of the three internal layers of block 4. We

TABLE IV
ACCURACY FOR DIFFERENT VGG16 BLOCKS.

Block with Fine-tuning	Trainable Params	Non Trainable Params	Computacional Time [s]	Accuracy
block1_conv1	14,714,688	0	5849.254	0.9764
block2_conv1	14,675,968	38,720	4395.321	0.9781
block3_conv1	14,454,528	260,160	2253.218	0.9792
block4_conv1	12,979,200	1,735,488	2877.035	**0.9808**
block5_conv1	7,079,424	7,635,264	2363.919	0.9803

TABLE V
ACCURACY FOR THE DIFFERENT LAYERS OF BLOCK 4.

Layer with Fine-tuning	Trainable Params	Non Trainable Params	Computacional Time [s]	Accuracy
block4_conv1	12,979,200	1,735,488	2877.035	0.9808
block4_conv2	11,799,040	2,915,648	2607.204	**0.9820**
block4_conv3	9,439,232	5,275,456	2462.092	0.9791

observed that the best *accuracy* was obtained by training from "block4_conv2". Therefore, in our proposal, we use the VGG16 architecture previously trained with the ImageNet dataset, and the fine-tuning technique is applied by training from "block 4_conv2".

E. Quantitative Results

The results obtained for our proposed method are presented in the last row of Table VI, in which we calculated the metrics exposed in Section II.B: ACC, PREC, REC, SPEC, F1, and MCC, from the TP, TN, FP, and FN cases, which are reported in the confusion matrix of Fig. 4. The AUC for each class, micro, and macro-average of the proposed framework are reported in the ROC curves conventions in Fig. 5. Otherwise, Fig. 6 shows the accuracy and loss curves of our model along the training epochs. These curves show a training model that converges smoothly and steadily towards weights that allow high precision and low error in both training and validation subsets. Besides, in Fig. 5, the AUC of this model is 0.99, which shows a high capacity to distinguish between the eight classes.

F. Comparisons

We compare the performance of our proposed method with existing related methods, which have been reported for endoscopic image classification, and described in Section I of this paper. As shown in Table VI, our method outperforms three of the baseline methods in terms of *accuracy*. The best results are shown in bold font, and the second-best is underlined. Although our method did not outperform the work in [19] by a difference of 0.28%, our method is significantly faster and less computationally complex. Specifically, the authors in [19] proposed a bilineal architecture that fuses extracted features with both Inception and ResNet networks. Such a method needs training over 55.8 million parameters to achieve the reported results. On the other hand, our method requires only 11.7 million trainable parameters to achieve very similar accuracy results.

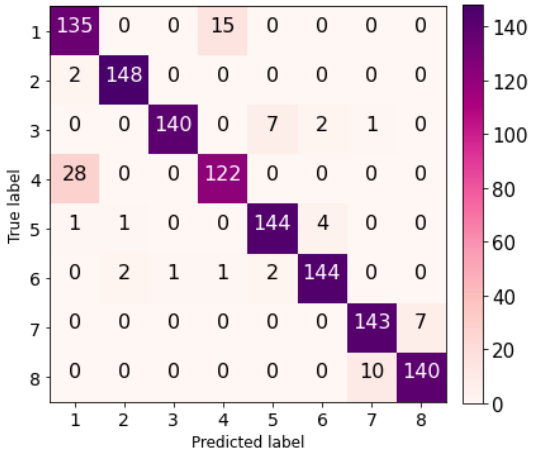

Fig. 4. Confusion matrix of the proposed classification approach for the Kvasir-V2 dataset.

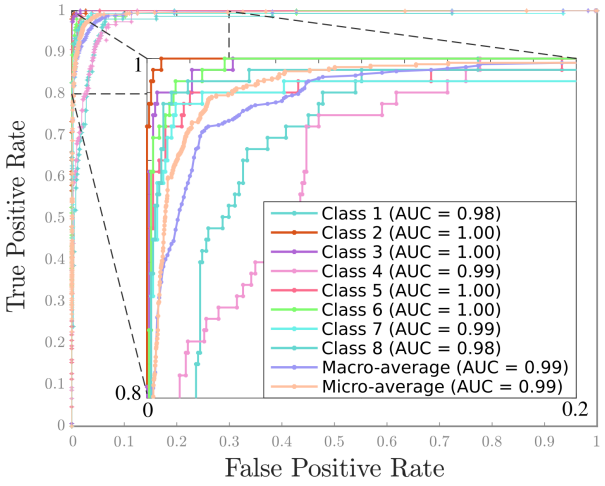

Fig. 5. AUC values and ROC curves for each class in the Kvasir-V2 dataset using the proposed classification framework.

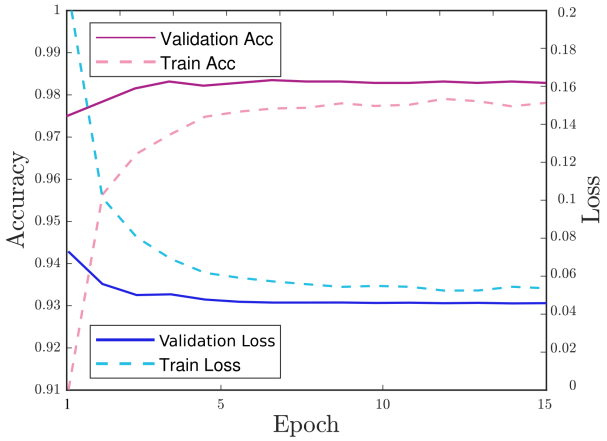

Fig. 6. Epochs vs. accuracy and loss classification curves

IV. CONCLUSIONS

In this work, we studied convolutional neural networks and transferred learning to classify gastrointestinal endoscopic

TABLE VI
APPLICATIONS OF MACHINE LEARNING IN GI TRACT ANALYSIS RESEARCH USING THE KVASIR-V2 DATASET.

Year	Method	Dataset Distribution			ACC	PREC	REC	SPEC	F1	MCC
		Train	Test	Validation						
2017	3 Layer CNN [18]	50%	50%	-	0.959	0.589	0.408	0.890	0.453	0.430
2019	Inception-ResNet-v2 [19]	85%	15%	-	**0.9848**	0.940	0.939	**0.991**	0.939	**0.930**
2019	ResNet50 CNN with Transfer Learning [20]	60%	30%	10%	0.878	-	-	-	-	-
2019	ANN with pre-trained CNN feature extractors [21]	80%	20%	-	0.9738	**0.9715**	**0.9727**	-	**0.9721**	-
2021	Proposed Method	80%	15%	5%	0.9820	0.9286	0.9275	0.99	0.9276	0.9173

images. We showed that a method based on pre-processing and fine-tuning transfer learning on a light convolutional neural network such as the VGG-16 could classify eight categories of diseases and abnormalities in the gastrointestinal tract with high precision. With experiments on the Kvasir-V2 dataset, we show that our proposed framework for endoscopic image processing and classification outperforms other state-of-the-art works. Furthermore, our method achieves an accuracy very similar to the most accurate work of the state-of-the-art, using only a fifth of the number of trainable parameters.

ACKNOWLEDGMENT

This work was supported by the Vicerrectoría de Investigación y Extensión from Universidad Industrial de Santander with the research project "**2707** - Procesamiento de imágenes y aprendizaje profundo como herramientas para la cobertura universal de la salud. Caso de estudio: seguimiento local y remoto del tratamiento de úlceras en extremidades inferiores en pacientes provenientes de comunidades rurales y municipios de Santander".

REFERENCES

[1] World Health Organization, "The top 10 causes of death," url: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2020.

[2] World Health Organization, "Diarrhoeal disease," url: https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease, 2017.

[3] Xuejiao Pang, Zijian Zhao, and Ying Weng, "The role and impact of deep learning methods in computer-aided diagnosis using gastrointestinal endoscopy," *Diagnostics*, vol. 11, no. 4, pp. 694, 2021.

[4] Mohamed Esmail Karar, Ezz El-Din Hemdan, and Marwa A Shouman, "Cascaded deep learning classifiers for computer-aided diagnosis of covid-19 and pneumonia diseases in x-ray scans," *Complex & Intelligent Systems*, vol. 7, no. 1, pp. 235–247, 2021.

[5] Soo-Yeon Kim, Yunhee Choi, Eun-Kyung Kim, Boo-Kyung Han, Jung Hyun Yoon, Ji Soo Choi, and Jung Min Chang, "Deep learning-based computer-aided diagnosis in screening breast ultrasound to reduce false-positive diagnoses," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[6] Peter Bossuyt, Gert De Hertogh, Tom Eelbode, Séverine Vermeire, and Raf Bisschops, "Computer-aided diagnosis with monochromatic light endoscopy for scoring histologic remission in ulcerative colitis," *Gastroenterology*, vol. 160, no. 1, pp. 23–25, 2021.

[7] Shuang Liang and Yu Gu, "Computer-aided diagnosis of alzheimer's disease through weak supervision deep learning framework with attention mechanism," *Sensors*, vol. 21, no. 1, pp. 220, 2021.

[8] Ibrahim Ibrahim and Adnan Abdulazeez, "The role of machine learning algorithms for diagnosing diseases," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 10–19, 2021.

[9] Francesco Piccialli, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino, "A survey on deep learning in medicine: Why, how and when?," *Information Fusion*, vol. 66, pp. 111–137, 2021.

[10] Jessica Paola Escobar, Natalia Gomez, Karen Sanchez, and Henry Arguello, "Transfer learning with convolutional neural network for gastrointestinal diseases detection using endoscopic images," in *2020 IEEE Colombian Conference on Applications of Computational Intelligence (IEEE ColCACI 2020)*. IEEE, 2020, pp. 1–6.

[11] Qingchen Zhang, Changchuan Bai, Zhikui Chen, Peng Li, Hang Yu, Shuo Wang, and He Gao, "Deep learning models for diagnosing spleen and stomach diseases in smart chinese medicine with cloud computing," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 7, pp. 1–1, 2021.

[12] Evgin Goceri, "Deep learning based classification of facial dermatological disorders," *Computers in Biology and Medicine*, vol. 128, pp. 104118, 2021.

[13] Karen Sanchez, Carlos Hinojosa, Henry Arguello, Simon Freiss, Nicolas Sans, Denis Kouamé, Olivier Meyrignac, and Adrian Basarab, "Subspace-based domain adaptation using similarity constraints for pneumonia diagnosis within a small chest x-ray image dataset," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 1232–1235.

[14] Hyun Yoo, Soyoung Han, and Kyungyong Chung, "Diagnosis support model of cardiomegaly based on cnn using resnet and explainable feature map," *IEEE Access*, vol. 9, pp. 55802–55813, 2021.

[15] Elizabeth Martinez, Camilo Calderón, Hans Garcia, and Henry Arguello, "Mri brain tumour segmentation using a cnn over a multi-parametric feature extraction," in *2020 IEEE Colombian Conference on Applications of Computational Intelligence (IEEE ColCACI 2020)*. IEEE, 2020, pp. 1–6.

[16] Yiting Wang, Shah Nazir, and Muhammad Shafiq, "An overview on analyzing deep learning and transfer learning approaches for health monitoring," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[18] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al., "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 164–169.

[19] Timothy Cogan, Maribeth Cogan, and Lakshman Tamil, "Mapgi: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning," *Computers in biology and medicine*, vol. 111, pp. 103351, 2019.

[20] Abel KahsayGebreslassie, Misgina Tsighe Hagos, et al., "Automated gastrointestinal disease recognition for endoscopic images," in *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. IEEE, 2019, pp. 312–316.

[21] Chathurika Gamage, Isuru Wijesinghe, Charith Chitraranjan, and Indika Perera, "Gi-net: Anomalies classification in gastrointestinal tract through endoscopic imagery with deep learning," in *2019 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2019, pp. 66–71.

[22] Laith Alzubaidi, Mohammed A Fadhel, Omran Al-Shamma, Jinglan Zhang, J Santamaría, Ye Duan, and Sameer R Oleiwi, "Towards a better understanding of transfer learning for medical imaging: A case study," *Applied Sciences*, vol. 10, no. 13, pp. 4523, 2020.