Gastrointestinal Disease Recognition Using Advanced Machine Learning Models

Zorawar Jaiswal
Department of Computer Science
California State University, Northridge
Northridge, CA, USA
zorawar.jaiswal.556@my.csun.edu

Abhishek Verma
Department of Computer Science
California State University, Northridge
Northridge, CA, USA
abhishek.verma@csun.edu

Abstract— **The Gastrointestinal tract (GI tract) often known as the digestive system is an organ system that is highly prone to cancers, polyps, and other diseases. Examination of this tract is not so straightforward and relies heavily on endoscopy. The real-time feed captured via endoscopy can last for hours and may have many pathological findings and anatomical landmarks. This research presents a deep learning based solution to classify images of the GI tract to highlight the endoscope reaching a landmark or a pathological region of interest. The experiments compare the prior benchmarks with CNN-based architectures (EfficientNet L2 and EfficientNet B7) and Transformer-based architecture (Hiera H). The experiment achieves an accuracy of 93.7%, which is an improvement upon the existing benchmarks. The research is important as it focuses on achieving high accuracy on a limited 2K size training set without overfitting.**

Keywords— Gastrointestinal Tract, Image Classification, Deep Learning, Transformers

I. Introduction

The gastrointestinal tract (GI tract) often referred to as the digestive system of the body is one of the most important organ systems of the human body. Pogorelov et al., labeled cancers of the GI tract as three-eighths of the most commonly occurring cancers [1]. A study based on the GLOBOCAN estimates from 2022 ranked cancers of the colon, stomach, and esophagus as the third, fifth, and eleventh most commonly occurring new cases, while they are the second, fifth, and seventh in terms of mortality [24].

Medical advancements over the years have led to a metamorphosis of gastrointestinal tract imaging and detection. Endoscopy began with wired optic-based systems and switched to capsule-based modules [8]. This, however, led to various challenges, such as the capsule missing the sight of concern or a lack of information about the direction of arrival. Various techniques were researched to mitigate these issues, such as estimating the location based on signal strength and introducing a magnet within the capsule that could be tracked with an external magnet [9]. Recent works focused on power efficiency and transmission speed [10].

Endoscopy images or frames set the background for our research. With a plethora of data available and to address emergencies in the medical profession, a need was felt to automate the observation of patients undergoing endoscopy with an event-driven framework that could notify the medical professionals about regions of interest both in terms of landmarks and pathological findings themselves. This paper bridges that gap by identifying the aforementioned regions based on the Kvasir dataset [1] using modern deep learning techniques.

The evolution of Deep Learning techniques from Convolutional Neural Networks (CNN) [11] to self-attention, Transformers [25] and Vision Transformers [16] have sparked innovation throughout the medical analysis domain [11].

This study aims to supplement GI tract-based disease and landmark detection to aid medical personnel. The study compares various CNN and Transformer-based deep learning architectures i.e. Hiera Huge, EfficientNet B7, and EfficientNet L2 with the top results on the Kvasir version 1 dataset.

The rest of this research is organized into sections exploring the literature review, dataset description, results and discussion, and conclusion.

II. Related Works

A. Deep Learning in Medical Diagnosis

Similar to the advancements in endoscopy, Deep Learning has found itself a plethora of applications within the medical domain. The application of deep learning ranges from tumor detection, skin lesion classification, anatomy classification, lung disorder inference from HRCT reports, ultrasound, and CT classification to the fabrication of additional training data [11, 12].

The application of Deep Learning has grown in both breadth and depth within the medical field [13]. Breadthwise, image classification and various types of image segmentation such as semantic segmentation (labeling a group of pixels), or labeling individual objects known as instance segmentation, or a combination of both these techniques known as panoptic segmentation [14]. Depth wise, the trend shifted from Convolutional Neural Networks (CNNs) [11, 13] to superior General Adversarial Networks (GANs) [15] to utilizing Transformers in Vision-related tasks [16]. Dosovitskiy et al., established the use of transformers for vision-related tasks. They claim that CNNs are powerful for smaller datasets since transformers lacked the inductive bias as CNNs gathered information using a kernel and neighboring pixels sharing information [17], hence, without larger datasets, the performance of transformers was inferior to traditional CNN models. This normalized the idea of training transformer models on larger datasets such as ImageNet and later fine-tuning the model to smaller datasets [15, 18]. Vision Transformer (ViT) models are superior to traditional CNN and Inception V3 models [19]. Armand et al. compared various UNet-based Transformer models across four datasets for medical image segmentation. They conclude that while individual architecture offers great accuracy, no clear winner could be established suggesting that the performance of architectures can depend on the use case i.e., the type of images encountered [20]. The trend finally shifted to creating hybrid models between Transformers and CNN architectures, leading to a boost in accuracy [17].

B. Research on the Kvasir Dataset

Most of the work on Kvasir has been carried out on the second variant. Kvasir V1, however, still stands out to be an important dataset since it compels one to work with limited data. Research on Kvasir follows the general trend of Deep Learning for Medical Diagnosis. Agrawal, Gupta, and Narayanan individually benchmarked VGGNet, ResNet50, Inception-V3, XceptionNet, and MobileNet, and implemented the feature fusion technique with final classification via Support Vector Machine (SVM) to achieve 83.8% on the Kvasir V2 [21].

Preprocessing techniques such as Canny Edge Detection were included along with feature fusion by Gupta et al., [22]. They extracted features from ResNet50 and EfficientNet B7 and performed max voting on various machine learning algorithms, including but not limited to Random Forest, Passive Aggressive Classifier, Ridge Classifier, and Logistic Regression to achieve above 88% accuracy while State Space Model (SSM) based Mamba architectures outperformed EfficientNet V2, Swin, and NextViT architectures with similar accuracy [23].

On the Kvasir dataset Huo et al., proposed HiFuse that extracted both local and global features, combining the advantages of CNN-based and Transformer-based models using a Hierarchical Feature Fusion module. They were able to achieve 85% accuracy on the V1 dataset at a 50/50 split with lower complexity [2]. Öztürk & Özkaya on the other hand, relied on feature fusion at the first, the middle, and the final pooling layer to capture the fundamental image information (the initial layers), and the semantic information (final layers) for better feature quality and pass these features into different layers of an LSTM. They compare AlexNet, GoogleNet and ResNet 50 with final classifiers based on either SVM or Artificial Neural Network (ANN). They were able to achieve a peak accuracy of 93.01% on the 2K training image variant [7]. The next section describes the Kvasir dataset in detail.

III. Dataset

The Kvasir Dataset was introduced by Pogorelov et al., [1] in 2017. This dataset comprises of eight classes and is available in two variants (4K images and 8K images). The Kvasir dataset is a balanced dataset vetted, analyzed, and annotated by medical specialists [1]. Fig 1 shows some sample data and the labels in Kvasir V1.

The dataset comprises anatomical landmarks, i.e., main features in the GI tract, namely Z-line, Pylorus, and Cecum. It also consists of pathological findings, i.e., abnormalities within the GI tract – Esophagitis, Polyps, and Ulcerative Colitis. Finally, the dataset contains Dyed and Lifted Polyps, and Dyed Resection Margins since they become regions of interest for various endoscopic treatments. There is an overlap between the images of Kvasir V1 and V2, hence, this research considers only Kvasir V1 – the original dataset for experiments with 4K images.

The dataset was split into a 50-50 training and testing split for direct comparison with the HiFuse Model [2].

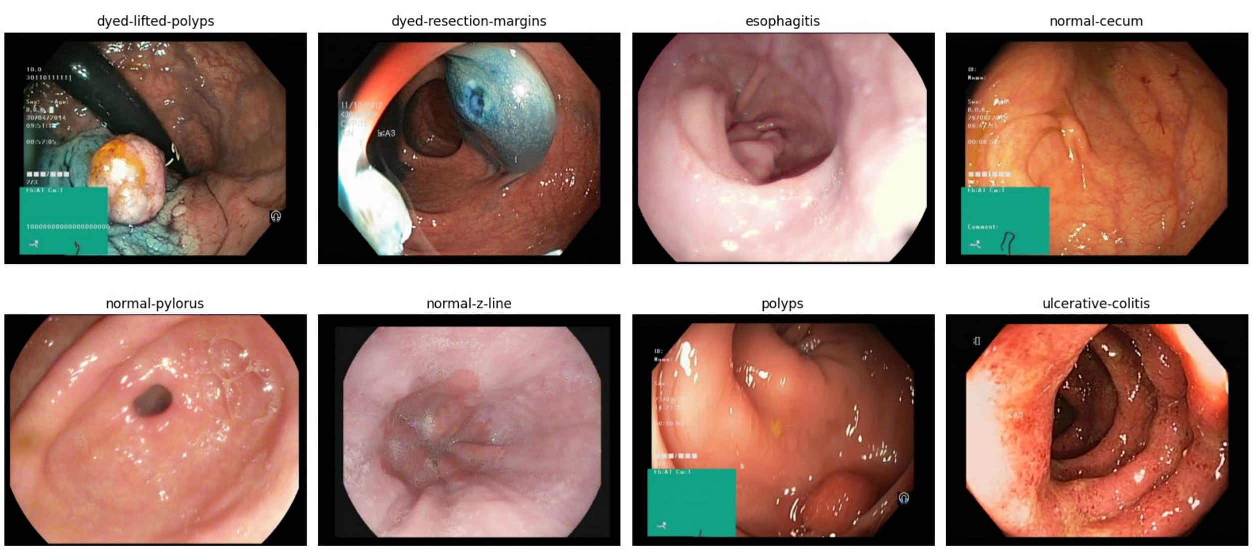

Fig 1. Sample data in Kvasir V1 and V2 [1].

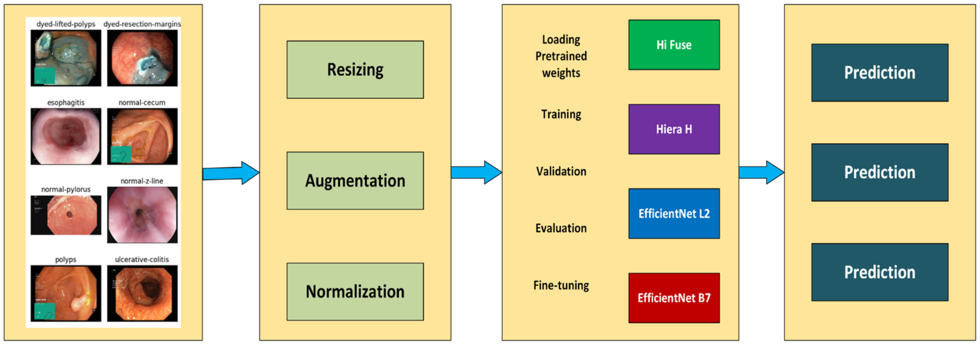

Fig 2: System pipeline of research methodology.

IV. Research Methodology

The Kvasir dataset [1] by default has images ranging in size from 400 to 700 Kb. The aspect ratio also varies across images and classes. The entire system pipeline is depicted in Fig 2 highlighting the preprocessing incorporated. The training split has preprocessing steps comprising of resizing the image, augmenting it, and normalizing the image. The exact steps of preprocessing for the training split included: (i) Randomly resizing and cropping a 224x224 pixel area. (ii) This is followed by a random horizontal flip with a 50% probability. (iii) These images are then converted to tensors and the image channels (Red, Green, Blue) are normalized.

The validation pipeline has similar steps involving: (i) Resizing the shorter side to 256 pixels. (ii) This was followed by a center crop of 224x224 pixels. (iii) The image was again converted to tensors and normalized in the same way as the training set.

These preprocessing steps were incorporated to accurately replicate the work of Huo et al. [2], and compare our implementation with theirs for a clear comparison. The resizing of the shorter side to 256 pixels during the validation ensures that regions of interest are not lost during center cropping. During the stage, however, random resize crop makes the system more robust against overfitting.

A. Deep Learning Models Used (EfficientNet B7, EfficientNet L2, Hiera H)

EfficientNet B7 was the most scaled-up model of Tan and Le's originally published work [3]. They experiment with scaling up convolutional neural networks in terms of depth, width, and resolution to improve accuracy while efficiently extracting more features. The system is considered efficient since it optimizes the process of finding the best height, width, and resolution parameters based on (1).

$$Depth * Width^2 * Resolution^2 \approx 2 \quad (1)$$

$$Such\ that\ Depth \geq 1,\ Width \geq 1,\ \&\ Resolution \geq 1$$

EfficientNet B7 has nearly 64M parameters and is nearly 245 Mb in terms of disc size. On the ImageNet dataset [4], EfficientNet B7 was 8.4 times smaller than the architecture with similar performance.

The EfficientNet L2 model is based on the Noisy Student architecture [5]. The authors implemented self-training on the EfficientNet architecture, where they first train a model, use it to label some unlabeled images, they then train a new model and again label the previously unlabeled images and finally train a new model on the larger corpus of images boosting available training data and feature availability [5]. The L2 is relatively larger and contains 474M parameters with a disc size of about 1.8 Gb.

Hiera is an approach to reducing the overheads on Transformer models for vision-related tasks [6]. Ryali et al., argue that using the same number of channels and spatial resolution throughout deep network was unnecessary and inefficient. They state the primary reason for these overheads is to account for the lack of inductive biases in Transformers. Instead of utilizing convolution layers to add these biases to Transformers, they utilize Masked Auto Encoders to learn spatial reasoning and speed up the overall process [6]. The Hiera has multiple variants and in our research Hiera H (huge) was utilized. Hiera H has 670M parameters and consumes 2.5 Gb disc space.

V. Experiment Setup

A. Model Training and Evaluation Metrics

Model training was set up on different GPUs ranging from Titan V (12 GB), A6000 (48 GB), and RTX 4070 (16 GB). The experiments involved training HiFuse (from scratch) for benchmarking and the remaining models (EfficientNet L2, EfficientNet B7, and Hiera H) for new research work. The splits remained the same – 50/50 to train and test on the Kvasir V1 dataset without any image augmentation or alteration to the dataset. The splitting of the dataset into test and train subsets was achieved using a random sampling method across the original dataset. Table III shows the time for training, the dataset

TABLE I. Validation Accuracy of Previous Work on 2000 Kvasir Training Images

Model	Validation Accuracy
HiFuse [2]	85%
Residual LSTM [7]	93.01%

TABLE II. CLASSWISE ACCURACY ON HIERA H.

Class Number	Class Name	Accuracy (%)
0	dyed-lifted-polyps (Class 0)	93.60%
1	dyed-resection-margins (Class 1)	93.20%
2	esophagitis (Class 2)	84.80%
3	normal-cecum (Class 3)	98.80%
4	normal-pylorus (Class 4)	100.00%
5	normal-z-line (Class 5)	92.00%
6	polyps (Class 6)	93.60%
7	ulcerative-colitis (Class 7)	96.40%
	Average Accuracy	94.05%

variant, the size on the disc, the number of parameters, and the best epoch out of 100 total epochs.

The evaluation metrics for performance included classification accuracy, the confusion matrix based on the true positive count, false negative count, false positive, and true negatives for the best epoch. Additionally, t-SNE plot and accuracy & loss curves were implemented for incremental epoch trends.

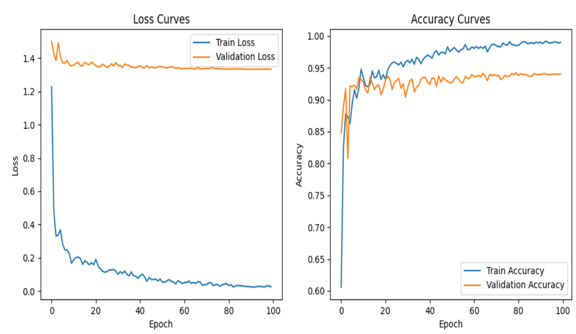

Fig 3: The loss and accuracy curve for Hiera H, Kvasir V1 (4K, 50/50)

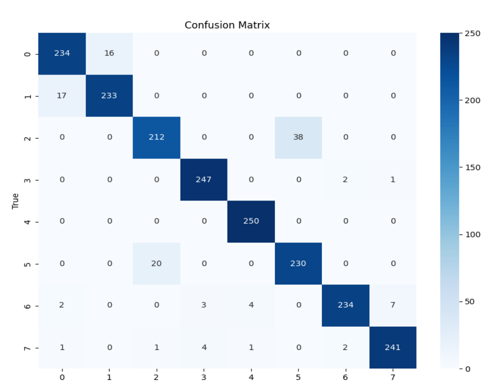

Fig 4: The confusion matrix for Hiera H, Kvasir V1 (4K, 50/50)

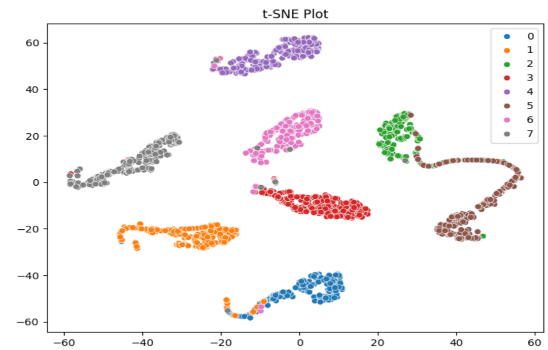

Fig 5: The t-SNE plot for Hiera H, Kvasir V1 (4K, 50/50)

B. *Software Setup*

The software setup for the system was simple. The programs were written in Python 3.10. Jupyter Notebooks [26] were the choice of IDE. Pytorch [27] was the choice of Deep Learning Framework to load the data, compute tensors and to establish and enhance the model parameters. Matplotlib [28] served as the primary tool for output visualization including display of randomly sampled data and the visualization of trends – class wise accuracy, and the accuracy and loss curves. Seaborn [29] supplemented the visualization by providing a concrete interface to visualize the confusion matrix and the t-SNE plot. Numpy [30] served as the go to utility to manipulate arrays and compute calculations such as accuracy from predictions and feature storage.

VI. RESULTS AND DISCUSSION

Hiera H was a clear winner in case of validation accuracy on its 22nd epoch. Figures 3-5 show the loss and accuracy curves, the t-SNE plot, the confusion matrix, and the class-wise accuracy respectively for all the experiments run.

Table I above shows the class wise accuracy for Hiera H's best epoch. The t-SNE plot and the confusion matrix for Hiera-H highlight the incorrect classification of classes 0 and 1, and classes 2 and 5. Classes 2 and 5 represent Esophagitis, and Normal Z line, respectively. The Normal Z line represents an anatomical landmark, while Esophagitis is a pathological finding [1], both of which reside within the esophagus or its

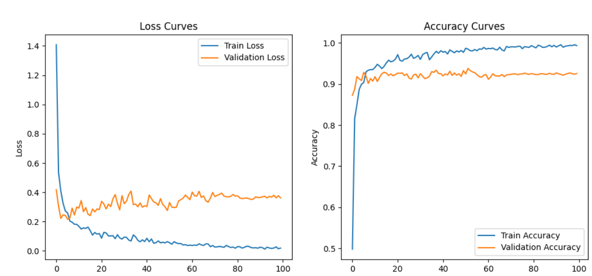

Fig 6: The loss and accuracy curve for EfficientNet L2, Kvasir V1 (4K, 50/50)

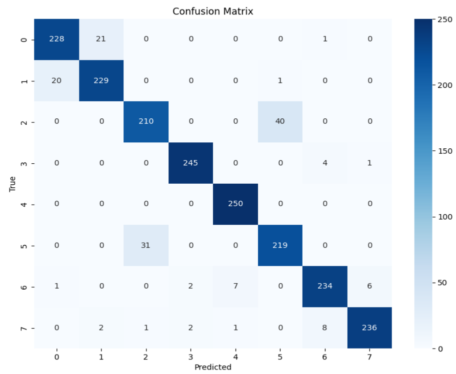

Fig 7: The confusion matrix for EfficientNet L2, Kvasir V1 (4K, 50/50)

border; hence, the visual similarity between the two leads to a decrease in accuracy. Classes 0 and 1, on the other hand, represent the Dyed Lifted Polyps and Dyed Resection Margins. The medical significance of Dyed Lifted Polyps is to highlight the presence of dyed polyps to be removed, while the Dyed Resection Margins is an evaluation of whether the polyp was fully disconnected or not. The resection may highlight an improperly removed polyp, leading to similarity between the two classes, additionally, the common factor between the two remains the dye. Images from both these classes contain a bluish dye as part of the procedure, which affects the feature extraction and may result in decreased accuracy.

Figures [6-11] show the loss and accuracy curves, the confusion matrix, and the t-SNE plot for EfficientNet L2 and B7 respectively. Both these models had high learning initially and with increasing epochs, the training and validation accuracy became divergent. The performance of both EfficientNet L2 and B7 is nearly identical with B7 surpassing the validation accuracy out of the two with a fractional cost of parameters showing superiority.

Table II highlights the performance of Hiera H among the state-of-the-art results of [2] and [7]. The accuracy of the HiFuse model benchmarked here correlates closely with the work reported [2]. Hiera H was the best-performing model with an

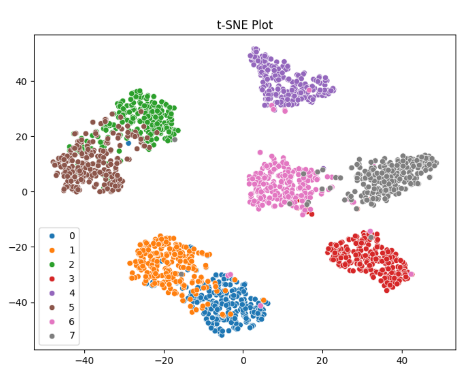

Fig 8: The t-SNE plot for EfficientNet L2, Kvasir V1 (4K, 50/50)

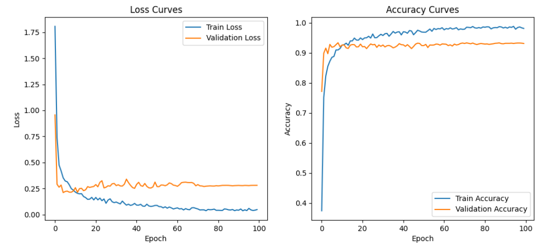

Fig 9: The loss and accuracy curve for EfficientNet B7, Kvasir V1 (4K, 50/50)

accuracy of 93.7% at its best epoch, surpassing all prior benchmarks [2, 7]; however, EfficientNet B7 offered great accuracy for the minimum training time. All experiments were carried out by unfreezing all layers of the models for transfer learning, other than HiFuse, which was trained from scratch.

Table III summarizes the training results for various models with the best epochs for the Kvasir V1 50/50 split dataset. HiFuse results reported in Table II are the results claimed by Huo et al., while the results in Table III are the ones benchmarked to compare the newly tested models – Hiera H, EfficientNet L2, and EffcientNet B7.

VII. Conclusion

Our research shows Hiera H to outperform its peers and previous techniques to classify images from the Kvasir V1 dataset by achieving an accuracy of 93.7%. EfficientNet B7 delivers exceptional performance in terms of accuracy, given the relatively low training time. The research further negates the claims that Vision Transformers perform inferior to CNN-based architectures on smaller datasets due to the lack of local inductive biases [17].

These promising results open the avenue of multiple use cases for the medical field. The simplest of all would be the direct utilization of the image classification models to aid

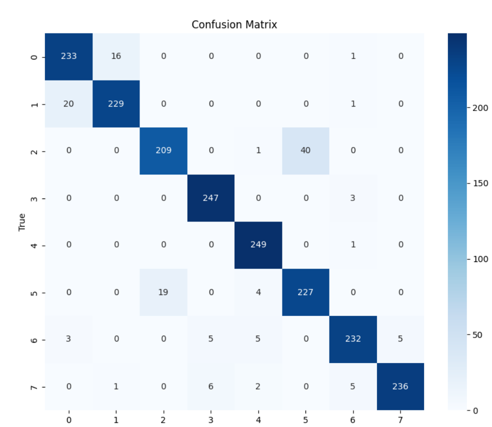

Fig 10: The confusion matrix for EfficientNet B7, Kvasir V1 (4K, 50/50)

TABLE III. THE TRAINING DETAILS FOR THE KVASIR V1 DATASET, 50/50 SPLIT FOR 100 EPOCHS.

Model	GPU Memory Used	Batch Size	Epoch	Total Training Time	Disk Space	No. of Parameters	Training Accuracy	Validation Accuracy
EfficientNet B7	11/12 GB	16	25	5807s	245 MB	63 M	94.9	92.95
Hiera Huge	48.3/49 GB	40	22	18324s	2.5 GB	670 M	94.65	93.7
EfficientNet L2	44.9/49 GB	20	15	9900s	1.8 GB	474 M	94.3	92.8
HiFuse	14.2/16GB	48	76	2361s	470 MB	123M	86.9	84.65

Fig 11: The t-SNE plot for EfficientNet B7, Kvasir V1 (4K, 50/50)

endoscopy. A capsule-based endoscopy may last for several hours [9, 11], thus manual monitoring schemes may not be the best in terms of resource utilization. The image classification models can act as breakpoints or event triggers that could alert the medical responders when the endoscope reaches a landmark of interest, and work as a replacement for magnet-based CEs, leading to reduced manufacturing costs of CEs.

References

[1] K. Pogorelov et al., "Kvasir," Proceedings of the 8th ACM on Multimedia Systems Conference, pp. 164–169, Jun. 2017. doi:10.1145/3083187.3083212

[2] X. Huo et al., HiFuse: Hierarchical multi-scale feature Fusion Network for Medical Image Classification, Sep. 2022. doi:https://arxiv.org/abs/2209.10218

[3] M. Tan and Q. V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 2020. doi:https://arxiv.org/abs/1905.11946

[4] O. Russakovsky et al., "Imagenet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, Apr. 2015. doi:10.1007/s11263-015-0816-y

[5] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet Classification," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, Jun. 2020. doi:10.1109/cvpr42600.2020.01070

[6] C. Ryali et al., Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles, 2023. doi:https://arxiv.org/abs/2306.00989

[7] Öztürk, Ş., & Özkaya, U. (2021). Residual LSTM layered CNN for classification of gastrointestinal tract diseases. Journal of Biomedical Informatics, 113, 103638. https://doi.org/10.1016/j.jbi.2020.103638

[8] M. Q.-H. Meng et al., "Wireless robotic capsule endoscopy: State-of-the-art and challenges," Fifth World Congress on Intelligent Control and Automation (IEEE Cat. No.04EX788), vol. 6, pp. 5561–5565, Jun. 2004. doi:10.1109/wcica.2004.1343799

[9] N. Dey, A. S. Ashour, F. Shi, and R. S. Sherratt, "Wireless capsule gastrointestinal endoscopy: Direction-of-arrival estimation based localization survey," IEEE Reviews in Biomedical Engineering, vol. 10, pp. 2–11, Apr. 2017. doi:10.1109/rbme.2017.2697950

[10] K. Kim et al., "A design of a high-speed and high-efficiency capsule endoscopy system," IEEE Transactions on Biomedical Engineering, vol. 59, no. 4, pp. 1005–1011, Dec. 2011. doi:10.1109/tbme.2011.2182050

[11] G. G et al., "Deep learning with unsupervised and supervised approaches in medical image analysis," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 1580–1584, Apr. 2022. doi:10.1109/icacite53722.2022.9823491

[12] M. Ayyannan et al., "Medical Image Classification using deep learning techniques: A Review," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), pp. 1327–1332, Mar. 2023. doi:10.1109/icears56392.2023.10084948

[13] M. Umer, S. Sharma, and P. Rattan, "A survey of deep learning models for medical image analysis," 2021 International Conference on Computing Sciences (ICCS), pp. 65–69, Dec. 2021. doi:10.1109/iccs54944.2021.00021

[14] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos, "Image segmentation using Deep Learning: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3523–3542, 1 July 2022, doi:10.1109/tpami.2021.3059968

[15] B. Kantheti and M. K. Javvaji, "Medical Image Classification for disease prediction with the aid of Deep Learning Approaches," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1442–1445, May 2022. doi:10.1109/iciccs53718.2022.9788144

[16] A. Dosovitskiy et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2021. doi:https://arxiv.org/abs/2010.11929

[17] D. Gandhi, V. Shah, and P. M. Chawan, "A Vision Transformer approach for classification an A small-sized medical image dataset," 2022 5th International Conference on Advances in Science and Technology (ICAST), pp. 519–524, Dec. 2022. doi:10.1109/icast55766.2022.10039593

[18] L. Gai, W. Chen, R. Gao, Y.-W. Chen, and X. Qiao, "Using vision transformers in 3-D medical image classifications," 2022 IEEE International Conference on Image Processing (ICIP), pp. 696–700, Oct. 2022. doi:10.1109/icip46576.2022.9897966

[19] K. Ranaweera and P. N. Pathirana, "Leveraging vision transformers for enhanced accuracy in pneumonia detection from medical imaging data," 2024 46th Annual International Conference of the IEEE Engineering in

Medicine and Biology Society (EMBC), pp. 1–4, Jul. 2024. doi:10.1109/embc53108.2024.10782752

[20] T. P. Armand, S. Bhattacharjee, H.-K. Choi, and H.-C. Kim, "Transformers effectiveness in medical image segmentation: A comparative analysis of unet-based architectures," 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 238–242, Feb. 2024. doi:10.1109/icaiic60209.2024.10463435

[21] T. Agrawal, R. Gupta, and S. Narayanan, "On evaluating CNN representations for Low Resource Medical Image Classification," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1363–1367, May 2019. doi:10.1109/icassp.2019.8682397

[22] D. Gupta, G. Anand, P. Kirar, and P. Meel, "Classification of endoscopic images and identification of gastrointestinal diseases," 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), pp. 231–235, May 2022. doi:10.1109/com-it-con54601.2022.9850571

[23] Tang, F., Yang, Q., Yu, L., & Long, J. (2024a). Mambaoutvision: An hybrid mambaout-transformer vision backbone for medical image classification. SSRN. https://doi.org/10.2139/ssrn.5044694

[24] F. Bray et al., "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 74, no. 3, pp. 229–263, Apr. 2024. doi:10.3322/caac.21834

[25] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, 2017. doi:arXiv:1706.03762

[26] T. Kluyver *et al.*, "Jupyter Notebooks - a publishing format for reproducible computational workflows," *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 2016. doi:10.3233/978-1-61499-649-1-87

[27] Paszke *et al.*, "PyTorch: an imperative style, high-performance deep learning library," *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 8026–8037, Dec. 2019. doi:10.5555/3454287.3455008

[28] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. doi:10.1109/mcse.2007.55

[29] M. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, Apr. 2021. doi:10.21105/joss.03021

[30] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. doi:10.1038/s41586-020-2649-2