Hybrid EfficientNet-B0, Vision Transformer, and LSTM Framework for Wireless Capsule Endoscopy Abnormality Detection

R. Vijayalakshmi
Department of Information Technology
Velammal College of Engineering and Technology
Madurai,India
rvl@vcet.ac.in

Hariharan M
Department of Information Technology
Velammal College of Engineering and Technology
Madurai,India
hariharan.m5577@gmail.com

Hariram T
Department of Information Technology
Velammal College of Engineering and Technology
Madurai,India
hariramhari569@gmail.com

Abstract*— The Wireless Capsule Endoscopy (WCE) procedure, also known as WCE, is one of the widely employed noninvasive visualization tools that can be applied for the diagnosis of GI ailments. Indeed, during a single exam, WCE is capable of capturing thousands of images; thus, even professionals face difficulties examining them due to their vast number and require quite some time for the task. Given the amount of visual data obtained, the probability of missing any vital abnormalities such as hemorrhage, ulcerations, polyps, and inflammation is rather high. As an attempt to address the challenge, the proposed research proposes a novel hybrid approach to automatic GI abnormalities detection that employs two models, namely EfficientNet-B0 along with Vision Transformer (ViT) and LSTM. While EfficientNet-B0 is responsible for extracting local spatial details of each WCE image, ViT focuses on modeling the global context. Then, these learned representations are aggregated and the LSTM layer is utilized further in order to continuously improve the feature sequence, thereby improving the general discriminative power and classification performance. This entire framework is programmed with the Python programming language and the experiments on the dataset from the Wireless Capsule Endoscopy images are performed through the use of various standard evaluation measures. The results of this experiment indicate that the model achieves classification accuracy exceeding 90%, implying that the approach of combining these three techniques indeed works well. In addition, this methodology enhances the effectiveness of abnormality detection and can thus be relied upon as an effective diagnostic tool for gastrointestinal problems. In addition, it is expected to alleviate some burdens for healthcare professionals and speed up their diagnostic work.

Keywords— Wireless Capsule Endoscopy, EfficientNet-B0, Vision Transformer, LSTM, Deep Learning, Gastrointestinal Abnormality Detection.

I. INTRODUCTION

Wireless Capsule Endoscopy (WCE) is what this kind of revolutionary diagnostic technique is all about because it enables non-invasive visualization of the gastrointestinal tract. The device consists of a camera, light source, battery, and wireless transmitter. Collectively, the elements of the device enable thousands of pictures to be taken as the capsule moves through the digestive tract. WCE has emerged as an effective tool in identifying different conditions of the gastrointestinal tract including but not limited to bleeding, ulcers, polyps, Crohn's disease, and other inflammatory lesions. Although WCE has made possible visualization of the interior of the gastrointestinal tract, reviewing thousands of images generated in such a process can become somewhat tedious and time-consuming as well. In fact, physicians have to review thousands of images each time, which could potentially result in errors during diagnosis. Thus, the topic of automatic anomaly detection techniques has gained much significance in medical image processing research.Recent advancements in the field of artificial intelligence, and more specifically deep learning, have contributed significantly towards the development of efficient disease detection techniques. CNNs have shown impressive performance in exploiting local spatial features, while transformer-based models have performed impressively in capturing contextual information using self-attention mechanisms. However, many existing models tend to focus mainly on a single perspective of the problem, be it feature extraction or context recognition. A proposed framework focuses on improving the performance of abnormality detection techniques.

II. LITERATURE SURVEY

Wireless Capsule Endoscopy (WCE) is a technique for detecting various gastrointestinal abnormalities including ulcers, polyps, bleeding, inflammatory diseases. In WCE exams., medical experts are required to manually examine recorded images and pinpoint potential diseases. Visual inspection is known to be an expert-driven, tedious task that suffers from human errors. To mitigate this problem, several artificial intelligence (deep learning) approaches have been introduced to assist or automatically discover the disease and classify them. Convolutional Neural Networks (CNNs) are widely known for their capability of learning specific local visual patterns which are abundant in medical images. On the other hand, Vision Transformers (ViTs) are capable of capturing global contextual representations from self-attention mechanism. In recent years, research on hybrid models consisting of CNNs and transformers, as well as other sequential learning algorithms have been explored in order to improve the accuracy and robustness of the models. In this survey, we review recent advances that focus on gastrointestinal disease detection from WCE exams., exploring feature extraction and image classification using transformer-based models for medical imaging.

Al-Otaibi et al. [1] (2024) suggested Efficient-Gastro, a deep learning framework based on EfficientNet for detection of gastrointestinal disease from Wireless Capsule Endoscopy (WCE) images. In this work, we have used datasets of WCE images and applied transfer learning and data augmentation techniques to improve the classification performance. The main contribution of this work is an optimized EfficientNet

for the computation efficiency of automatic diagnosis of gastrointestinal diseases.

Varam et al. [2] , in 2024, looked into how Vision Transformer models are in practice deployed for gastrointestinal disease identification, specifically via the Kvasir-Capsule dataset. However, the most important part is that attention was paid mainly to the quantization of transformers and the operation of these systems on edge devices. It can be said that the main idea is to understand the fact that transformer-based models are able to provide good results in diagnostics but need less computational resources.

Sharmila and Geetha, [3] (2024), introduced a Hierarchical Spatio Pyramid TranfoNet structure; however, it was combined with Efficient-CondConv SwishNet structure for gastrointestinal disease classification. It can be noticed that they combined hierarchical feature extraction and transformer-based contextual awareness, and hence the resultant system is more robust to noise and outliers. All in all, this study introduces a hybrid system, which is more effective to identify different types of gastrointestinal diseases.

Das et al. [4] (2024) introduced a CapsuleNet built on top of EfficientNet-B7, and for the gastrointestinal disease classification problem , they used capsule endoscopy images as inputs. In their pipeline they kind of used deep feature extraction , plus capsule learning mechanisms, so that the model can better distinguish diseases across several abnormality classes. The main thing they claim as a contribution is improved classification performance especially when the medical datasets are imbalanced , which is often the case in real world scenarios.

Wang Yang and Tang [5] (2023) kind a proposed a Vision Transformer with Hybrid Shifted Windows, aimed at gastrointestinal endoscopy image classification, yes . In their setup they tested the approach on gastrointestinal image datasets and they use combined local as well as global contextual learning, which helps the network to “see” fine details together with broader cues. In summary, the proposed method was more successful in terms of classification results when compared to others, and it proves that transformers do work effectively in the diagnosis of gastrointestinal diseases.

Oukdach et al. [6] in 2023 proposed ConV-ViT, which is the combination of Convolutional Neural Networks and Vision Transformers, for Wireless Capsule Endoscopy image classification. The idea is to merge the local CNN features with context-aware transformer-based features, and thereby the model will be able to capture both the near textures and far-off information. In any case, the results presented in this paper proved to be better in terms of detecting anomalies through such a complementary methodology.

Bissoonauth-Daiboo et al. [7] (2023) considered the Vision Transformers architecture applied to the problem of classifying endoscopic images. Compared to more common CNNs used in such tasks, the researchers pitted transformer models against the latter, and, yes, they were able to show that their performance was fairly comparable, which was measured by their effectiveness in classification. The contribution of this paper was basically to confirm the viability of transformers on actual medical images.

Regmi et al. [8] in 2023 studied Vision Transformers for classifying GI images using the Kvasir dataset. Here, they compared transformer structures against CNN architectures that are considered state-of-the-art in terms of performance, and concluded that it was very successful, with the F1-score reaching 94.36%. Thus, the overall idea from this research indicates the usefulness of learning-based transformer methods to detect diseases, particularly gastroenterological disorders.

Wu et al. [9] (2023), for instance, seem to have put forward FLATer, which is a relatively light transformer architecture designed using local feature attention to diagnose digestive system diseases. The model seems to be optimized regarding efficiency and accuracy, therefore appearing better suited to medical practice applications. The article further revealed the possibility of practically implementing transformer-based diagnostic models in the medical sphere.

Thai et al. [10] (2023) kinda developed a multimodal learning framework for gastro intestinal visual question answering. In their design the model integrated transformer based visual learning with image enhancement things , to better support diagnostic understanding of the gastrointestinal images. Overall , the key contribution was improving how multimodal medical image interpretation is handled, like in a more sensible way for the task.

Dosovitskiy et al. [11] , (2021) put forward the Vision Transformer (ViT), which is basically a transformer-style image recognition setup trained using the ImageNet-21K dataset. In their design, the usual convolution operations are swapped out for self-attention mechanisms and somehow it reached state-of-the-art image classification results. Overall this study laid the groundwork for transformer based medical image analysis, and it feels like a turning point in that direction as well.

Tan and Le [12] (2019) put forward EfficientNet, like a kind of convolutional neural networks family where scaling is done in a compound way for depth , width , and the image resolution. When evaluated on ImageNet, EfficientNet reportedly got higher classification accuracy, and all that with notably fewer parameters. The overall architecture then got quite widely adopted for medical image classification problems.

Redmon and Farhadi [13] in 2018 came up with YOLOv3, which is a real time object detection framework, tested on the COCO dataset. This method uses one stage object localization process which can detect the presence of objects both fast and accurately. Then the entire architecture was later repurposed for detecting anomalies in medical images and also localizing anomalies.

He et al. [14] introduced ResNet, which was a deep residual learning framework, which introduced residual connections and hence helped in training very deep networks quite easily.

This approach resulted in better classification performance, and soon became the basic architecture used for analyzing medical images.

Hochreiter & Schmidhuber [15] in 1997,came up with the Long Short Term Memory (LSTM) network designed to deal with sequential learning. For some reason, it had an immense effect on the world. The memory cells along with its gating mechanisms were introduced in order for it to be able to understand long-term dependencies in the data. Today,LSTM remains one of the most popular architectures of the recurrent neural networks applied in deep learning.

III. METHODOLOGY

This architecture is sort of a combination of deep learning model which makes use of EfficientNet-B0 together with a Vision Transformer model along with the LSTM block for the purpose of automated abnormality detection of images taken through Wireless Capsule Endoscopy (WCE). The major reason behind this is that the model will be able to acquire local and global cues from the WCE images, thereby making it useful in diagnosing Gastrointestinal diseases. Generally, the whole process begins with image preprocessing and then feature extraction and fusion of features. After this follows feature refining as well as classification and evaluation of performance. The complex features of the lesions are acquired through CNNs in addition to transformers.

A. Image Preprocessing

Pre-processing of images is performed to ensure that the data somehow remains consistent, and that the quality of data becomes much improved, okay. All the Wireless Capsule Endoscopy images are initially down-sampled to 224 × 224 pixels, and then normalized before being fed into the deep learning network. The purpose of normalization is to minimize the variance in the intensity of the pixels, which in return results in an efficient training procedure, relatively less jittery. Let us denote the incoming image as:

$$I \in \mathbb{R}^{224\times224\times3}$$

Where,

- I represents the normalized RGB image.

B. Local Feature Extraction Using EfficientNet-B0

EfficientNet-B0 is employed as the backbone for extracting efficient convolutional features as it has been proved to be capable of achieving high levels of accuracy while maintaining the number of parameters relatively low, as well. Since compound scaling principle has been applied in EfficientNet, tuning of the depth, width, and input resolution of the network happens simultaneously, instead of separately. EfficientNet-B0 extracts local spatial features from the gastrointestinal images in the form of texture, lesions borders, color variation, and irregularities, and they can be represented as in:

$$F_{local} = f_{EfficientNet}(I)$$

Where,

- I stands for the input WCE images.
- F_{local} stands for the local feature representation extracted using EfficientNet-B0.

C. Vision Transformer

However, even though convolutional neural networks are efficient at extracting local patterns within images, they are unable to capture long-range correlations within distant patches of an image. Therefore, a Vision Transformer is added to this process to help address this problem. First, the image is divided into patches and each of them represented by an embedding vector. Afterwards, the encoder transformer layers step in and begin working on these embedding vectors. In this case, the role of self-attention is to help the model understand the global relations within all image patches. The formula for self-attention can be described by :

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where:

- Q= Query matrix
- K= Key matrix
- V= Value matrix
- d_k= Dimension of key vectors

The equation above facilitates learning of relationships among image patches within the Vision Transformer model.

A Vision Transformer creates contextual features defined by the following expression:

$$F_{global} = f_{ViT}(I)$$

In which:

- F_{global} is the global contextual features that the Vision Transformer extracts

D. Feature Fusion Strategy

To exploit the somewhat complementary strengths of EfficientNet-B0 along with a Vision Transformer, the extracted feature vectors are brought together through a feature fusion mechanism . EfficientNet-B0 delivers quite crisp local details, while the Vision Transformer brings broader, more global context, you know in a sense. As a result, the final fused feature representation is formed as:

$$F_{fused} = F_{local} \oplus F_{global}$$

Where:

- F_{local}= Local features from EfficientNet-B0
- F_{global}= Global features from Vision Transformer
- $\oplus$= Feature concatenation operation

The fused feature vector combines both local and contextual information.

E. LSTM-Based Feature Refinement

After feature fusion, the resulting feature vector is fed into a Long Short-Term Memory (LSTM) network, for feature refinement. The LSTM layer helps with the enhancement of the extracted features by keeping useful information and at the same time trimming over- redundant parts, sometimes this makes the representation more stable. As a result, the classification robustness improves, together with the ability to represent features in a better way. The refined feature representation can be written as:

$$F_{refined} = f_{LSTM}(F_{fused})$$

Where:

- F_{fused}= Combined feature vector
- $F_{refined}$= Refined feature representation generated by the LSTM network

The LSTM enhances the quality of learned features before classification.

F. Classification Layer

The refined feature vector is then passed into a fully connected dense layer, and after that it goes through a Softmax activation, for the last classification. Here, the class probability is basically computed like this:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

Where:

- $P(y_i)$= Probability of class i
- z_i= Output score of class i
- C= Total number of classes

The class with the highest probability is selected as the final prediction.

G. Model Training

The framework that we propose is trained with the Adam optimization algorithm because it has an adaptive learning style and it tends to converge efficiently too, not always but in practice. For the loss part we use Categorical Cross-Entropy, so the classification performance can be optimized in a stable way. The loss function is set up like this:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Where:

- L= Cross-Entropy loss
- y_i= True class label
- $\hat{y}_i$= Predicted probability
- N= Number of training samples

The objective is to minimize the loss during training.

H. Performance Evaluation

The effectiveness of the proposed EfficientNet–ViT–LSTM framework is evaluated using standard classification metrics, including Accuracy, Precision, Recall, and F1-Score.

Accuracy measures how the classifier works overall, and it is usually computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

Precision tells us the share of the samples that were predicted as positive but are actually correct :

$$Precision = \frac{TP}{TP + FP}$$

Where:

- TP = True Positive
- FP = False Positive

Recall is related to how well the system is able to recognize true positives, which means that it tests true positive recognition:

$$Recall = \frac{TP}{TP + FN}$$

Where:

- TP = True Positive
- FN = False Negative

The F1-Score provides a balanced measure of Precision and Recall and is calculated as:

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

In which,

- Precision = Positive predictive value
- Recall = Detection capacity of the model

These performance measurement metrics give us a complete understanding of the model's detection capability for GI problems.

I. Overall Workflow of the Proposed System

Workflow for the Proposed Framework involves steps such as obtaining images through the process of Wireless Capsule Endoscopy and pre-processing of the images. Following that, the pre-processed images will be simultaneously passed to both EfficientNet-B0 and the Vision Transformer where the former concentrates on extracting the local cues while the latter is concerned with global cues. Once both sets of local and global features have been obtained, they are combined to form an even more unified feature set after which an LSTM layer is employed for refining the combined local and global cues. Finally, the obtained refined local and global cues are fed to the Softmax classifier layer in order to classify gastrointestinal abnormalities or, at least, make an informed decision about their possible identities. The effectiveness of this whole procedure is measured through Accuracy,

Precision, Recall, and F1-Score measures. The main strength of the proposed framework lies in its combination of a convolutional learning paradigm, transformers' global understanding of the images, as well as feature extraction and refining mechanisms.

IV. RESULTS AND DISCUSSION

The mentioned above EfficientNet–ViT–LSTM framework was sort of tested for its efficiency in detecting gastrointestinal diseases through the use of WCE images. Training of the model was performed based on pre-processed WCE images, and the evaluation process was carried out by means of such metrics as Accuracy, Precision, Recall, and F1-Score. According to the experimental findings, the combination of EfficientNet-B0, Vision Transformer, and LSTM appears very useful for classification purposes, as the combination involves the use of local feature extraction, broader context learning, and, finally, feature refinement.

A. Training Performance Analysis

Figure 1 depicts the performance accuracy and validation accuracy obtained during training. It is evident from the figure that there is an increase in accuracy gradually and after some epochs the accuracy remains constant, thereby implying that the framework learns important patterns that make the model capable of generalizing. In addition, the relatively smooth convergence of the curves depicts stability in the framework that is able to extract discriminating patterns from wireless capsule endoscopy images and not noise. Lastly, the final accuracy of the model stands at 92.8%.

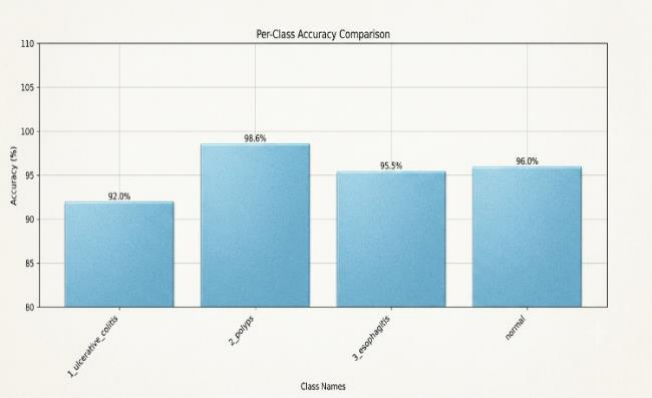

Fig. 1. Accuracy comparision

B. Abnormality Detection Results

Figure 2 highlights some outputs of abnormality detection using the proposed framework, and all in all, it seems rather promising.The model was able to detect various kinds of gastrointestinal abnormalities, including bleeding ulcers and polyps, as well as inflammatory lesions despite their non-obvious appearances. In particular, what works here is the extraction of the local features from the EfficientNet-B0 and the global contextual features from the Vision Transformer, which together allow the network to effectively detect abnormal patterns. Hence, it can be stated that the proposed hybrid architecture has proven itself capable of detecting important abnormalities in WCE images.

Fig. 2. Types Of Abnormal Detection

C. Normal Image Classification Results

Figure 3 represents some normal gastrointestinal images, which have been identified properly through the proposed system. It turns out that the model does a fairly good job distinguishing between normal tissue structures and abnormal tissues; hence, there are fewer instances of false positives. Identifying normal tissues becomes very crucial, since it enhances diagnostic accuracy, while also helping healthcare professionals concentrate on the abnormal ones. In any case, the results reveal that the proposed framework performs reliably on both normal and abnormal WCE images, albeit the circumstances being somewhat different.

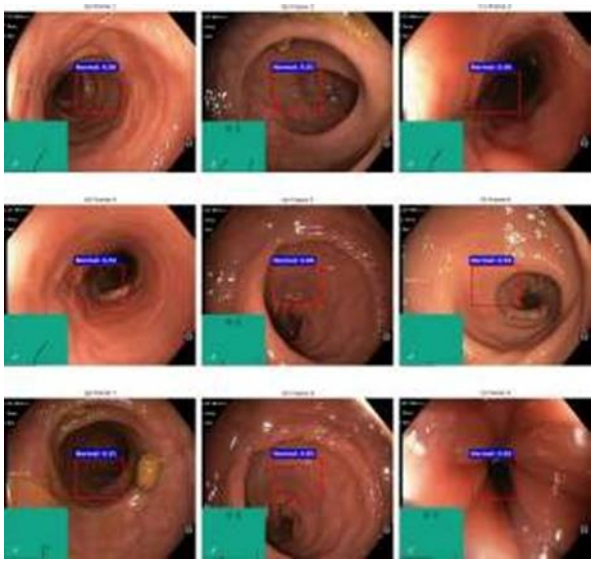

Fig. 3 . Normal Detection

D. Performance Evaluation

Table I demonstrates how good the model under discussion compares to baseline deep learning models. Overall, the

performance of the proposed EfficientNet-ViT-LSTM model is better than all the metrics used to evaluate its performance.

TABLE I. PERFORMANCE EVALUATION USING BASELINE MODELS

Model	Accuracy	Precision	Recall	F1-Score
CNN	84.2%	83.5%	82.9%	83.1%
ResNet50	87.1%	86.8%	86.2%	86.5%
Vision Transformer	89.3%	88.9%	88.5%	88.7%
Proposed EfficientNet + ViT + LSTM	92.8%	92.1%	91.9%	92.2%

The result indicates that the suggested hybrid architecture achieves highest classification accuracy, which is 92.8%. The EfficientNet B0 model is effective in capturing small scale texture and lesion information from the images, while the Vision Transformer captures broader and more contextual links among image regions in parallel. Afterward, the LSTM layer refines the combined features representation, thus classification performance will be more stable and accuracy is improved. Compared to standard CNN models and purely transformer-based models, the suggested architecture seems to have higher abnormality detection performance.

E. Ablation Analysis

The ablation study to analyze the contribution of each individual part of architecture involved incrementally adding EfficientNet-B0, Vision Transformer, and LSTM.

TABLE II. ABLATION STUDY

Model configuration	Accuracy
EfficientNet Only	86.4%
Vision Transformer Only	88.1%
EfficientNet + ViT	91.2%
EfficientNet + ViT + LSTM	92.8%

From the results of ablation study, it can be seen that the inclusion of each component brings positive changes to the model as a whole. The accuracy of EfficientNet-B0 alone amounts to 86.4% whereas Vision Transformer has an accuracy level of 88.1%. In the case when both components are combined, the classification accuracy grows to 91.2%, proving the benefits of combining local cues with global feature extraction. Finally, with the additional LSTM layer, the score raises to 92.8%, practically proving the effectiveness of feature refinement phase.

F. Discussion

These results confirm that the application of this proposed EfficientNet-ViT-LSTM combination is indeed effective for gastrointestinal abnormality recognition tasks, as expected. The EfficientNet-B0 model performs the role of extracting features efficiently, the Vision Transformer, on the other hand, contributes towards identifying the global or contextual relations within an image. The role of the LSTM is to optimize the representation of fused features prior to classification. When combined in a way where each complements one another, it allows this network model to perform better when distinguishing between normal and abnormal gastrointestinal diseases. Furthermore, this novel framework can potentially assist physicians in saving time during the tedious process of examining the WCE images, and improve the performance of CAD systems as well.

V. CONCLUSION

The current paper introduced a combination approach based on EfficientNet-B0, Vision Transformer (ViT), and LSTMs to automatically detect abnormalities in the gastrointestinal tract using Wireless Capsule Endoscopy (WCE). The first was employed to extract local details, and the second model to capture larger context, while the third network was incorporated to refine feature fusion to achieve higher accuracy and stability in the end.

According to the experiments conducted on this work, it achieved a performance rate of 92.8%. Precision and recall are both considered high as well. In general, the results show that by integrating convolutional processing and transformer-style attention along with sequence modeling technologies, a higher level of performance in terms of abnormality detection is possible. Moreover, there may be an improvement in the reliability of classification. In practical terms, it will allow doctors to go through vast data sets of WCE imaging, which makes it easier to identify diseases of the digestive tract with accuracy and reliability.

VI. REFERENCES

[1] S. Al-Otaibi, A. Rehman, M. Mujahid, S. Alotaibi, and T. Saba, "Efficient-gastro: Optimized EfficientNet model for the detection of gastrointestinal disorders using transfer learning and wireless capsule endoscopy images," *PeerJ Computer Science*, vol. 10, p. e1902, 2024.

[2] D. Varam, L. Khalil, and T. Shanableh, "On-Edge Deployment of Vision Transformers for Medical Diagnostics Using the Kvasir-Capsule Dataset," *Applied Sciences*, vol. 14, no. 18, p. 8115, 2024.

[3] V. Sharmila and S. Geetha, "Gastro Intestinal Disease Classification Using Hierarchical Spatio Pyramid TranfoNet With PitTree Fusion and Efficient-CondConv SwishNet," *IEEE Access*, 2024.

[4] A. Das, A. Singh, N. Kumar, and S. Prakash, "CapsuleNet: A Deep Learning Model To Classify GI Diseases Using EfficientNet-B7," 2024.

[5] W. Wang, X. Yang, and J. Tang, "Vision Transformer With Hybrid Shifted Windows for Gastrointestinal Endoscopy Image Classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4452–4461, 2023.

[6] Y. Oukdach, Z. Kerkaou, M. El Ansari, and L. Koutti, "ConV-ViT: Feature Fusion-based Detection of Gastrointestinal Abnormalities

using CNN and ViT in WCE Images," in *Proceedings of WINCOM*, 2023.

[7] P. Bissoonauth-Daiboo, et al., "Endoscopic Image Classification Using Vision Transformers," in *International Conference on Advances in Artificial Intelligence*, 2023.

[8] S. Regmi, A. Subedi, U. Bagci, and D. Jha, "Vision Transformer for Efficient Chest X-ray and Gastrointestinal Image Classification," 2023.

[9] S. Wu, et al., "High-Speed and Accurate Diagnosis of Gastrointestinal Disease: Learning on Endoscopy Images Using Lightweight Transformer with Local Feature Attention," *Bioengineering*, vol. 10, no. 12, 2023.

[10] T. M. Thai, A. T. Vo, H. K. Tieu, L. N. P. Bui, and T. T. B. Nguyen, "UIT-Saviors at MEDVQA-GI 2023: Improving Multimodal Learning with Image Enhancement for Gastrointestinal Visual Question Answering," in *ImageCLEF MEDVQA-GI Challenge*, 2023.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

[13] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[15] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.