Multi-Class Abnormality Classification Task in Video Capsule Endoscopy

Dev Rishi Verma, Vibhor Saxena, Dhruv Sharma, Arpan Gupta

Institute of Engineering and Technology, JK Lakshmipat University, Jaipur, 302026
Rajasthan, India

Email: {devrishiverma, vibhorsaxena, dhruvsharma, arpan.gupta}@jklu.edu.in

Abstract

In this work we addressed the challenge of multi-class anomaly classification in Video Capsule Endoscopy (VCE)[1] with a variety of deep learning models, ranging from custom CNNs to advanced transformer architectures. The purpose is to correctly classify diverse gastrointestinal disorders, which is critical for increasing diagnostic efficiency in clinical settings. We started with a baseline CNN model and improved performance with ResNet[2] for better feature extraction, followed by Vision Transformer (ViT)[3] to capture global dependencies. We further improve the results by using Multiscale Vision Transformer (MViT)[4] for improved hierarchical feature extraction, while Dual Attention Vision Transformer (DaViT) [5] delivered best results by combining spatial and channel attention methods. Our best balanced accuracy was **0.8592** and Mean AUC was **0.9932**. This methodology enabled us to improve model accuracy across a wide range of criteria, greatly surpassing all other methods.

1 Introduction

The entire gastrointestinal (GI) tract, especially the small intestine, which is difficult to reach with traditional endoscopy, can be examined with Video Capsule Endoscopy (VCE) [6], a non-invasive diagnostic technique, whereby a capsule fitted with a camera captures images (or videos) at regular intervals as it passes through the GI tract. The data is captured at a rate ranging from 2 to 35 FPS, having a resolution of around 256×256 ($H \times W$). However, manual analysis of this captured data is time-consuming and prone to errors due to the enormous volume of video frames produced by every procedure. Learning-based automated systems have enormous potential to improve the precision and effectiveness of identifying gastrointestinal anomalies such as ulcers, polyps, and bleeding [1]. Recent developments in transformer architectures have demonstrated incredible promise in this area, especially with models such as the Vision Transformer (ViT) [3], which use attention mechanisms [7] and hierarchical processing of visual information to dramatically improve performance on image recognition tasks. [3].

In order to solve this issue, the Capsule Vision 2024 Challenge promotes the creation of sophisticated AI models that can classify abnormalities across multiple classes. The dataset

provided consists of images from 3 publicly available datasets, SEE-AI [8], KID [9], KVASIR-capsule [10] and one private AIIMS [11] dataset. In order to improve the classification performance, we experiment with a variety of deep learning architectures, starting with a custom CNN and working our way up to more complex models like the Vision Transformer (ViT) and its variations. By combining convolutional techniques with attention mechanisms, our method takes advantage of the spatial-temporal complexity of VCE image data. This improves the model's capacity to identify both local and global dependencies within the video frames. Furthermore, prior research has shown that deep learning methods can successfully detect anomalies in VCE, which offers a strong basis for our DaViT [5] model's performance.

We present a detailed analysis of our experiments and the associated results. In Section 2, we provide the methods and models used for training, along with the details of data augmentation. Section 3 and 4 present our results and related discussions, respectively. Finally, conclusion is given in Section 5.

2 Methods

Figure 1 illustrates our experimental setup. The steps and models are described in the subsequent subsections.

2.1 Preprocessing

To improve performance in medical imaging, the model's preprocessing involves particular transformations for the training and validation datasets. Images in the training set are resized to 224 $\times$ 224 pixels and rotated up to 45 degrees in addition to being randomly flipped horizontally and vertically. These methods aid in enhancing the data and strengthen the model's resistance to scale and orientation changes. In order to stabilize model learning, images are also transformed into tensors and normalized using standard RGB mean and standard deviation values.

2.2 CNN (Convolutional Neural Network)

With three convolutional layers that detected spatial information, the first model made use of a Convolutional Neural Network (CNN) designed for structured data, such as photographs. Each layer maintains input dimensions while using a 3x3 kernel, stride of 1, and padding of 1. Complex data associations are captured by the non-linearity introduced by the ReLU activation function.As seen with max pooling layers, the spatial dimensions are cut in half while maintaining the crucial information by employing a 2x2 kernel and stride of 2. This enables the model to concentrate on crucial elements of feature maps for examination. The number of classes in our dataset corresponds to the projected class scores.

2.3 ResNet (Residual Network)

We refined a pretrained ResNet50 model for image classification, known for its deep network architecture and use of residual connections to address the vanishing gradient problem in very deep networks. By allowing some layers to skip steps and maintain a steady flow of gradients, the model can effectively learn even with many layers, improving overall performance. We modified the ResNet code provided by the organizers by changing the input image size to 224x224, enhancing the model's accuracy in capturing fine-grained features and improving classification performance. Our ResNet-50 setup includes an initial layer for basic feature extraction, residual blocks for learning complex patterns with better connections, and a final layer customized for accurate classification based on specific classes in the dataset.

2.4 ViT (Vision Transformer)

The Vision Transformer model processes images by breaking them down into patches and reshaping them for processing by the Transformer. A trainable linear projection maps patches to a fixed size, resulting in patch embeddings. Position embeddings are added for positional information. The Transformer encoder consists of layers alternating between self-attention and multi-layer perceptron blocks. ViT has less built-in image-specific knowledge compared to CNNs, learning spatial relationships from scratch. Feature maps from a CNN can also be used as input patches for flexibility in patch size and representation. This approach allows ViT to learn spatial relationships between patches independently from pre-established image features.

2.5 MViT (Multiscale Vision Transformer)

The Multiscale Vision Transformer (MViT) architecture utilizes transformer blocks in different stages to process images efficiently. The model, based on the pre-trained mvit v2 large, adjusts to input complexities by operating at various spatial and temporal resolutions. At the core of MViT is the Multi Head Pooling Attention (MHPA) mechanism, which pools input sequences before attention calculations to handle varying resolutions effectively. The pooling operator adjusts sequence length through specified stride and padding, preserving crucial information while reducing computational load. Attention mechanism processes shortened tensors, capturing essential data relationships for effective learning. Parallelization of attention across multiple heads enhances feature learning, enabling MViT to identify diverse characteristics in the data. [4].

2.6 DaViT (Dual Attention Vision Transformers)

The Dual Attention Vision Transformers (DaViT) architecture efficiently captures fine details and overall patterns in visual data through four stages. It incorporates dual attention blocks that use spatial window attention to focus on specific regions within small patches of an image for computational efficiency. Additionally, channel group attention enables different channels to interact and gather global information. DaViT offers various configurations, such as DaViT-Tiny, DaViT-Small, and DaViT-Base, to optimize performance according to task complexity

and dataset size. This versatility makes DaViT a promising solution for a wide range of vision-related challenges, striking a balance between efficiency and effectiveness in processing visual information. [5].

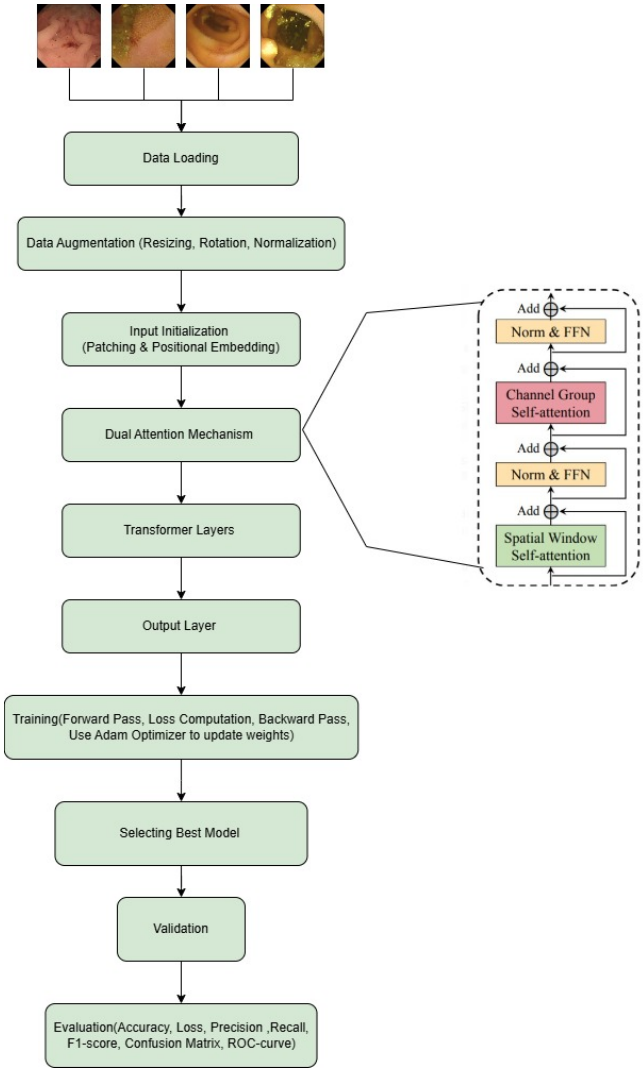

Figure 1: Block diagram of the developed DaViT pipeline. DaViT model same as in [5].

3 Results

A number of important metrics, such as the precision-recall curve, ROC curve, and per-class precision, recall, and F1 score, were used to assess the DaViT model's performance on the validation dataset.[12]. These metrics assess the model's overall efficacy across a range of abnormalities and offer insight into how well it can differentiate between classes.

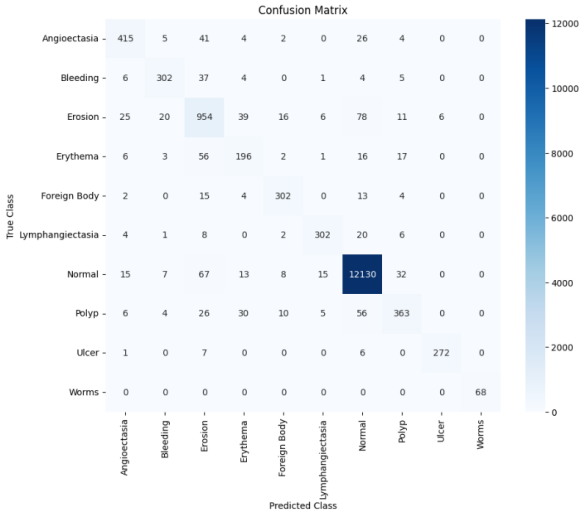

Figure 2: Confusion Matrix for the DaViT Model on the validation set.

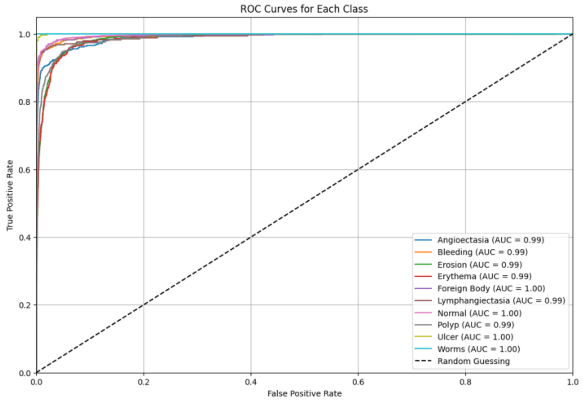

Figure 3: ROC Curve for the DaViT Model on the validation set.

In our experiments, the DaViT model outperformed traditional methods across key metrics. Table 1 give the classification metrics for the 10 classes when our finetuned DaViT model is evaluated on the validation set (having 16132 samples). The detailed performance results when comparing different models, are shown in Table 2.

Classes	**Precision**	**Recall**	**F1-Score**	**Specificity**
Angioectasia	0.8645	0.8350	0.8495	0.9958
Bleeding	0.8830	0.8412	0.8616	0.9974
Erosion	0.7877	0.8259	0.8064	0.9828
Erythema	0.6758	0.6599	0.6678	0.9940
Foreign Body	0.8830	0.8882	0.8856	0.9974
Lymphangiect	0.9151	0.8804	0.8974	0.9982
Normal	0.9822	0.9872	0.9847	0.9430
Polyp	0.8212	0.7260	0.7707	0.9949
Ulcer	0.9784	0.9510	0.9645	0.9996
Worms	1	1	1	0.990

Table 1: Classification report for all classes on validation set

	Mean Specificity	**Mean Avg. Precision**	**Mean Sensitivity**	**Mean F1-score**	**Mean AUC**	**Balanced Acc.**
Custom CNN	0.9642	0.6545	0.5703	0.6265	0.9524	0.5703
ResNet50	0.9844	0.8685	0.7963	0.8192	0.9900	0.7963
ViT	0.9803	0.8206	0.7505	0.7710	0.9847	0.7505
mViTv2-l	0.9890	0.9121	0.8307	0.8505	0.9933	0.8307
DaViT-s	**0.9904**	**0.9147**	0.8595	**0.8688**	**0.9932**	0.8595
DaViT-s (WRS)	0.9900	0.8989	0.8604	0.8600	0.9902	0.8604

Table 2: Performance comparison of models on key metrics.

4 Discussion

This study addressed the multi-class anomaly classification difficulty in Video Capsule Endoscopy (VCE), with a focus on transformer model performance. While our early models, the custom CNN and ResNet50, offered a solid platform for analysis (accuracies of **0.8502** and **0.9308**, respectively), ResNet50 outperformed the ViT model, which achieved an accuracy of **0.9138**. MViT increased our results to **0.9440**, demonstrating the transformer designs' ability to handle complicated data distributions and extract the nuanced patterns required for accurate anomaly identification. DaViT small outperformed all other models, with an outstanding accuracy of **0.9487**. We applied weighted random sampling (WRS) to our DaVit model which showed an accuracy of **0.9433**, showing persistently high performance. The evaluation metrics for transformer models provide more information about their capabilities. MViT had a mean specificity of 0.9890, a mean F1-score of 0.9933, and a balanced accuracy of 0.8307, indicating that it can effectively detect both positive and negative cases. Similarly, DaViT Small achieved a mean AUC of 0.9932 and a balanced accuracy of 0.8595, demonstrating its skill in discriminating between distinct classes within the dataset. Interestingly, the addition of WRS did not result in significant performance improvements for DaViT small, implying that the model's intrinsic strengths and architectural design are critical for classification success. Finally, the findings of this study demonstrate the transformational potential of sophisticated transformer models in medical diagnostics. DaViT Small's strong performance validates its applicability for the Capsule Vision 2024 challenge and proves its ability to improve accuracy in automated diagnoses, ultimately leading to better patient outcomes and operational efficiency in healthcare. Future research should focus on these designs and their implementations in a variety of diagnostic scenarios in order to attain even greater accuracy and dependability.

5 Conclusion

In conclusion, our study on RGB medical picture classification shows how advanced transformer-based architectures, especially the DaViT model, may achieve excellent accuracy and robustness. By methodically examining a number of models, such as a custom CNN, ResNet50, Vision Transformer (ViT), mViT, and DaViT, we found that although more straightforward architectures like CNNs and ResNet offered fundamental insights, they were unable to handle intricate visual patterns in medical data. But with a accuracy of **0.9466** and notable gains in recall, precision, F1-score, and AUC metrics, DaViT turned out to be our best model. These findings demonstrate DaViT's ability to efficiently capture both local and global information, which makes it an appealing option for medical picture classification tasks.

6 Acknowledgments

As participants in the Capsule Vision 2024 Challenge, we fully comply with the competition's rules as outlined in [1]. Our AI model development is based exclusively on the datasets provided in the official release in [1].

References

[1] Palak Handa, Amirreza Mahbod, Florian Schwarzhans, Ramona Woitek, Nidhi Goel, Deepti Chhabra, Shreshtha Jha, Manas Dhir, Deepak Gunjan, Jagadeesh Kakarla, et al. Capsule vision 2024 challenge: Multi-class abnormality classification for video capsule endoscopy. *arXiv preprint arXiv:2408.04940*, 2024.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL `https://arxiv.org/abs/2010.11929`.

[4] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4814, June 2022.

[5] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 74–92. Springer, 2022.

[6] Joshua Melson, Guru Trikudanathan, Barham K. Abu Dayyeh, Manoop S. Bhutani, Vinay Chandrasekhara, Pichamol Jirapinyo, Kumar Krishnan, Nikhil A. Kumta, Rahul Pannala, Mansour A. Parsi, Amrita Sethi, Arvind J. Trindade, Rabindra R. Watson, John T. Maple, and David R. Lichtenstein. Video capsule endoscopy. *Gastrointestinal Endoscopy*, 93 (4):784–796, Apr 2021. ISSN 0016-5107. doi: 10.1016/j.gie.2020.12.001. URL `https://doi.org/10.1016/j.gie.2020.12.001`.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[8] Akihito Yokote, Junji Umeno, Keisuke Kawasaki, Shin Fujioka, Yuta Fuyuno, Yuichi Matsuno, Yuichiro Yoshida, Noriyuki Imazu, Satoshi Miyazono, Tomohiko Moriyama, Takanari Kitazono, and Takehiro Torisu. Small bowel capsule endoscopy examination and open access database with artificial intelligence: The see-artificial intelligence project. *DEN Open*, 4(1):e258, 2024. doi: https://doi.org/10.1002/deo2.258. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/deo2.258`.

[9] Anastasios Koulaouzidis, Dimitris K Iakovidis, Diana E Yung, Emanuele Rondonotti, Uri Kopylov, John N Plevris, Ervin Toth, Abraham Eliakim, Gabrielle Wurm Johansson, Wojciech Marlicz, et al. Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. *Endoscopy international open*, 5(06):E477–E483, 2017.

[10] Pia H. Smedsrud, Vajira Thambawita, Steven A. Hicks, Henrik Gjestang, Oda Olsen Nedrejord, Espen Næss, Hanna Borgli, Debesh Jha, Tor Jan Derek Berstad, Sigrun L. Eskeland, Mathias Lux, Håvard Espeland, Andreas Petlund, Duc Tien Dang Nguyen, Enrique Garcia-Ceja, Dag Johansen, Peter T. Schmidt, Ervin Toth, Hugo L. Hammer, Thomas de Lange, Michael A. Riegler, and Pål Halvorsen. Kvasir-capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1):142, May 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00920-z. URL `https://doi.org/10.1038/s41597-021-00920-z`.

[11] Nidhi Goel, Samarjeet Kaur, Deepak Gunjan, and S. J. Mahapatra. Dilated cnn for abnormality detection in wireless capsule endoscopy images. *Soft Computing*, 26(3):1231–1247, Feb 2022. ISSN 1433-7479. doi: 10.1007/s00500-021-06546-y. URL `https://doi.org/10.1007/s00500-021-06546-y`.

[12] Dev Rishi Verma, Vibhor Saxena, Dhruv Sharma, and Arpan Gupta. Capsule commandos. `https://github.com/devrishivermaa/capsule-commandos`, 2024.