Multi-Class Abnormality Classification in Video Capsule Endoscopy Using Deep Learning

Arnav Samal [⋆], Ranya [†]

[⋆] Department of Computer Science and Engineering, National Institute of Technology Rourkela, India

[†] Department of Artificial Intelligence and Data Sciences, Indira Gandhi Delhi Technical University for Women, Delhi, India

Email: samalarnav@gmail.com, ranya102btcsai22@igdtuw.ac.in

Abstract

This report outlines Team Seq2Cure's deep learning approach for the Capsule Vision 2024 Challenge, leveraging an ensemble of convolutional neural networks (CNNs) and transformer-based architectures for multi-class abnormality classification in video capsule endoscopy frames. The dataset comprised over 50,000 frames from three public sources and one private dataset, labeled across 10 abnormality classes. To overcome the limitations of traditional CNNs in capturing global context, we integrated CNN and transformer models within a multi-model ensemble. Our approach achieved a balanced accuracy of 86.34% and a mean AUC-ROC score of 0.9908 on the validation set, with significant improvements in classifying complex abnormalities. Code is available at this https URL.

1 Introduction

1.1 Background & Challenges

Gastrointestinal (GI) endoscopy is a critical diagnostic tool for identifying and managing various GI disorders, including celiac disease, gastrointestinal bleeding (GIB), esophagitis, and malignancies. Traditional endoscopy, while effective, is invasive and often requires sedation, which can be uncomfortable for patients.

Video Capsule Endoscopy (VCE), introduced in 2001, offers a non-invasive alternative. Patients swallow a small, wireless capsule that captures thousands of images as it moves through the GI tract, allowing for detailed visualization of small intestinal disorders like bleeding, erosions, and ulcers without the need for sedation.

Despite its advantages, VCE presents challenges, primarily due to the vast volume of data generated. Each procedure produces thousands of images, requiring significant time and expertise to analyze. The rapid movement of the capsule often leads to complex and stochastic imagery, increasing the likelihood of missed diagnoses. These challenges highlight the need for automated analysis tools to support efficient and accurate interpretation of VCE data.

To address these challenges, the Capsule Vision 2024 Challenge [1] was launched. The aim of the challenge is to provide a platform for the development, testing, and evaluation of AI models for automatic classification of abnormalities captured in VCE frames.

1.2 Previous Works

Recent studies have explored the application of artificial intelligence and machine learning tools (AIMLT) to enhance the efficiency and accuracy of VCE image analysis. Various research groups have employed convolutional neural networks (CNNs) to tackle challenges in VCE. Xie et al. [2] trained a CNN on nearly 3,000 capsule studies, achieving a significant increase in the detection rate of small bowel pathologic findings while reducing reading time by 89.3% compared to conventional human reading, though the study faced limitations due to the heterogeneity of VCE readers involved.

Building on this foundation, Afonso et al. [3] developed a CNN specifically focused on identifying ulcers and erosions in the small bowel, achieving impressive results with a sensitivity of 90.8%, specificity of 97.1%, and an overall accuracy of 95.6%. Further advancing this line of research, Ferreira et al. [4] demonstrated similar success in ulcer detection, achieving an accuracy of 92.4%.

1.3 Limitations and Proposed Solutions

Previous research in VCE analysis has encountered several key challenges. The need for extensive pre-processing and labeling of VCE images to create reliable training datasets remains resource-intensive. Additionally, traditional CNN-based approaches struggle with capturing long-range dependencies due to fixed receptive fields, limiting their ability to detect subtle abnormalities in VCE frames. The high computational requirements of advanced machine learning models further hinder their use in clinical settings.

Our project addresses these challenges by implementing a multi-model ensemble approach that combines CNN and transformer architectures. This approach enhances the model's ability to capture global dependencies and contextual information, leveraging the strengths of both architectures to improve classification accuracy and robustness.

2 Methodology

2.1 Dataset

The dataset [5] was developed using a combination of three publicly available sources—SEE-AI project, KID, and Kvasir-Capsule—along with a private dataset from AIIMS. In total, the training set comprised 37,607 frames, and the validation set included 16,132 frames, each annotated with one of 10 abnormality classes: angioectasia, bleeding, erosion, erythema, foreign body, lymphangiectasia, polyp, ulcer, worms, and normal. For both training and inference phases, we utilized the same dataset, which has been made publicly available on Kaggle through the following links: training dataset and inference dataset

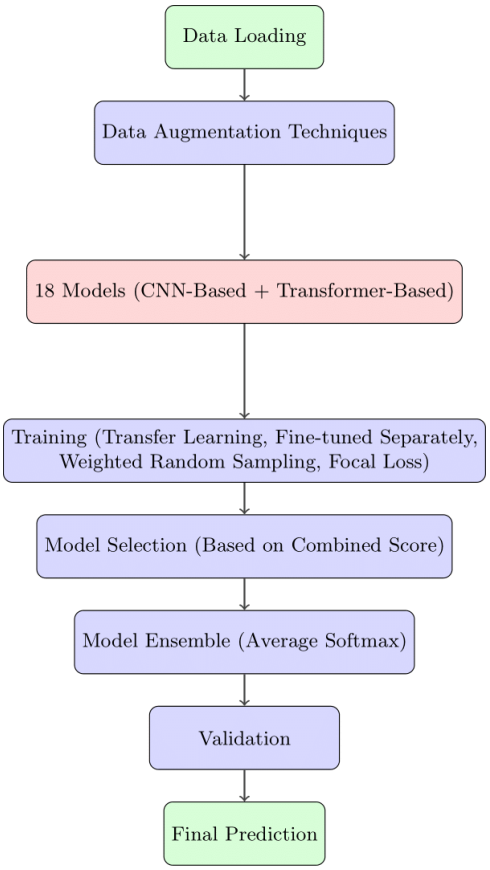

Figure 1: Block diagram of the developed pipeline.

2.2 Data Preprocessing

To enhance model generalization, a comprehensive data augmentation pipeline was implemented, artificially expanding the training dataset through several spatial and intensity transformations:

2.2.1 Spatial transformations

- **Resize & Flip:** Images were resized to (224, 224) pixels, with a 50% chance of horizontal flip and a 30% chance of vertical flip.
- **Rotation & Affine:** Random rotations up to $\pm 15^{\circ}$, with affine transformations including translation (up to 10%) and scaling (90-110%).
- **Perspective Distortion:** A 50% chance of applying perspective distortion to simulate variations in camera angles.

2.2.2 Intensity transformations

- **Color Jitter & Normalization:** Adjusted brightness, contrast, saturation, and hue, followed by normalization (mean and std [0.5, 0.5, 0.5]).

- **Random Erasing & Gaussian Blur:** A 20% chance of erasing parts of the image and a 30% chance of applying Gaussian blur for robustness.

2.3 Model Architecture

This project employs a multi-model ensemble approach that integrates both traditional Convolutional Neural Network (CNN) architectures and transformer-based models for effective video capsule endoscopy (VCE) frame classification. This ensemble method aims to leverage the strengths of each model type, enhancing overall classification performance.

Traditional CNN Models

The following CNN architectures are utilized due to their proven efficacy in image classification tasks:

- **EfficientNet:** Optimizes the network architecture by balancing depth, width, and resolution through a compound scaling method [6].
- **ResNet:** Known for its deep residual learning framework, which mitigates the vanishing gradient problem, enabling the training of very deep networks [7].
- **MobileNetV3:** Designed for mobile and edge devices, focusing on high efficiency and performance with limited computational resources [8].
- **RegNet:** A family of models emphasizing design simplicity and efficiency with a straightforward, regularized architecture [9].
- **DenseNet:** Utilizes dense connectivity patterns to improve feature propagation and reduce the number of parameters [10].
- **InceptionV4:** Enhances classification performance through a multi-path architecture incorporating various convolutional operations in parallel [11].
- **ResNeXt:** Combines the strengths of ResNet with the cardinality concept to create models that achieve improved accuracy and efficiency [12].
- **WideResNet:** Focuses on widening layers rather than deepening them to enhance performance with fewer layers [13].
- **MNASNet:** Optimized for mobile devices, using a reinforcement learning approach for architectural search [14].
- **SEResNet50:** Integrates Squeeze-and-Excitation blocks into the ResNet architecture, improving representational capacity [15].
- **ConvNeXt:** A modernized CNN architecture that adopts concepts from transformers to enhance performance [16].

Transformer-based Models

The following transformer architectures are employed for their advanced capabilities in handling image data:

- **Vision Transformer (ViT):** Processes images as sequences, leveraging the self-attention mechanism to capture global dependencies effectively [17].
- **Swin Transformer:** Introduces hierarchical feature maps and shifted windows to improve computational efficiency and performance in vision tasks [18].
- **DeiT (Data-efficient Image Transformers):** Designed to work effectively with fewer data samples while maintaining competitive performance [19].
- **BEiT (Bidirectional Encoder Representation from Image Transformers):** Integrates a vision transformer with masked image modeling to provide rich contextual representations [20].
- **CaiT (Class-Attention in Image Transformers):** Enhances the representation of image classes using attention mechanisms within the transformer framework [19].
- **TwinsSVT (Spatially Separable Vision Transformer):** Utilizes spatial separability to improve efficiency while retaining high performance [21].
- **EfficientFormer:** Combines efficiency and performance by focusing on the optimization of transformer structures for image classification tasks [22].

The model parameters are available at this https URL.

2.4 Training

To optimize performance and mitigate overfitting, the following techniques and configurations were applied:

2.4.1 Transfer Learning & Fine-tuning

Pre-trained models were used for transfer learning, with **full model fine-tuning** performed on the video capsule endoscopy (VCE) dataset. All layers were fine-tuned to adapt the networks to domain-specific features, ensuring effective learning from the VCE data.

2.4.2 Early Stopping

Early stopping [23] was employed with a patience of 5 epochs, halting training if validation performance did not improve for 5 consecutive epochs. This approach prevented overfitting and unnecessary training cycles.

2.4.3 Optimizer & Training Configuration

The models were optimized using the **AdamW optimizer** [24], with the following settings:

- **Learning rate:** 1e-4
- **Weight decay:** 0.05

The training process involved:

- **Batch size:** 32

- **Epochs:** 20
- **Hardware:** 4 Nvidia Tesla P100 GPUs (16 GB memory each)
- **Training time:** 15-16 hours

2.5 Class Imbalance Mitigation

To address class imbalance, the following techniques were applied:

2.5.1 Weighted Random Sampling

Weighted Random Sampling [25] was used to ensure more frequent sampling of underrepresented classes, reducing bias toward overrepresented classes. The sampling weight w_i for class i was defined as:

$$w_i = \frac{1}{N_i}$$

where N_i is the number of samples in class i. This ensures that the probability of selecting samples from underrepresented classes is higher, promoting a balanced representation of all classes during training.

2.5.2 Focal Loss

Focal Loss [26] was applied to prioritize difficult-to-classify examples, effectively down-weighting well-classified samples and focusing more on hard examples. The focal loss for a binary classification task is defined as:

$$\mathcal{L}_{\text{FL}}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

where:

- p_t is the predicted probability for the true class.
- α_t is a balancing factor for the class (used to address class imbalance).
- γ is the focusing parameter, which adjusts the rate at which easy examples are down-weighted. A common value is $\gamma = 2$.

This loss function ensures that harder examples receive more focus by reducing the contribution of well-classified samples, effectively handling the class imbalance problem.

2.6 Model Selection & Checkpointing

Model selection was based on a balanced evaluation metric that combined both **Balanced Accuracy** and **Mean AUC Score**, ensuring robust performance across all abnormality classes. This approach mitigated overfitting to dominant classes and enhanced the model's ability to detect rare abnormalities.
The best models were saved using a **Combined Score**, defined as:

$$\text{Combined Score} = \frac{\text{Balanced Accuracy} + \text{Mean AUC Score}}{2}$$

A tolerance of 1e-4 was used to track performance improvements, ensuring that only models demonstrating substantial gains were checkpointed, thus preserving the most optimal versions.

2.7 Model Ensembling

To implement the ensemble inference for the validation and test dataset, we used the following procedure, which leverages multiple models to generate predictions through softmax probability averaging. The pseudocode for the ensemble inference is as follows:

```
function ensemble(models, dataloader, device):
    all_predictions = []
    all_image_paths = []

    # Set models to evaluation mode
    for model in models:
        model.eval()

    # Inference with ensemble
    for (X, image_paths) in dataloader:
        X = X.to(device)

        # Compute average softmax predictions across models
        ensemble_preds = mean(softmax(model(X)) for model in models)

        all_predictions.append(ensemble_preds.cpu())
        all_image_paths.extend(basename(path) for path in image_paths)

    predictions = concatenate(all_predictions)
    return predictions, all_image_paths
```

This method ensures that predictions from each model are combined, enhancing the overall robustness of the inference process through softmax probability averaging.

3 Results

The evaluation of the ensemble model on the validation dataset demonstrates strong overall performance in classifying video capsule endoscopy (VCE) images. The model achieved a balanced accuracy of 0.8634 and a mean AUC-ROC of 0.9908, indicating its effectiveness in distinguishing various abnormalities. Notably, the model excelled in identifying worms, normal instances, and ulcers. However, it revealed areas for improvement, particularly in classifying erythema and polyps, which exhibited lower F1 scores. For detailed performance metrics, including class-wise precision, recall, specificity, and AUC-ROC scores, refer to here.

3.1 Achieved Results on the Validation Dataset

The performance of the ensemble model was assessed on the validation dataset, yielding the key metrics outlined in Table 1.

Class	Precision	Recall	F1-Score	Specificity
Angioectasia	0.87	0.84	0.85	0.99
Bleeding	0.87	0.85	0.86	0.99
Erosion	0.80	0.80	0.80	0.98
Erythema	0.69	0.64	0.66	0.99
Foreign Body	0.86	0.92	0.89	0.99
Lymphangiect.	0.85	0.93	0.89	0.99
Normal	0.98	0.98	0.98	0.94
Polyp	0.78	0.73	0.75	0.99
Ulcer	0.98	0.93	0.96	1.00
Worms	0.99	1.00	0.99	1.00
Macro Avg	0.87	0.86	0.86	0.99

Table 1: Performance Metrics for Each Class

Method	Avg. Acc.	Avg. Prec.	Avg. AUC	Avg. Rec.	Avg. F1	Bal. Acc.
SVM (baseline)	0.82	0.83	0.94	0.41	0.49	0.41
ResNet50 (baseline)	0.76	0.60	0.87	0.32	0.37	0.32
VGG16 (baseline)	0.69	0.52	0.92	0.54	0.48	0.54
Custom CNN (baseline)	0.46	0.10	0.31	0.10	0.09	0.10
Ensemble Model (Ours)	**0.946**	**0.905**	**0.990**	**0.863**	**0.864**	**0.863**

Table 2: Validation results and comparison to baseline methods from the Capsule Vision 2024 challenge.

4 Discussion

Based on our previous studies on medical image classification using CNN, we introduced a methodology that implements a multi-model ensemble approach for video capsule endoscopy frame classification. This approach combines the strengths of various CNN and transformer architectures to enhance the classification accuracy of VCE images.

4.1 Rationale for using multi-model ensemble approach

The decision to implement a multi-model ensemble approach for video capsule endoscopy (VCE) frame classification was driven by several factors. This approach enhances accuracy and robustness by leveraging various CNN and transformer architectures. It improves classification performance by combining diverse model capabilities to capture a broader range of features and reduce errors through averaged predictions, thereby increasing output reliability.

4.2 Methodological Considerations

Training models using GPUs incurred significant computational costs, presenting barriers to efficient model development. Although the models were pretrained, they were often trained on large datasets that contained only a few similar samples to the target dataset, limiting the effectiveness of transfer learning.

Advanced training techniques like Knowledge Distillation, K-Fold Cross-Validation, and Self-Supervised Pre-training methods such as Masked Auto-Encoding and Contrastive Learning can enhance model performance and generalization.

4.3 Challenges and Future Directions

Despite promising results, challenges persist in applying AI in VCE. The dependence on extensive human annotation for training datasets is labor-intensive, potentially hindering clinical integration. We aim to develop a user-friendly annotation tool to enhance efficiency in dataset preparation and improve supervised CNN accuracy. Additionally, addressing the "black box" nature of AI remains essential; incorporating visual representations of areas of interest, like Grad-CAMs, can enhance model explainability. Finally, the imbalanced nature of VCE datasets calls for innovative approaches, such as few-shot learning, to mitigate the impact of limited pathology examples.

5 Conclusion

In conclusion, our study highlights the potential of employing multiple CNNs to enhance the accuracy of lesion detection in the gastrointestinal tract using video capsule endoscopy, despite challenges associated with data limitations. The implementation of our AI approach offers promising pathways for improving diagnostic accuracy while alleviating burdens on medical professionals. Future research should focus on validating our approach across diverse datasets and refining our annotation tools for broader clinical application.

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 [5].

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] Weiling Xie, Yangtian Zhao, Peng Wang, and Zhiwei Zhang. Real-time small bowel disease detection and classification using capsule endoscopy with deep learning: A multi-center study. *Gastrointestinal Endoscopy*, 91(6):AB301, 2020. doi: 10.1016/j.gie.2020.03.1918.

[3] João Afonso, Carla Rolanda, Raquel Gonçalves, and Marta Silva. Deep learning for small bowel capsule endoscopy: A systematic review and meta-analysis. *Gastrointestinal Endoscopy*, 90(4):668–679, 2019. doi: 10.1016/j.gie.2019.06.018.

[4] Francisco Ferreira, Jorge Sousa, and Miguel Mascarenhas-Saraiva. Artificial intelligence system for detection and classification of intestinal ulcers in capsule endoscopy. *Digestive Diseases and Sciences*, 66(8):2714–2721, 2021. doi: 10.1007/s10620-020-06634-3.

[5] Palak Handa, Amirreza Mahbod, Florian Schwarzhans, Ramona Woitek, Nidhi Goel, Deepti Chhabra, Shreshtha Jha, Manas Dhir, Deepak Gunjan, Jagadeesh Kakarla, and Balasubramanian Raman. Training and validation dataset of capsule vision 2024 challenge. *Figshare*, 7 2024. doi: 10.6084/m9.figshare.26403469.v1. URL https://figshare.com/articles/dataset/Training_and_Validation_Dataset_of_Capsule_Vision_2024_Challenge/26403469.

[6] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, and Mingxing Tan. Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019.

[9] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10415–10424, 2020.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[11] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[12] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017.

[13] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[14] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2820–2828, 2019.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[16] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. Convnext: Revisiting convolutions for vision. *arXiv preprint arXiv:2201.03545*, 2022.

[17] 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. *arXiv preprint arXiv:2010.11929*, 2020.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[19] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2021.

[20] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.09785*, 2021.

[21] Xiangxiang Zhang, Furu Wei, and Li Dong. Twins: Revisiting the design of spatial attention in vision transformers. *arXiv preprint arXiv:2104.13840*, 2021.

[22] Tingting Lin, Yixuan Wang, Xiaoxi Liu, and Xiaolong Qiu. Efficientformer: Vision transformers in the real world. *arXiv preprint arXiv:2106.13319*, 2021.

[23] Lutz Prechelt. Automatic early stopping using cross-validation: quantifying the criteria. In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019.

[25] Joanna Byrd and Zachary C. Lipton. What is the effect of importance weighting in deep learning? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 872–881. PMLR, 2019.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.