Multi-Class Abnormality Detection for Video Capsule Endoscopy Using EfficientViT, Weighted Loss, and Weighted Sampling Techniques

Girin Chutia

HCL Technologies Ltd.

Email: girin.iitm@gmail.com

Abstract

This report presents an approach to the Capsule Vision 2024 Challenge, which focuses on multi-class abnormality classification in video capsule endoscopy (VCE) frames. The challenge aims to develop AI models capable of automatically classifying abnormalities captured in VCE video frames into ten distinct categories. To address this, a deep learning-based solution was developed. The model was trained on a diverse dataset comprising 37,607 VCE frames and validated on an additional 16,132 frames. Through extensive hyperparameter tuning, data augmentation, and sampling techniques, the model achieved a mean AUC of 0.98 and a balanced accuracy of 0.83 on the validation dataset. These results demonstrate the effectiveness of this approach in accurately classifying VCE abnormalities, potentially reducing the diagnostic burden on clinicians and improving patient outcomes.

1 Introduction

Video Capsule Endoscopy (VCE) has revolutionized the non-invasive visualization of the gastrointestinal (GI) tract, particularly for diagnosing small bowel-related diseases. This technology has gained importance due to the global increase in GI and liver diseases, attributed to changing environmental factors, nutrition, and antibiotic use. VCE offers several advantages over conventional endoscopy, including its non-invasive nature, the absence of sedation-related complications, and an improved ability to detect GI tract anomalies. However, VCE's full potential is currently limited by several factors:

- Time-consuming analysis: A typical VCE procedure generates thousands of frames, requiring hours of retrospective analysis by experienced gastroenterologists.
- Human bias and false-positive rates: The analysis is subject to errors due to factors such as bubbles, debris, and intestinal fluid.
- Inadequate doctor-to-patient ratio: This global issue further delays the analysis process.
- Hardware limitations: Challenges include capsule retention, battery limitations, and bowel obstructions.

To address these challenges, there is growing interest in leveraging Artificial Intelligence (AI) for VCE technology, particularly in abnormality classification. The development of robust, user-friendly, and interpretable AI-based models for multi-class abnormality classification has the potential to reduce the burden on gastroenterologists, decrease inspection time, and maintain high diagnostic precision.

This work aims to present an AI-based solution for multi-class abnormality classification in VCE frames. By doing so, we seek to contribute to the advancement of VCE technology and improve the efficiency of GI tract disease diagnosis and management.

2 Methods

2.1 Data Description

The dataset provided for the challenge consists of 37,607 training and 16,132 validation VCE frames, mapped to 10 class labels : angioectasia, bleeding, erosion, erythema, foreign body, lymphangiectasia, polyp, ulcer, worms, and normal. The test dataset consists of 4,385 images. The class distribution of the training and validation datasets is shown in Figure 1.

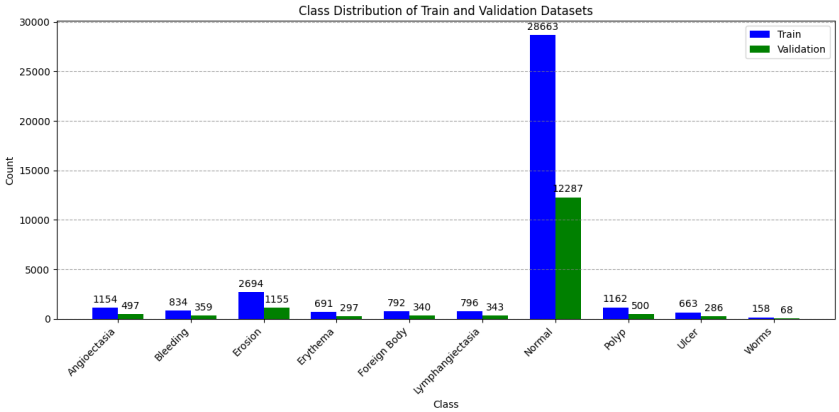


Figure 1: Training and Validation Class distribution

2.2 Preprocessing and Data Augmentations

For training the model, the images were resized to 224x224 pixels. Normalization was done using a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225], following ImageNet standards to ensure compatibility with pre-trained models for better performance. Data augmentation techniques, including RandomHorizontalFlip, RandomVerticalFlip, RandomErasing (probability 0.1), RandomAdjustSharpness (sharpness factor 1.2, probability 0.1), RandomRotation (0 to 180 degrees), and RandomAutocontrast (probability 0.1) and MixUp [1] (alpha=0.2, probability 0.3) were applied to improve model robustness. Figure 2 shows the effect of each of these augmentation on a sample training image :

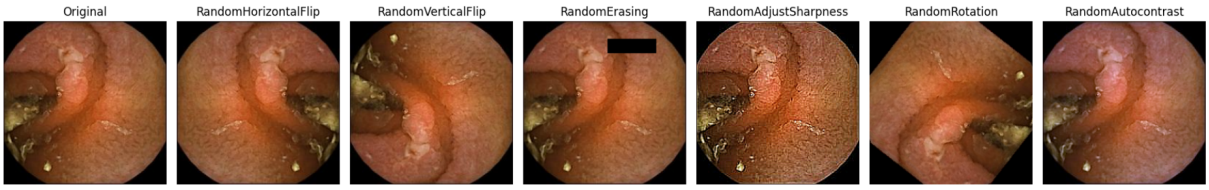


Figure 2: Illustration of RandomHorizontalFlip, RandomVerticalFlip, RandomErasing, RandomAdjustSharpness, RandomRotation and RandomAutocontrast augmentations

Figure 3: Illustration of MixUp[1] augmentation : Image 1 and Image 2 are mixed up with alpha=0.2

2.3 Model

EfficientViT-l2[2] is used to develop the solution for the challenge. In this approach, the backbone of the model is kept fixed, and only the classification head is trained. EfficientViT [2] has emerged as a notable family of models that introduce a novel lightweight multiscale attention mechanism. Unlike previous ViT models that rely on computationally intensive self-attention, large-kernel convolutions, or complex topologies, EfficientViT achieves a global receptive field and multi-scale learning through lightweight and hardware-efficient operations achieving state of the art results. An illustration of building blocks of EfficientViT architecture are shown in Figure 4

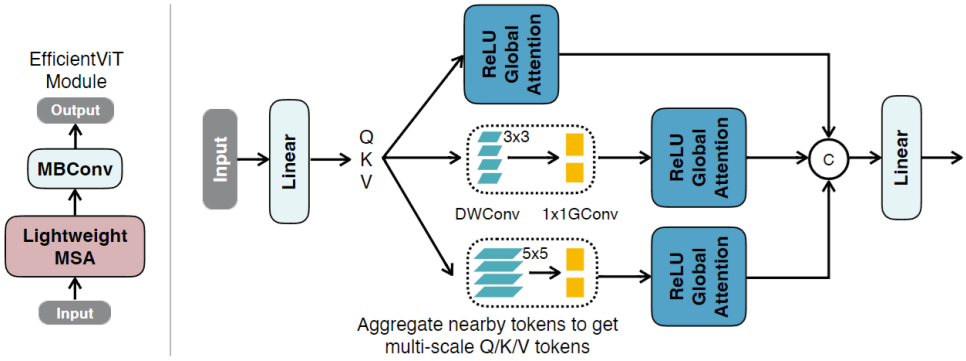


Figure 4: Illustration of EfficientViT's building blocks. The left side depicts a building block comprising a lightweight Multi-Scale Attention (MSA) module and an MBConv[3] layer, where the MSA captures context information and the MBConv[3] captures local information. The right side shows the Lightweight Multi-Scale Attention process, which involves generating multi-scale tokens from Q/K/V tokens via small-kernel convolutions, applying ReLU-based global attention, and concatenating outputs for feature fusion through a final linear projection layer.

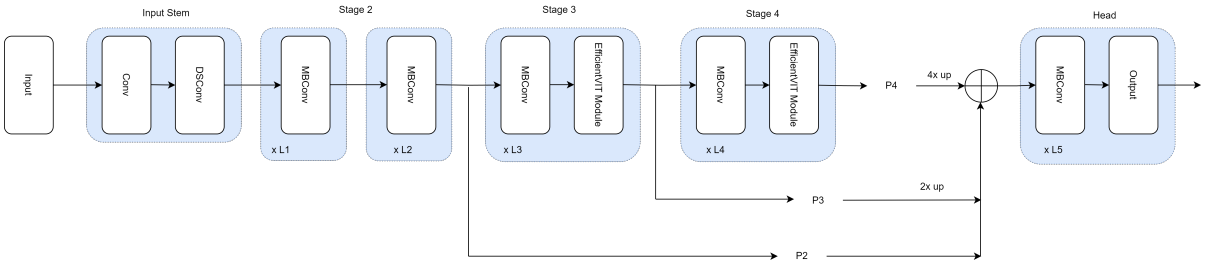


Figure 5: Illustration of the macro structure of EfficientViT. For the EfficientViT-l2 model : L1 = 1, L2 = 2, L3 = 2, L4 = 8, and L5 = 8. The EfficientViT architecture incorporates MBConv[3] and DSConv[4] blocks.

2.4 Training Pipeline

The model was trained for 26 epochs using the AdamW optimizer [5] with an initial learning rate of 0.0001. A batch size of 32 was used. To address the class imbalance in the dataset, a weighted sampler was employed for data sampling during training, and weighted cross-entropy loss was used as the loss function. The training pipeline flow is shown in Figure 6. The model was trained on an NVIDIA GeForce GTX 1660 Ti (6 GB VRAM) GPU.

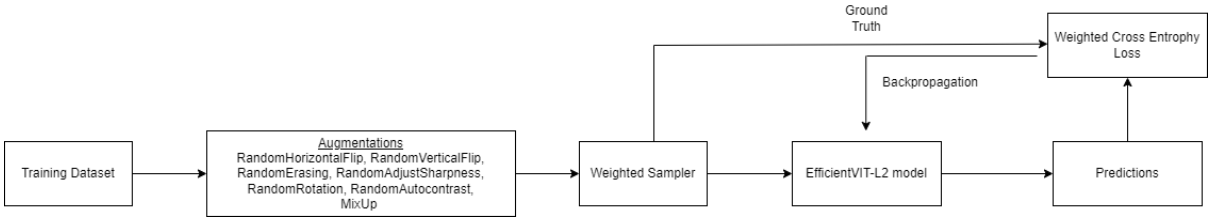


Figure 6: Training Pipeline.

2.4.1 Weighted Sampler

A weighted sampler is a crucial tool for addressing class imbalance, particularly when certain classes are underrepresented. By assigning higher sampling probabilities to minority classes, the weighted sampler ensures that these classes are more frequently represented during training. This approach helps prevent models from becoming biased towards majority classes, which can dominate the learning process in imbalanced datasets. The sampling probabilities are calculated using a power transformation given by:

$$\text{weight}_i = \left(\frac{1}{n_i}\right)^{\alpha} \tag{1}$$

where α is the power transformation parameter. In our case, $\alpha = 0.35$. α controls how aggressively the weighted sampler addresses class imbalance, balancing between maintaining class diversity and avoiding over-emphasis on minority classes. n_i = number of training images of class i. The weights are then normalized to sum to 1:

$$\text{normalized\_weight}_i = \frac{\text{weight}_i}{\sum_j \text{weight}_j} \tag{2}$$

These normalized weights are used to sample the dataset, ensuring that minority classes are more frequently represented during training.

2.4.2 Weighted Cross Entropy Loss

Weighted cross-entropy loss is widely used to address class imbalance in classification tasks. By assigning different weights to each class, it ensures that underrepresented classes have a greater influence on the loss calculation. This prevents the model from becoming overly biased towards majority classes, which could dominate the learning process. The loss function penalizes misclassifications of minority classes more heavily through these weights, leading to improved performance on imbalanced datasets. The weights are calculated using the power transformation as discussed in Section 2.4.1. The weighted cross-entropy loss for a multi-class classification problem is given by:

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

Where:

- L is the loss.
- N is the number of classes.
- w_i is the weight for class i.
- y_i is the true label for class i (1 if the class is correct, otherwise 0).
- $\hat{y}_i$ is the predicted probability for class i.

2.5 Evaluation Metrics

The primary metrics used for evaluation were the mean AUC and balanced accuracy. The mean AUC (Area Under the ROC Curve) is a metric that summarizes the model's performance across all classification thresholds. For multi-class classification, the AUC can be computed using a one-vs-all approach, where the AUC is calculated for each class separately and then averaged. Balanced accuracy is particularly useful for handling imbalanced datasets, as it accounts for the unequal distribution of classes. It is the average of recall obtained on each class, ensuring that the performance on minority classes is not overshadowed by the majority classes.

3 Results

The performance results on training and validation datasets are shown in Table 1.

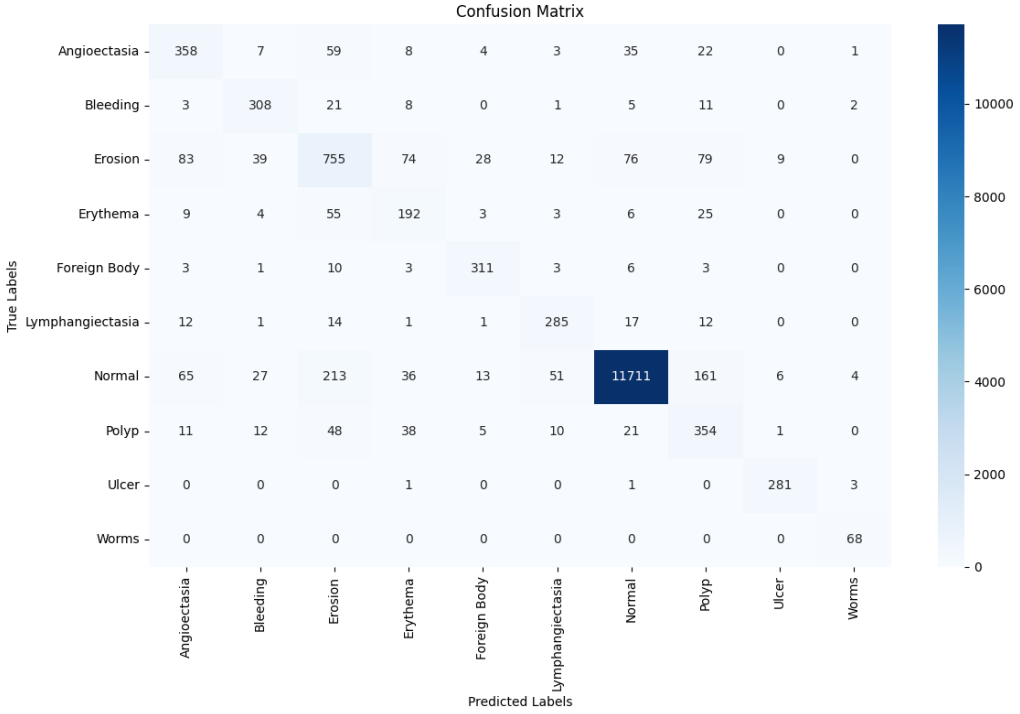
Table 1: Model Performance Metrics

Model	Training Dataset Mean AUC	Validation Dataset Mean AUC	Training Dataset Balanced Accuracy	Validation Dataset Balanced Accuracy
EfficientViT-l2	1	0.98	0.95	0.83

Confusion matrix and classification report of the trained model on the validation dataset are shown in Figure7,

	precision	recall	f1-score	support
Angioectasia	0.66	0.72	0.69	497
Bleeding	0.77	0.86	0.81	359
Erosion	0.64	0.65	0.65	1155
Erythema	0.53	0.65	0.58	297
Foreign Body	0.85	0.91	0.88	340
Lymphangiectasia	0.77	0.83	0.80	343
Normal	0.99	0.95	0.97	12287
Polyp	0.53	0.71	0.61	500
Ulcer	0.95	0.98	0.96	286
Worms	0.87	1.00	0.93	68
accuracy			0.91	16132
macro avg	0.76	0.83	0.79	16132
weighted avg	0.92	0.91	0.91	16132

(a) Classification Report (Validation dataset)



(b) Confusion Matrix (Validation Dataset)

Figure 7: Confusion matrix and Classification report on validation dataset.

4 Discussion

The results indicate that EfficientViT-l2 performs very well on the VCE multi-class abnormality classification task, achieving a mean AUC of 0.98 and a balanced accuracy of 0.83 on the validation dataset. This performance can be attributed to the EfficientViT-l2 model [2], along with the use of weighted sampling, weighted cross-entropy loss, and augmentations such as MixUp [1] in the solution development.

5 Conclusion

This study proposes a solution for multi-class abnormality classification in Video Capsule Endoscopy (VCE), aimed at enhancing the diagnosis of gastrointestinal conditions. Leveraging deep learning techniques, the developed model demonstrates superior performance in classifying various abnormalities. By utilizing a curated dataset along with effective data augmentation and sampling techniques, the model generalizes well across different video frames, ensuring reliable classifications and potentially improving patient outcomes.

6 Acknowledgments

As participants in the Capsule Vision 2024 Challenge, I fully comply with the competition's rules as outlined in [6]. My AI model development is based exclusively on the datasets provided in the official release in [7].

References

[1] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2017. URL `https://api.semanticscholar.org/CorpusID:3162051`.

[2] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17302–17313, 2023.

[3] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. URL `https://api.semanticscholar.org/CorpusID:4555207`.

[4] Marcelo Gennari, Roger Fawcett, and Victor Adrian Prisacariu. Dsconv: Efficient convolution operator. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5147–5156, 2019. URL `https://api.semanticscholar.org/CorpusID:57573783`.

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL `https://api.semanticscholar.org/CorpusID:53592270`.

[6] 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.

[7] 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. *Fishare*, 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.