Capsule Vision Challenge 2024: Multi-Class Abnormality Classification for Video Capsule Endoscopy

Aakarsh Bansal [a], Bhuvanesh Singla [a], Raajan Rajesh Wankhade [a], Dr. Nagamma Patil [a]

[a] Dept. of Information Technology, National Institute of Technology, Karnataka

Email: raajanrajeshwankhade@gmail.com

Abstract

This study presents an approach to developing a model for classifying abnormalities in video capsule endoscopy (VCE) frames. Given the challenges of data imbalance, we implemented a tiered augmentation strategy using the albumentations library to enhance minority class representation. Additionally, we addressed learning complexities by progressively structuring training tasks, allowing the model to differentiate between normal and abnormal cases and then gradually adding more specific classes based on data availability. Our pipeline, developed in PyTorch, employs a flexible architecture enabling seamless adjustments to classification complexity. We tested our approach using ResNet50 and a custom ViT-CNN hybrid model, with training conducted on the Kaggle platform. This work demonstrates a scalable approach to abnormality classification in VCE.

1 Introduction

Video capsule endoscopy (VCE) is a minimally invasive diagnostic technique that involves swallowing a small, camera-equipped capsule to capture images of the gastrointestinal (GI) tract as it passes through. This procedure allows for detailed visualization of the small intestine, an area challenging to reach with traditional endoscopic methods.

VCE has proven useful for identifying various GI abnormalities, including bleeding, erosions, angioectasia, polyps, and foreign bodies, among others. Despite its benefits, VCE produces a vast amount of image data, which requires time-consuming analysis by medical professionals to detect abnormalities effectively. Given these challenges, there is growing interest in developing AI-based systems to automatically classify abnormalities in VCE images, thereby aiding clinicians in their diagnostic efforts.

The huge amount of data requires careful labeling by medical experts which adds to the cost. There is also a concern of privacy while sharing these images. All these factors lead to formation of a dataset which is imbalanced towards certain classes and rarity of certain diseases increases the dataset's skewness which makes it difficult to be used for training. To address this, we applied a systematic approach to data augmentation using the albumentations library [1]. Augmentation techniques were grouped into three levels, Heavy, Medium, and Light, based on their intensity. These augmentations included

transformations such as horizontal and vertical flips, rotations, color jittering, and Gaussian blur. By expanding the diversity of minority-class images through this approach, we aimed to balance the training data distribution and enhance model performance on underrepresented classes.

Additionally, we structured the learning process to begin with a straightforward task, classifying Normal vs. Abnormal cases. We gradually increased task complexity by adding specific abnormal classes one at a time. This progressive approach allowed the model to build a strong foundational understanding before tackling more complex, data-scarce classes. Experiments with both ResNet50 [2] and a custom ViT-CNN hybrid architecture were conducted. This strategy highlights an adaptable framework for developing AI-based VCE classification models capable of addressing data imbalance and learning challenges, fostering improved generalization across VCE datasets.

2 Methods

Our methodology mainly revolves around two aspects:

- Addressing the pre-existing imbalance in the train dataset.
- Addressing learning issues that the model might face when learning on imbalanced data.

2.1 Addressing Data Imbalance

The dataset in Handa et al. was heavily imbalanced, containing 28,663 Normal images for training and only 158 images for the Worms class. All abnormal classes had less than 3,000 images compared to 28,663 for the Normal class.

To address this imbalance, we developed three levels of augmentations adapted from the albumentations library:

- Heavy Augmentations
  - HorizontalFlip
  - VerticalFlip
  - RandomRotate90
  - ColorJitter
  - ShiftScaleRotate
  - GaussianBlur
- Medium Augmentations
  - HorizontalFlip
  - VerticalFlip
  - RandomRotate90
  - ColorJitter
- Light Augmentations
  - HorizontalFlip

– VerticalFlip

Class Distribution in train set before the augmentations are shown in Figure 1.

Class Distribution in train set before augmentations, excluding the normal class, has been shown in Figure 2

Class Distribution in train set after augmentations, excluding Normal class, has been shown in Figure 3.

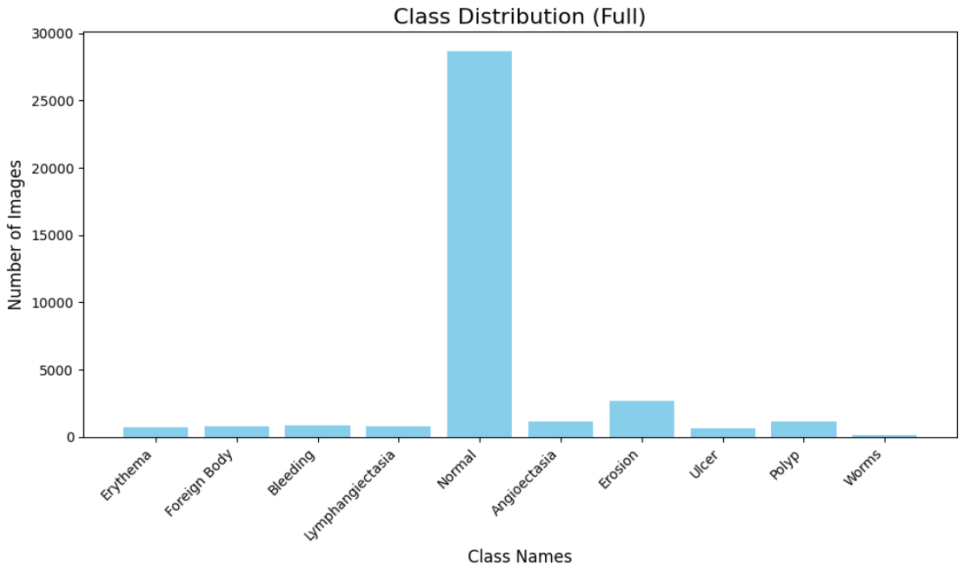

Figure 1: Pre-augmentation Class Distribution (with Normal class)

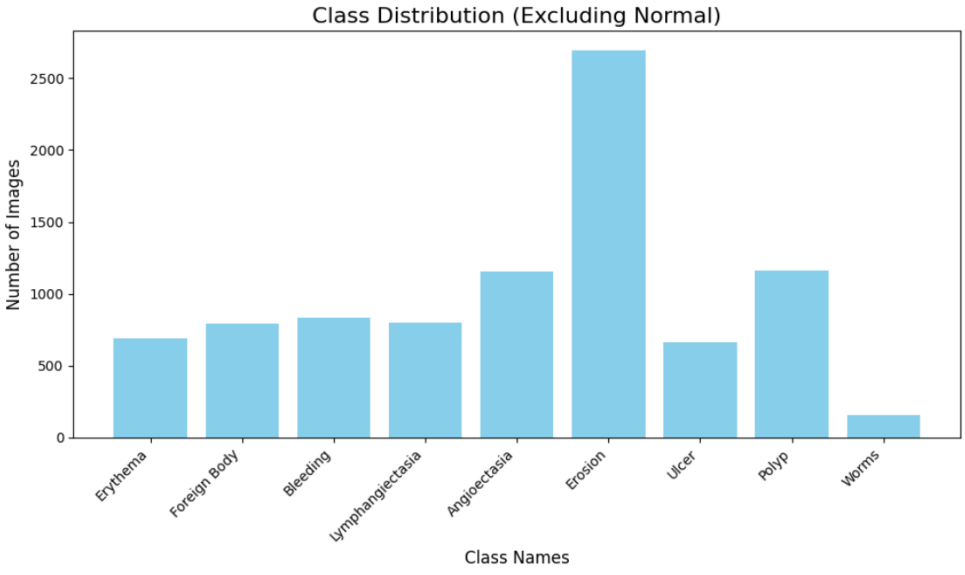

Figure 2: Pre-augmentation Class Distribution (without Normal class)

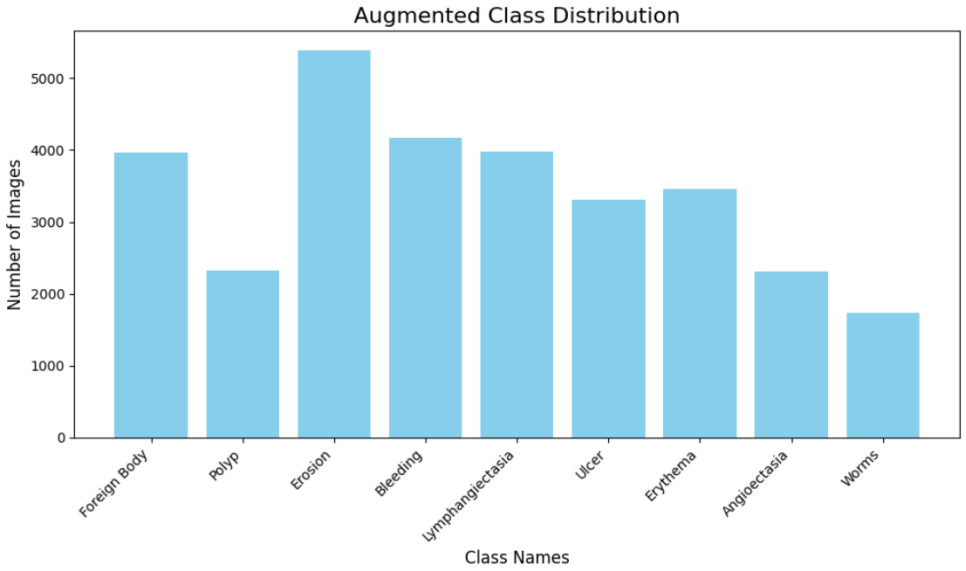

Figure 3: Post-augmentation Class Distribution (without Normal class)

2.2 Addressing Learning Challenges

When working with an imbalanced dataset, it becomes difficult for the model to learn effectively. To handle this, we used ideas from [4] and adapted them for VCE Image Classification. The main idea is to start with the simplest task, which is to classify Normal vs. Abnormal cases, and then gradually introduce one abnormal class at a time in each step. A similar method is used in [5], where a "harder" task is based on how much annotators agree. In our case, however, we define a "harder" task by the number of training images available, with the class with the fewest images being the hardest.

By slowly increasing the complexity of the task, beginning with the easiest, we aim to help the model learn better in the presence of imbalanced data.

2.3 Final Pipeline and Implementation

The final pipeline has been implemented in PyTorch. We have defined a custom dataloader that is flexible in handling the classes on which the model is being trained. This allows us to simply change the classification head of the model, retain the backbone and introduce new, harder classes in the same code.

We have experimented with the ResNet50 model and a custom ViT-CNN Hybrid (Figure 4) architecture. The hybrid architecture consists of a ViT branch [6] and a ResNet34 branch using which features are extracted. These features are then passed through a classification head to obtain the output. We have used the Adam [7] optimizer with a learning rate of 1e-3. The models were trained on the Kaggle platform using the P100 GPU.

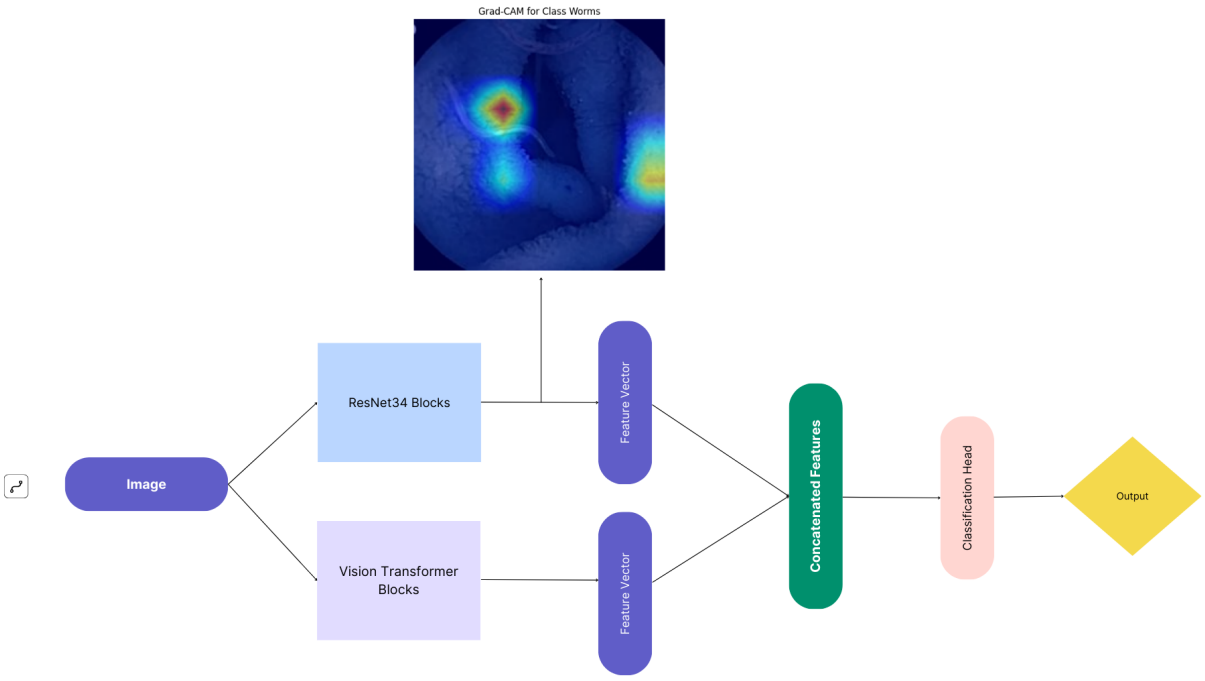

Figure 4: ViT-CNN Architecture and GradCam

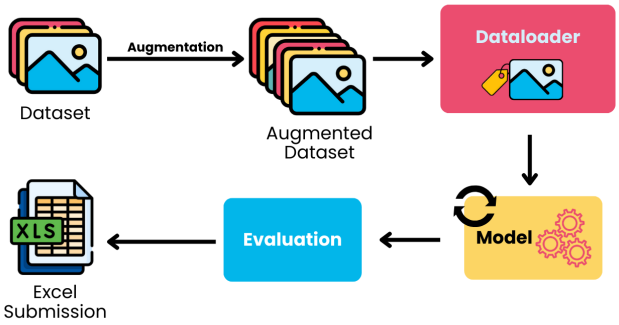

Figure 5: Block diagram of the developed pipeline.

2.4 Explainabiltiy

Although applications of AI in medical imaging are increasing, it is still mostly represented as a blackbox [8]. Hence, it becomes necessary to have some understanding behind the model's decisions. There are several ways of doing this, some methods add explainability from the beginning of training (ante-hoc) while some post training (post-hoc). Since we are using pretrained models with fine tuning, we have implemented a post-hoc technique called Gradient-weighted Class Activation Mapping (GradCAM) [9]. This technique, as per Varam et al. has performed well for medical imaging. Therefore, we applied this technique to the CNN (pretrained ResNet34) branch of our hybrid ViT-CNN model in

order to gain insights behind the decisions taken by the model, an example is shown in Figure 4.

3 Results

The results obtained showed a notable increase in validation accuracy and F1 scores when trained using our training method as compared to direct training.

We have obtained results for standard training using ResNet50 and ViTCNN model on the augmented dataset and compare it with our training procedure on the augmented dataset.

3.1 Achieved results on the validation dataset

We can see that the inclusion of our Methodology described in sections 2.1 and sections 2.2 increases the overall performance of the model. A comparative analysis has been shown in Table 1.

Table 1: Validation results and comparison to the baseline methods reported by the organizing team of Capsule Vision 2024 challenge.

Method	Avg. ACC	Avg. Precision	Avg. Recall	Avg. F1-score
ResNet50 (baseline)	76%	0.78	0.76	0.76
SVM (baseline)	82%	0.81	0.82	0.78
ResNet50 (Our Method)	89.57%	0.893	0.895	0.894
ViT-CNN Hybrid (Our Method)	**89.79%**	**0.909**	**0.897**	**0.902**

Note: All averages here are weighted averages.

4 Discussion

Our methodology tries to structure the learning process such that the model starts with an easy tasks and the complexity of the classification tasks increase gradually. This method proved effective as is seen in Table1 where we can see a notable increase in the F1 scores of the model. This signifies that the model has become better at handling minority classes (the harder tasks in our case).

5 Conclusion

In summary, the approach described here successfully tackles the challenges of classifying imbalanced image data for the Capsule Vision 2024 challenge. By starting with an easier task, separating Normal from Abnormal, and gradually adding more specific categories, the model learns to handle the complexity of the data step by step. This strategy, inspired by curriculum learning, helps the model adjust to the difficulty of each class, especially

those with fewer examples. Overall, this method shows promise for improving results across all classes, as seen in the results compared to baseline models.

6 Acknowledgments

As participants in the Capsule Vision 2024 Challenge, we fully comply with the competition's rules as outlined in [10]. Our AI model development is based exclusively on the datasets provided in the official release in [3].

References

[1] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. ISSN 2078-2489. doi: 10.3390/info11020125. URL https://www.mdpi.com/2078-2489/11/2/125.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[3] 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.

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

[5] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Mustafa Nasir-Moin, Naofumi Tomita, Lorenzo Torresani, Jason Wei, and Saeed Hassanpour. Learn like a pathologist: Curriculum learning by annotator agreement for histopathology image classification, 2020. URL https://arxiv.org/abs/2009.13698.

[6] 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, 2021. URL https://arxiv.org/abs/2010.11929.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[8] Dara Varam, Rohan Mitra, Meriam Mkadmi, Radi Aman Riyas, Diaa Addeen Abuhani, Salam Dhou, and Ayman Alzaatreh. Wireless capsule endoscopy image classification: An explainable ai approach. *IEEE Access*, 11:105262–105280, 2023. doi: 10.1109/ACCESS.2023.3319068.

[9] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL `http://arxiv.org/abs/1610.02391`.

[10] 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.