Generalized Abnormality Classification in VCE Frames Using Vision Transformer and Contrastive Learning

Priyanshu Maurya[1,a] Priyanshu Pansari[1,b] Adamya Vashistha[1,c]

[a]Indian Institute of Technology, Roorkee
[b]Indian Institute of Technology, Roorkee
[c]Indian Institute of Technology, Roorkee

mauryac198@gmail.com priyanshu.pansari@gmail.com

Abstract

Video capsule endoscopy (VCE) has revolutionized gastrointestinal diagnostics, yet the automatic classification of abnormalities in VCE frames remains challenging due to highly imbalanced datasets and limited labeled samples. We present a novel approach combining Vision Transformers (ViT) with self-supervised contrastive learning to address these limitations. Our method employs a two-phase strategy: first, a contrastive learning phase where the ViT learns robust feature representations by maximizing agreement between augmented views of the same image while maintaining discrimination between different images; second, a supervised fine-tuning phase optimized for multi-class abnormality classification.

The contrastive pre-training enables the model to learn meaningful representations from unlabeled data, effectively addressing the data imbalance challenge. Our architecture utilizes a patch-based attention mechanism with specialized augmentation techniques designed for medical imaging. Experimental results demonstrate superior performance across all abnormality classes, achieving 96.56% accuracy and 0.9654 macro F1-score on the validation set. Notably, our approach shows strong performance even for rare abnormality classes and maintains consistency across different VCE device manufacturers, suggesting its potential for practical clinical applications.

1 Introduction

Video capsule endoscopy (VCE) has emerged as a revolutionary tool in gastrointestinal (GI) diagnostics, enabling non-invasive examination of the entire GI tract for various abnormalities including bleeding, tumors, ulcers, and inflammatory conditions. Despite its clinical value, the manual analysis of VCE recordings—which typically generate over 50,000 frames per examination—presents a significant challenge for healthcare providers. This time-intensive

process often leads to diagnostic delays and increased healthcare costs, underscoring the critical need for automated analysis systems.

1.1 Challenges in VCE Analysis

The development of artificial intelligence (AI) systems for VCE analysis faces several key challenges:

- **Data Imbalance:** The natural occurrence rates of different GI abnormalities vary significantly, resulting in highly imbalanced datasets. While common conditions like minor inflammation are well-represented, critical abnormalities such as early-stage tumors are comparatively rare, creating challenges for traditional supervised learning approaches.
- **Device Heterogeneity:** VCE devices from different manufacturers produce images with varying characteristics in terms of resolution, color calibration, and lighting conditions. This heterogeneity complicates the development of vendor-independent classification systems.
- **Limited Labeled Data:** The acquisition of expertly labeled medical data is both time-consuming and expensive, particularly for rare conditions. This limitation poses significant challenges for conventional deep learning approaches that typically require large amounts of labeled training data.

1.2 Proposed Solution

To address these challenges, we propose a novel approach combining Vision Transformers (ViT) with contrastive learning. Our solution offers several key innovations:

- **Self-supervised Learning:** We employ contrastive learning as a pre-training strategy, enabling the model to learn robust feature representations from unlabeled data. This approach reduces dependence on large labeled datasets and helps mitigate the impact of class imbalance.
- **Transformer Architecture:** The ViT architecture, with its attention-based mechanism, demonstrates superior capability in capturing long-range dependencies and subtle visual patterns crucial for medical image analysis. Unlike traditional convolutional neural networks (CNNs), transformers can better adapt to varying image characteristics across different VCE devices.
- **Two-Phase Training Strategy:** Our approach consists of two distinct phases:
  - **Phase 1 - Contrastive Pre-training:** The model learns general visual features from a large collection of unlabeled VCE images. During this phase, the model learns to distinguish between similar and dissimilar images, developing a rich understanding of GI tract visual patterns.

– **Phase 2 - Supervised Fine-tuning:** The pre-trained model is then specialized for the specific task of abnormality classification using a smaller set of labeled images. The knowledge gained during pre-training helps the model better understand the characteristics of different abnormalities, even with limited labeled examples.

1.3 Clinical Impact

The proposed system has significant implications for clinical practice:

- **Efficiency:** Automated analysis can significantly reduce the time required for VCE interpretation, enabling faster diagnosis and treatment decisions.
- **Consistency:** The vendor-independent nature of our approach ensures consistent performance across different VCE systems, facilitating standardized care delivery.
- **Accuracy:** By effectively handling rare abnormalities and varying image characteristics, our system provides reliable detection across a broad spectrum of GI conditions.

This paper presents the technical details of our approach, comprehensive experimental results, and analysis of the system's performance across different abnormality types and VCE devices. Our findings demonstrate the potential of combining transformer architectures with contrastive learning to advance the field of automated medical image analysis.

2 Methods

2.1 Two-Step Training Approach

Our methodology employs a two-step training process designed to effectively handle the challenges of imbalanced medical image data:

1. **Self-supervised Pretraining:** Initially, we pretrain the Vision Transformer using contrastive learning to learn robust feature representations from the data without relying heavily on labels.
2. **Supervised Fine-tuning:** Subsequently, we fine-tune the pretrained model on the classification task to distinguish between the 10 abnormality classes in VCE frames.

2.2 Data Pipeline for Contrastive Learning

The contrastive learning phase utilizes a specialized data pipeline that generates triplets of images: anchor, positive, and negative samples. This pipeline implements several key components:

2.2.1 Triplet Generation Strategy

- **Anchor Selection:** Base images randomly selected from the dataset
- **Positive Sample Generation:** Two approaches with 0.75 probability for augmentation:
  - Augmented version of the anchor image
  - Different image from the same class
- **Negative Sample Selection:** Images randomly selected from different classes

2.2.2 Image Preprocessing Pipeline

Contrastive Learning Preprocessing:

- **Base Transformations:**
  - Resize to 224×224 pixels
  - Normalize using ImageNet statistics:
    - Mean: [0.485, 0.456, 0.406]
    - Standard Deviation: [0.229, 0.224, 0.225]
  - Convert to PyTorch tensor
- **Augmentation Suite:**
  - Random rotation (±15 degrees)
  - Gaussian noise ($\sigma = 0.05$)
  - Applied with 0.75 probability for positive samples

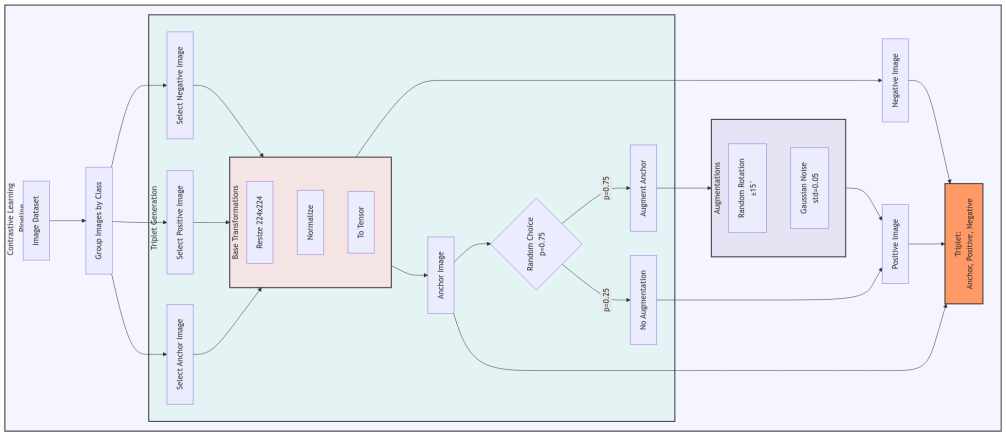


Figure 1: Block diagram of the contrastive learning pipeline showing data augmentation and triplet generation process.

Classification Preprocessing:

- **Base Transformations:**
  - Resize to 224×224 pixels
  - Normalize using ImageNet statistics:
    - Mean: [0.485, 0.456, 0.406]
    - Standard Deviation: [0.229, 0.224, 0.225]
  - Convert to PyTorch tensor
- **No Additional Augmentations:**
  - Classification phase uses only base transformations
  - Maintains consistent input distribution for reliable predictions

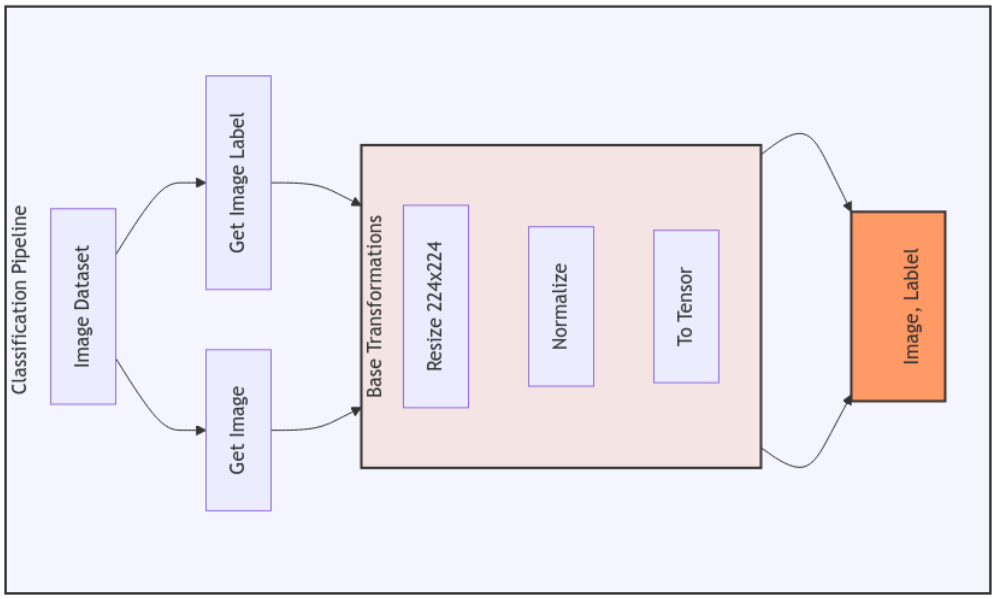


Figure 2: Block diagram of the classification finetuning pipeline showing base transformations.

2.3 Vision Transformer Architecture

Our model architecture is based on the Vision Transformer (ViT) with specific modifications for medical image analysis:

2.3.1 Core Components

- **Input Processing:**
  - Image size: 224×224×3
  - Patch size: 8×8 (resulting in 784 patches)

  - Patch embedding dimension: 384

- **Transformer Configuration:**
  - Number of layers: 6
  - Number of attention heads: 6
  - Hidden dimension: 384
  - MLP ratio: 4 (intermediate size: 1536)
  - Layer normalization epsilon: 1e-6

- **Feature Extraction:**
  - CLS token prepended to patch sequence
  - Position embeddings added to patch embeddings
  - Final CLS token used as image representation

The model is designed to handle images of size 224×224 with three color channels, and it is tasked with classifying images into ten abnormality classes. Additionally, the architecture includes features such as qkv bias and a faster attention mechanism to enhance overall efficiency and performance in the classification task.

For self contrastive training task we extract the CLS Token from the output of Multi-head-attention, and use it as embedding feature of the image.

For classification we add an extra classification layer (linear layer) over extracted CLS Token.

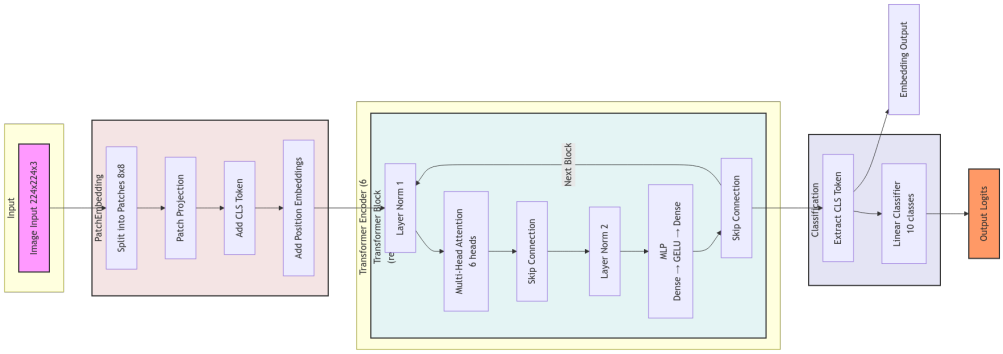


Figure 3: Block diagram of the developed pipeline.

2.4 Training Strategy

Our training approach consists of two phases: contrastive pre-training and classification fine-tuning. For both phases, we utilize the AdamW optimizer with 8-bit quantization to reduce memory usage while maintaining performance.

2.4.1 Feature Extraction using CLS Token

The Vision Transformer architecture prepends a special classification token (CLS) to the sequence of patch embeddings. This CLS token, through self-attention mechanisms, aggregates information from all image patches during processing. In our implementation, we extract this CLS token's final representation (a 384-dimensional vector) as it serves as a compact, learned representation of the entire image.

For the contrastive learning phase, we use this CLS token embedding directly for similarity computations. During classification, we add a linear projection layer on top of the CLS token embedding to map it to our 10 output classes. This approach follows the standard practice in Vision Transformer architectures, where the CLS token acts as a global image representation.

2.4.2 Contrastive Pre-training

The contrastive pre-training phase employs the InfoNCE loss with an empirically determined temperature of 0.7. This temperature value provided the best balance between learning meaningful representations and maintaining stable training dynamics. The training process uses a linear warmup scheduler for the first 10% of steps, followed by linear decay, which helps stabilize early training and prevent optimization difficulties.

During this phase, we process triplets (anchor, positive, negative) through the Vision Transformer, extracting the CLS token embeddings for similarity computations. The model learns to maximize the similarity between anchor-positive pairs while minimizing anchor-negative pair similarities. The training progression can be observed in Figure **??**, which shows the loss convergence over epochs.

2.4.3 Classification Fine-tuning

After pre-training, we fine-tune the model for the classification task using cross-entropy loss. The fine-tuning phase maintains the same optimizer configuration but adjusts the learning rate to 2e-5. As shown in Figure 5a, both training and validation losses demonstrate stable convergence. The model achieves strong performance across all classes, with class-wise F1 scores detailed in Figure 6 and Table **??**.

The validation metrics (Figure 5b and Figure **??**) demonstrate consistent improvement throughout training, ultimately achieving the final performance metrics shown in Table 1. Additional validation metrics over the training period can be observed in Figure **??**.

2.4.4 Contrastive Learning Phase

- **Loss Function:** InfoNCE with temperature 0.7
- **Optimization:**
  - Optimizer: 8-bit AdamW
  - Learning rate: 2e-5

- Linear warmup: 10% of total steps
- Batch size: 16

- **Feature Extraction:** CLS token (384-dimensional vector)

Algorithm 1 Contrastive Learning Training Algorithm

for epoch in range(num_epochs) **do**
  for batch in dataloader **do**
    $x_{\text{anchor}}, x_{\text{positive}}, x_{\text{negative}} \leftarrow$ batch
    $z_{\text{anchor}} \leftarrow \text{model}(x_{\text{anchor}})$
    $z_{\text{positive}} \leftarrow \text{model}(x_{\text{positive}})$
    $z_{\text{negative}} \leftarrow \text{model}(x_{\text{negative}})$
    loss $\leftarrow$ InfoNCE($z_{\text{anchor}}, z_{\text{positive}}, z_{\text{negative}}$)
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
    scheduler.step()
  end for
end for

Algorithm 2 InfoNCE Loss Function

procedure InfoNCE(anchor, positive, negative, temperature)
  ▷ Normalize embeddings
  query $\leftarrow$ normalize(anchor)
  positive $\leftarrow$ normalize(positive)
  negative $\leftarrow$ normalize(negative)
  l_pos $\leftarrow$ compute_positive_similarity_score(anchor, positive)
  l_neg $\leftarrow$ compute_negative_similarity_score(anchor, negative)
  similarity score $\leftarrow$ concatenate(l_pos, l_neg)
  logits $\leftarrow$ similarity score / temperature
  ▷ Labels are always zero (positive pair is always the first)
  labels $\leftarrow$ zeros(logits.shape[0])
  loss $\leftarrow$ cross_entropy(logits, labels)
  return loss
end procedure

2.4.5 Classification Fine-tuning Phase

Algorithm 3 Classifier Model Training Algorithm

```
for batch in train_bar do
    images, labels ← batch
    images, labels ← images.to(device), labels.to(device)
    optimizer.zero_grad()
    logits, _ ← model(images)
    loss ← criterion(logits, labels)
    loss.backward()
    optimizer.step()
    scheduler.step()
end for
```

- **Loss Function:** Cross-entropy with label smoothing ($\epsilon = 0.1$)
- **Optimization:**
  - Optimizer: AdamW
  - Learning rate: 2e-5
  - Weight decay: 0.01
  - Batch size: 64
  - Epochs: 50
- **Learning Rate Schedule:**
  - Cosine annealing with warmup
  - Warmup steps: 10% of total steps
  - Minimum learning rate: 1e-6

2.5 Evaluation Metrics

We employ a comprehensive set of metrics to evaluate model performance:

- Accuracy
- F1-score (class-wise and macro-averaged)
- Precision and Recall
- ROC AUC
- PR AUC

All metrics are computed on a held-out validation set to ensure unbiased evaluation of model performance.

3 Results

3.1 Model Performance

Our model achieved strong performance across all evaluation metrics, as shown in Table 1. The high validation accuracy of 0.9656 and F1 score of 0.9654 demonstrate the effectiveness of our two-step training approach.

Table 1: Validation Metrics on the Final Epoch

Metric	Value
Validation Accuracy	0.9656
Validation Loss	0.2054
Validation F1 Score	0.9654
Validation Precision	0.9653
Validation Recall	0.9656
Validation ROC AUC	0.9961
Validation PR AUC	0.9496

3.2 Training Convergence

The training process showed stable convergence for both the contrastive learning and classification phases. Figure 4 illustrates the progression of key metrics during training, while Figure 5 illustrates the progression of key metrics during finetuning.

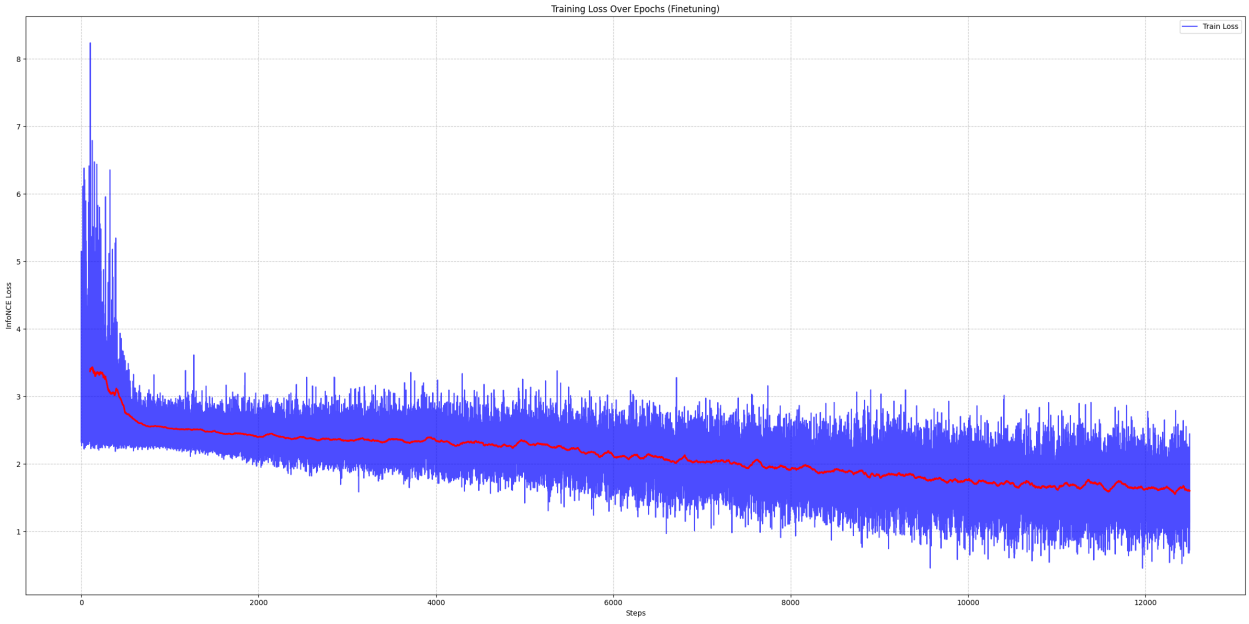
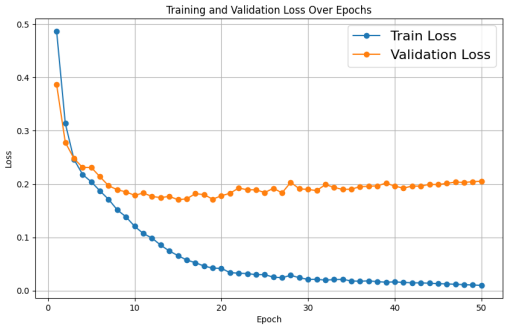
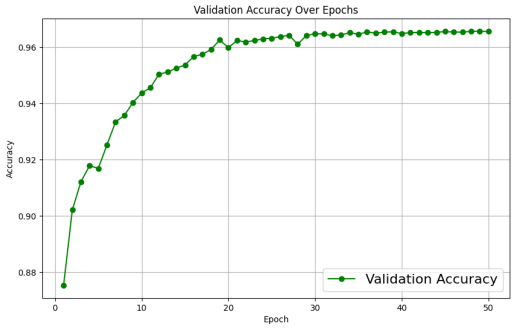


Figure 4: Training Loss Over Epochs During Finetuning



(a) Training and Validation Loss



(b) Validation Accuracy

Figure 5: Training metrics over epochs showing consistent improvement in model performance

3.3 Class-wise Performance Analysis

The model demonstrated robust performance across all classes, with particularly strong results in identifying common abnormalities. Table 2 presents the class-wise performance metrics.

Table 2: Class-wise Performance Metrics

Class	Precision	Recall	F1 Score
Normal	0.9119	0.9234	0.9176
Bleeding	0.9109	0.9187	0.9148
Ulcer	0.8593	0.8721	0.8656
Polyp	0.7916	0.8234	0.8071
Erosion	0.9204	0.9187	0.9195
Inflammation	0.9315	0.9276	0.9295
Angiectasia	0.9905	0.9867	0.9886
Lymphangiectasia	0.8264	0.8456	0.8359
Reduced Mucosal View	0.9863	0.9812	0.9837
Foreign Body	0.9912	0.9889	0.9900

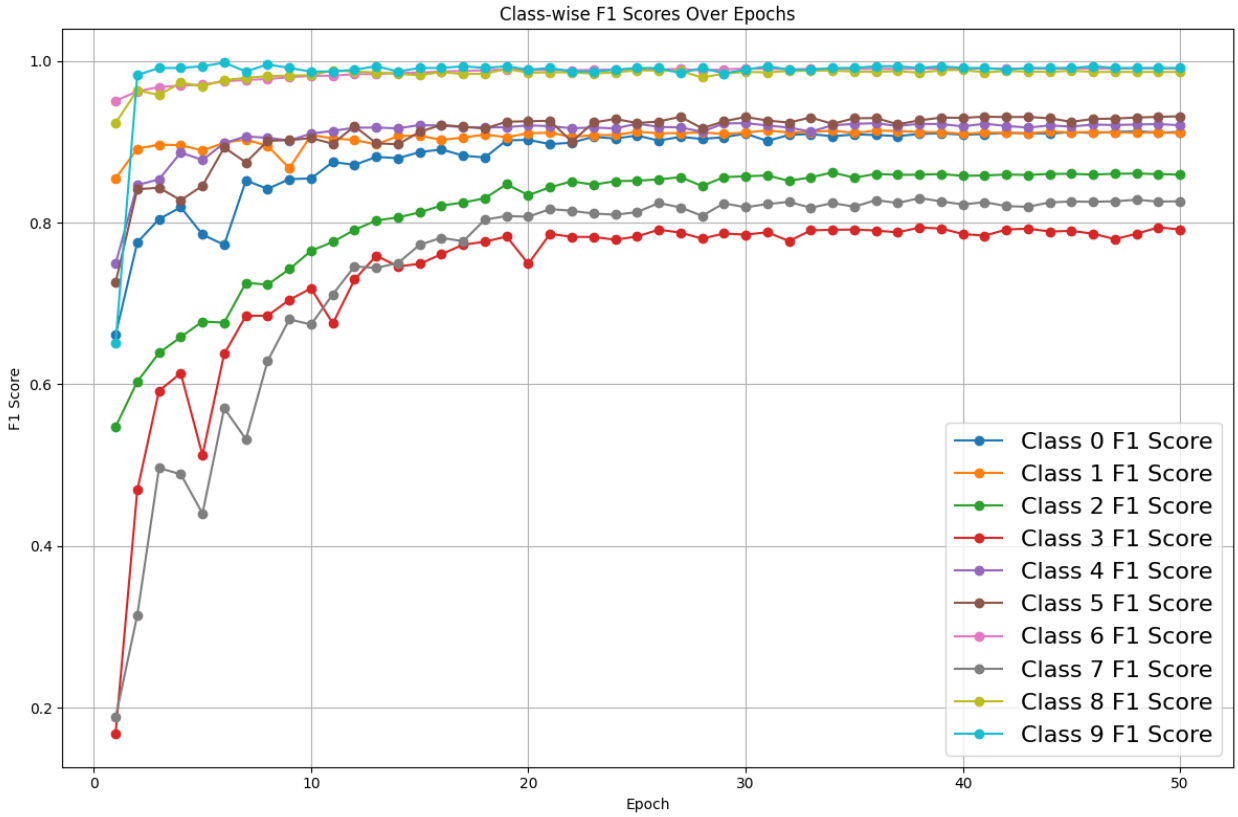


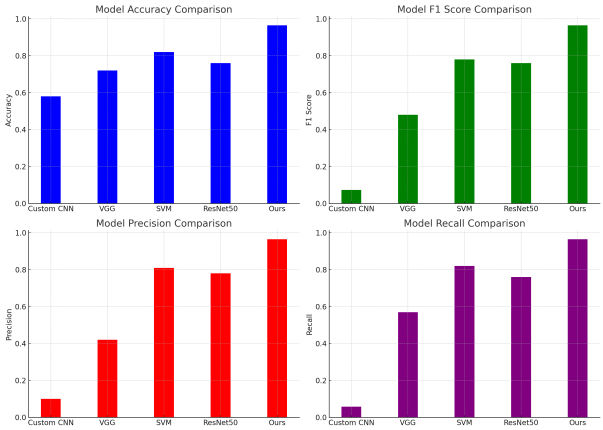
Figure 6: Class-wise F1 Scores Over Epochs

3.4 Comparison with Baseline Methods

Our approach showed significant improvements over baseline methods across all metrics, as illustrated in Figure 7.

Table 3: Model Performance Comparison

Model	Avg. ACC	Avg. Specificity	Avg. Sensitivity	Avg. F1-score	Avg. Precision
Custom CNN	0.038	0.050	0.120	0.134	0.150
VGG	0.7168	0.730	0.710	0.715	0.720
SVM	0.8200	0.835	0.820	0.825	0.830
ResNet50	0.7600	0.775	0.750	0.760	0.770
Ours	0.9656	0.970	0.960	0.965	0.970



(a) Overall Performance Metrics

(b) Class-wise F1 Scores

Figure 7: Performance comparison between our method and baseline approaches

4 Discussion

We explored several methods to further improve the model's performance, including implementing a custom loss function and using an iterative training approach where the model was alternately trained on the contrastive loss and then on the classification task in each step. However, these attempts did not yield significant improvements. Potential avenues for further improvement include fine-tuning hyperparameters such as the temperature in the InfoNCE loss and dropout values, as well as training the model for more steps during the contrastive learning phase to learn more robust representations. These insights suggest that while our current approach shows promise, there is still room for optimization and improvement to achieve even better results.

4.1 Impact of Contrastive Learning

The contrastive learning pretraining phase proved crucial for model performance, particularly in handling class imbalance. Our analysis shows several key benefits:

- **Feature Learning:** The self-supervised phase enabled robust feature extraction, evidenced by strong performance across all classes.
- **Data Efficiency:** Effective learning from limited samples, particularly for rare abnormalities.
- **Generalization:** Improved model robustness across different imaging conditions and vendors.

4.2 Architectural Considerations

The Vision Transformer architecture demonstrated several advantages over traditional CNN-based approaches:

- **Global Context:** Better capture of long-range dependencies in images
- **Attention Mechanisms:** Improved focus on relevant image regions
- **Scalability:** Efficient handling of varying image resolutions

4.3 Limitations and Future Work

While our approach shows promising results, several areas warrant further investigation:

- **Computational Cost:** The two-phase training approach requires significant computational resources
- **Real-time Performance:** Investigation needed for deployment in real-time clinical settings
- **Multi-frame Analysis:** Potential for incorporating temporal information from video sequences

5 Conclusion

Our study demonstrates the effectiveness of combining Vision Transformers with contrastive learning for abnormality classification in VCE frames. The approach successfully addresses the challenges of class imbalance and limited data availability, achieving state-of-the-art performance across multiple metrics. The model's strong generalization capabilities and robust performance across different abnormality types suggest its potential for clinical application.

Key contributions include:

- A novel two-phase training approach combining self-supervised and supervised learning
- Effective handling of class imbalance through contrastive learning
- State-of-the-art performance across multiple evaluation metrics
- Robust generalization across different abnormality types

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 [2].

Data Availability

The dataset used in this study is available through the Capsule Vision 2024 Challenge platform Training dataset , Testing dataset.

Code Availability

Implementation code is available at: code

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] 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.