IT5811 PROJECT II

MULTI-CLASS ABNORMALITIES CLASSIFICATION AND SEGMENTATION FOR GASTROINTESTINAL DIAGNOSIS IN CAPSULE ENDOSCOPY

Presented By

Nithinsrivatsan S (2021506058)

Vishnuvasan M (2021506125)

Shiek sajnathul faizana (2021506096)

Madeshwaran (2021506313)

Under the Guidance of

Dr. M. Hemalatha – Asst. Professor

Department of Information Technology, MIT

Problem Statement:

Capsule endoscopy creates large amounts of video data. Analyzing these frames manually is difficult and time-consuming. The task is to create an efficient and reliable system capable of automatically classifying and segmenting abnormalities captured in video capsule endoscopy. The solution must ensure accurate classification of abnormalities, as well as precise segmentation, to highlight areas of concern for further analysis by medical professionals.

Working of Capsule Endoscopy:

- A small vitamin-sized capsule contains a camera used for endoscopy.
- After swallowing, the capsule travels through the digestive tract and captures thousands of images.
- The images are recorded by a device worn around the waist.
- It provides visuals of the small intestine, which is hard to access with traditional endoscopy.
- The capsule exits the body naturally after around 8 hours through a bowel movement.

Applications:

- Early detection of abnormalities
- Efficiency in medical diagnosis
- Remote healthcare support
- Improved patient monitoring
- Integration into telemedicine platforms.

Need for The System:

- Doctors need to manually review 50,000 to 80,000 frames from just one capsule endoscopy. This process is extremely time-consuming and slows down the overall diagnosis and treatment workflow.
- Continuous visual analysis can lead to fatigue and missed abnormalities, affecting diagnostic accuracy.
- The manual process places a heavy burden on medical professionals, affecting their efficiency and focus.

Objective:

The project aims :

- To develop a system that automatically identifies and classifies abnormalities in video capsule endoscopy frames, as well as segments the affected areas for detailed analysis.
- To ensure the system accurately classifies abnormalities and precisely segments them, aiding in improved diagnosis and decision-making.
- To create a user-friendly and intuitive system that is easy for medical professionals to operate, facilitating quick and efficient adoption.

Scope of the project:

- The system processes individual frames captured by a swallowed capsule camera to detect and analyze gastrointestinal conditions..
- It classifies frames into ten categories like bleeding, ulcer, polyp, and more, using advanced deep learning models and the system segments the abnormal regions within the frames, helping doctors clearly see the affected areas..
- By automatically identifying and highlighting abnormalities, the system supports early diagnosis and timely medical action.

Dataset:

- The Kvasir dataset contains endoscopy images verified by medical experts, collected from hospitals in Norway.
- For classification, it includes 52,315 frames across 10 abnormality classes with proper train, test, and validation splits.
- For segmentation, it provides 1,000 labeled images with ground truth masks for detecting abnormal regions.

Literature Survey / Existing Systems:

Full Literature Survey Document

Shift from Machine Learning to Deep Learning:

- Early methods: Handcrafted features (SIFT, MPEG-7, LBP) + SVM
- Limited generalization and manual feature extraction
- Deep Learning (CNNs) replaced manual feature extraction, improving classification and generalization.

Advancements in CNN Architectures:

- Popular models: ResNet50, DenseNet121, VGG16
- Feature fusion techniques improved performance with larger datasets
- Better generalization and feature learning for complex tasks.

Efficient Feature Extraction with Modern Models

- EfficientNet: Optimized feature extraction with minimal computational cost
- Transformer-based models (ViT) improved long-range feature dependencies and accuracy.

Literature Survey / Existing Systems:

Full Literature Survey Document

Saliency and Attention-Based Methods:

- LSST and SALLC: Focused on detecting important regions in images.
- Attention-based models (SE-ResNet, SCE-ResNet) dynamically highlight key features for better interpretability

Hybrid & Transfer Learning Approaches:

- ResNet50 + Transfer Learning: Improved feature learning for medical images
- Capsule Networks + CNNs: Retained spatial hierarchies and improved classification robustness

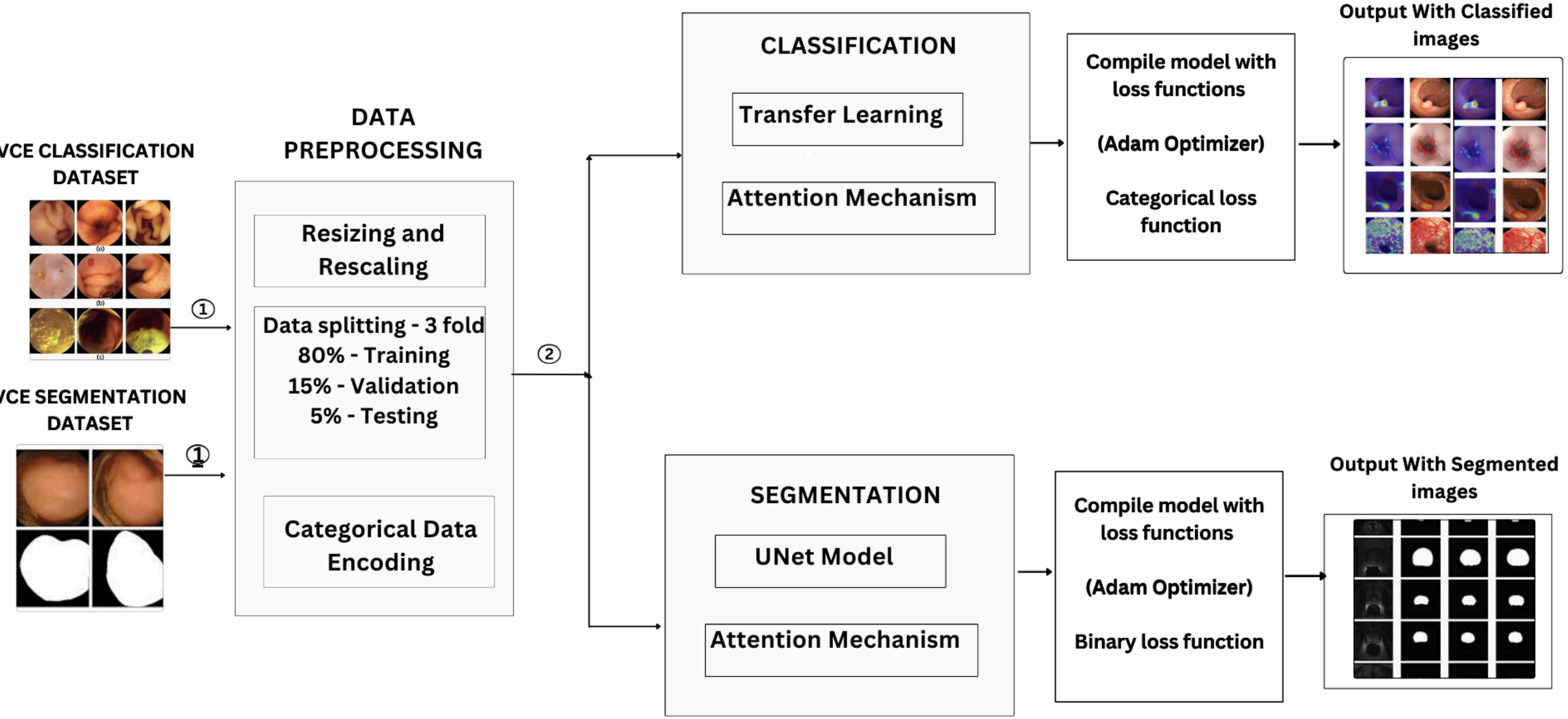
Limitations:

- Models struggle to detect lesions with unclear textures, leading to low sensitivity and missed abnormalities.
- Traditional approaches treat features independently and fail to capture local relationships in image data.
- Lesions with subtle boundaries are often poorly segmented, reducing diagnostic accuracy.
- Small and noisy datasets increase the chances of overfitting and reduce overall performance.

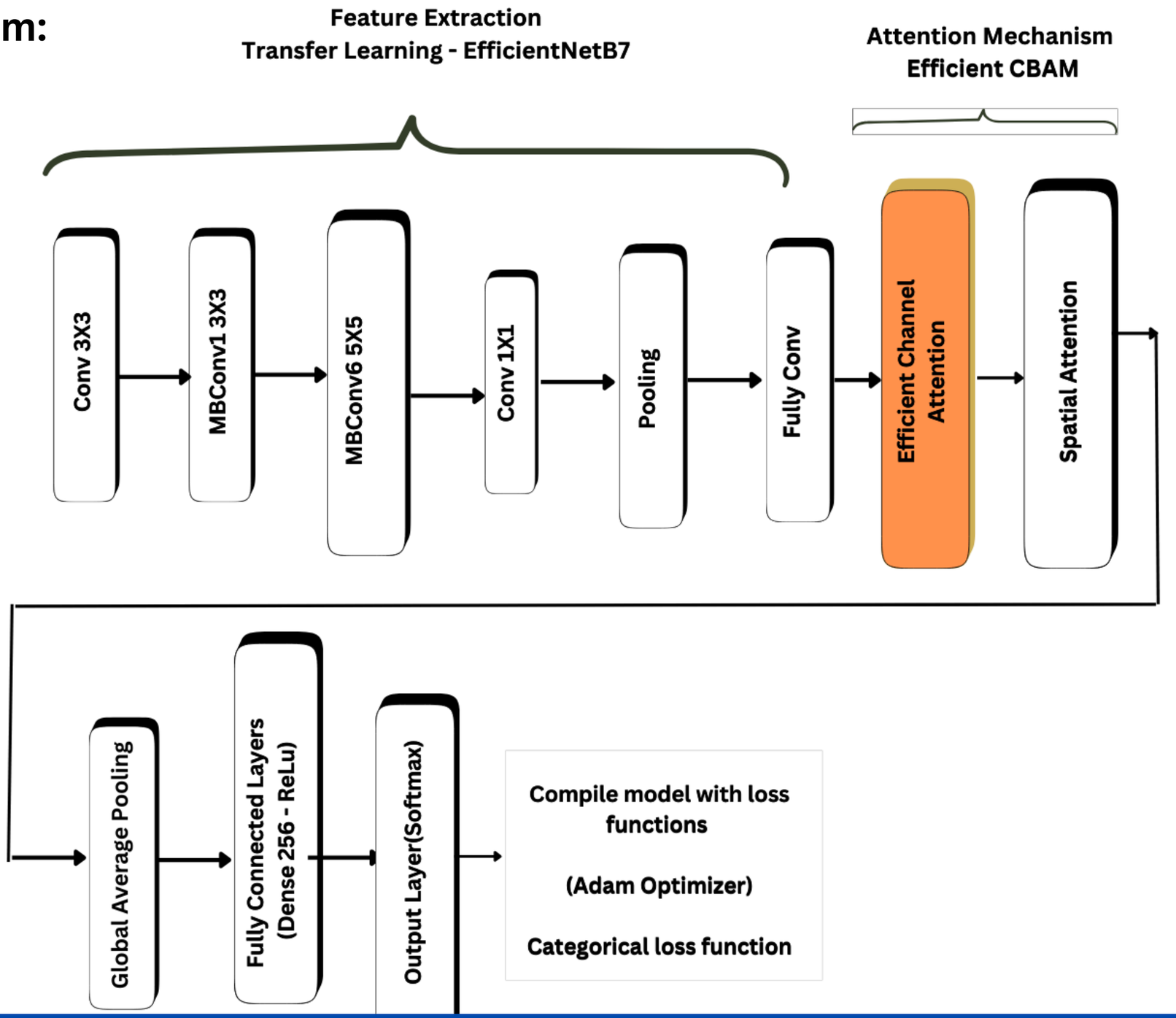
Proposed System Overview:

- The system uses two datasets: one for classification and one for segmentation, both containing VCE images of the intestine.
- Images are resized, rescaled, and split into training (80%), validation (15%), and testing (5%) sets; labels are encoded for model compatibility.
- Transfer learning is applied using pretrained CNN models like EfficientNetB7 to extract meaningful features efficiently only for classification.
- An attention module combining channel and spatial attention helps the system focus on the most relevant image regions.
- For classification, features pass through global average pooling, dense layers, and a Softmax layer to predict the image class.
- For segmentation, refined features are processed using a UNet architecture to produce precise segmentation maps of abnormal areas.
- Both models are compiled using the Adam Optimizer—categorical loss for classification and binary loss for segmentation—and output class predictions with highlighted abnormal regions.

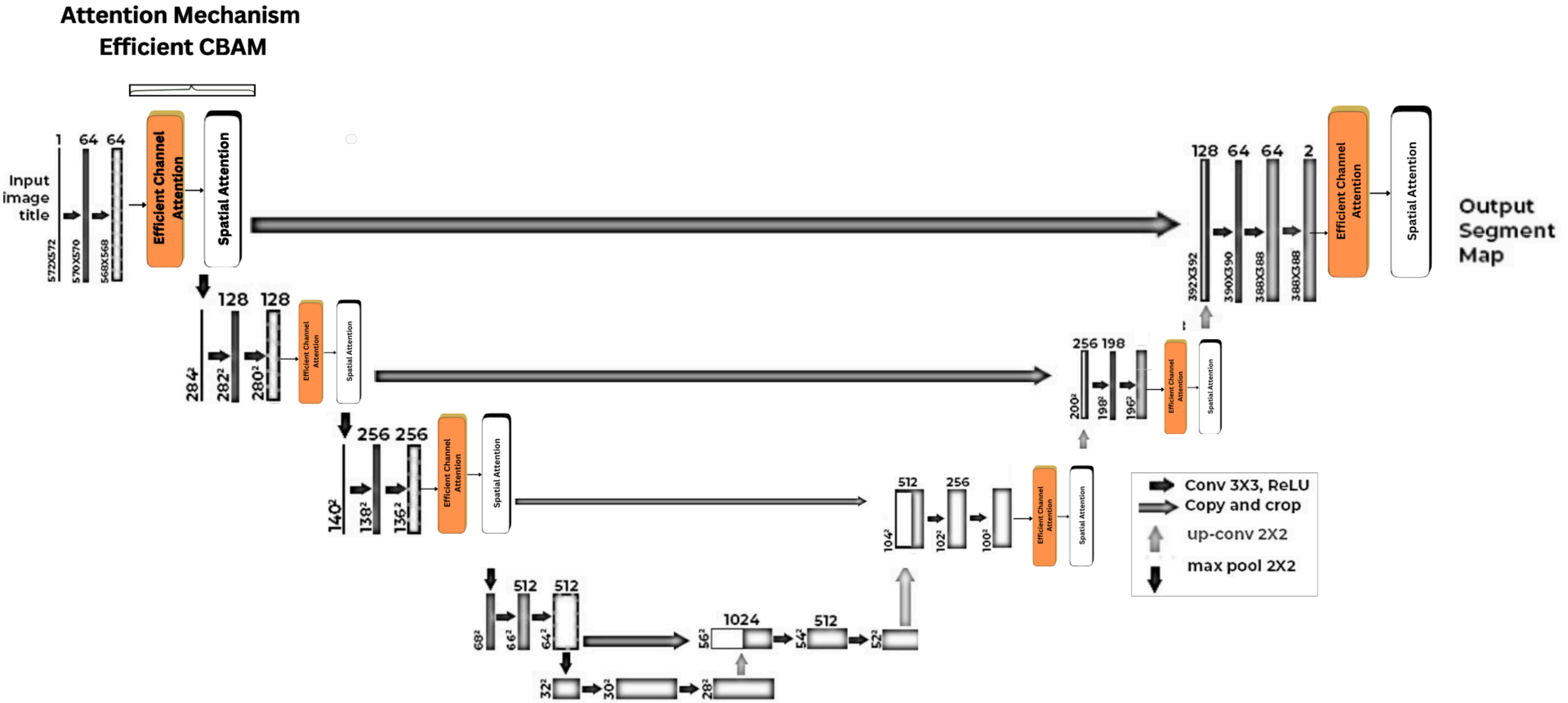
Architecture Diagram:



Classification Diagram:



Architecture Diagram:



Module Design:

1. Input Module:

- Classification: VCE images labeled with abnormality class.
- Segmentation: VCE images with binary masks for abnormal regions.

2. Preprocessing Module:

- Resize images/masks (e.g., 224×224 or 256×256).
- Normalize pixel values to [0,1].
- Split: 80% train, 15% validation, 5% test.
- Encode labels: One-hot (classification), binary (segmentation).

3. Feature Extraction with Transfer Learning:

- Classification Only:
  - Pretrained EfficientNetB7 extracts rich visual features (edges, textures).
  - Reduces training time and improves accuracy by leveraging learned weights from ImageNet.

4. Attention Mechanism: (Applied to both tasks):

- Efficient Channel Attention (ECA): Focuses on relevant feature channels using lightweight 1D convolution.
- Spatial Attention: Focuses on key spatial regions in the image for improved attention localization.

5. Task-Specific Architecture:

- Classification Path:
  - Global Average Pooling → Dense Layer (ReLU) → Softmax Layer → Class Probabilities.
- Segmentation Path:
  - U-Net Architecture:
    - Encoder: Convolution + pooling layers capture semantic features.
    - Decoder: Up-sampling with skip connections for precise localization.
    - Final Layer: 1×1 convolution with Sigmoid activation for binary segmentation map.

6. Training Setup:

- Optimizer: Adam.
- Loss: Categorical Cross-Entropy (classification), Binary Cross-Entropy (segmentation).
- Metrics: Accuracy, Precision, Recall, F1 (classification); IoU, Dice, Pixel Accuracy (segmentation).

7. Output Module:

- Classification: Predicts image class.
- Segmentation: Highlights abnormal regions in binary output mask.

Implementation:

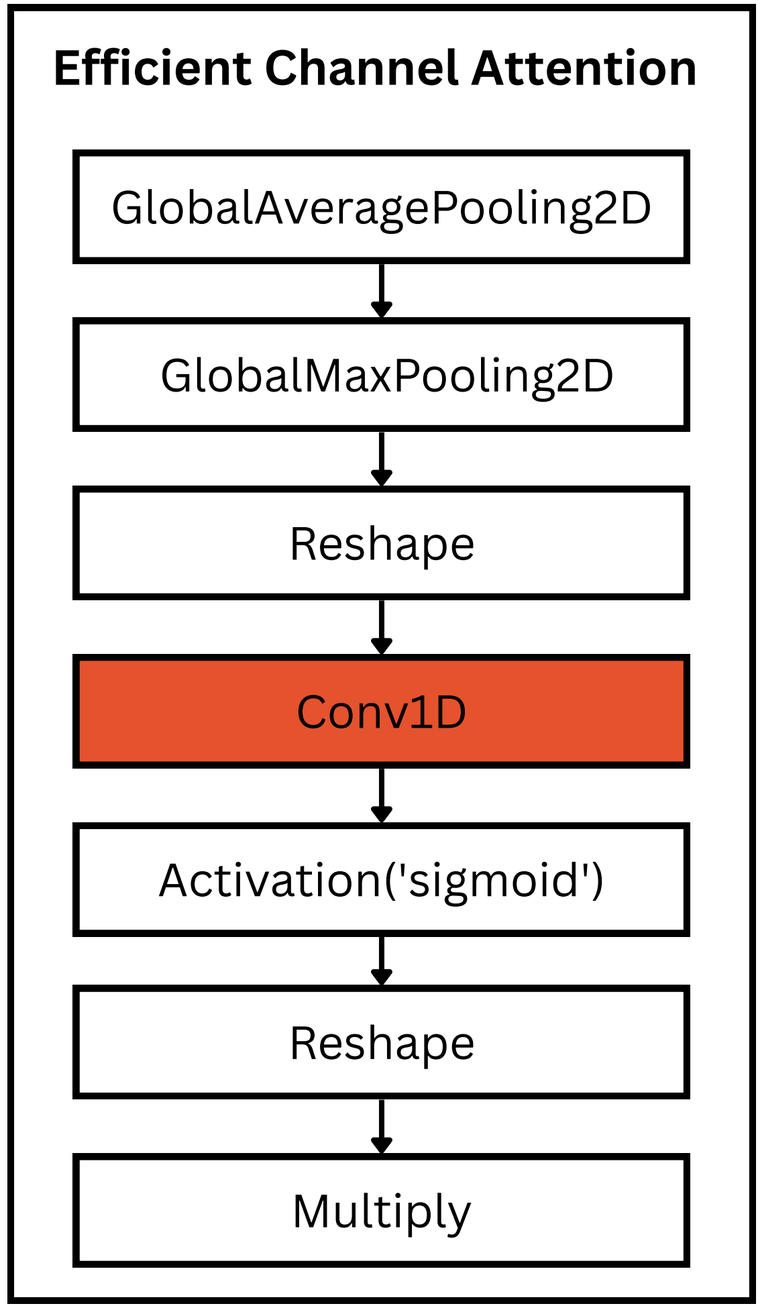
Classification

- A custom attention block combines Efficient Channel Attention (ECA) using Conv1D and CBAM-style spatial attention using Conv2D after average and max pooling.
- EfficientNetB7 (pretrained on ImageNet) is used as the base model with all layers set to be trainable.
- The ECA + CBAM attention block is applied to the output of EfficientNetB7 to enhance feature representation.
- The output is processed through GlobalAveragePooling, a Dense layer with 256 units, BatchNormalization, and a final softmax layer for classification.

Segmentation

- A CBAM block is used after each encoder and decoder stage, applying both channel attention (via MLP on global pooled features) and spatial attention (via Conv2D on pooled maps).
- The architecture follows the standard U-Net structure with symmetric encoder-decoder paths and skip connections.
- Each convolution block consists of two Conv2D layers followed by BatchNormalization and attention refinement through CBAM.
- The model outputs a single-channel mask using a sigmoid-activated Conv2D layer and is compiled with binary cross-entropy and BinaryIoU metrics.

Implementation:



Algorithm 3 Attention Module for Feature Refinement

Procedure: `Apply_Attention(X)`

Input: Feature map $X \in \mathbb{R}^{B\times H\times W\times C}$

Output: Refined feature map $X'' \in \mathbb{R}^{B\times H\times W\times C}$

Step 1: Channel Attention via ECA

Apply Global Average Pooling (GAP): $Z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j,c} \quad \forall c \in \{1,\ldots,C\}$

Output shape after GAP: $Z \in \mathbb{R}^{B\times C}$

Reshape for 1D convolution: $Z \to \tilde{Z} \in \mathbb{R}^{B\times C\times 1}$

Apply 1D convolution and sigmoid activation: $s = \sigma(\text{Conv1D}(\tilde{Z})) \in \mathbb{R}^{B\times C\times 1}$

Reshape s to match input: $s \to \hat{s} \in \mathbb{R}^{B\times 1\times 1\times C}$

Apply channel-wise multiplication: $X' = X \cdot \hat{s}$

Step 2: Spatial Attention

Apply average pooling along channels: $M_{\text{avg}}(i,j) = \frac{1}{C}\sum_{c=1}^{C} X'_{i,j,c}$

Apply max pooling along channels: $M_{\text{max}}(i,j) = \max_{c=1}^{C} X'_{i,j,c}$

Output maps: $M_{\text{avg}}, M_{\text{max}} \in \mathbb{R}^{B\times H\times W\times 1}$

Concatenate pooled maps: $M = \text{Concat}(M_{\text{avg}}, M_{\text{max}}) \in \mathbb{R}^{B\times H\times W\times 2}$

Apply 2D convolution and sigmoid activation: $S = \sigma(\text{Conv2D}(M)) \in \mathbb{R}^{B\times H\times W\times 1}$

Final refined output: $X'' = X' \cdot S$

Ablation study - Classification:

Model	Accuracy	Precision	Recall	F1-Score
1. CNN Model	81.08	78.24	81.08	75.9
2. Squeeze and excitation with MobileNetV2	90.50	91.36	90.5	90.53
3. Self Attention with ResNet	92.40.	92.26	92.4	92.20
4. EfficientNet B7 without CBAM.	93.77	93.8	93.77	93.75
5. CBAM with EfficientNet B7.	93.88	93.7	93.88	93.80
6. Efficient CBAM with EfficientNet B7.	**94.25**	**94.30**	**94.25**	**94.2**

the combination of EfficientNet B7's powerful feature extraction and Efficient CBAM's attention which integrates Efficient Channel Attention and spatial Attention mechanism that enhances important spatial and channel features, leading to better performance with lower computational cost.

Ablation study - Segmentation:

Model	Accuracy	IoU	mDice
1. UNet Model	91.3	70.7	79.7
2. CBAM with UNet	92.15	70.8	81.09
3. Self Attention with ResNet	93.1	76.1	83.9
4. Data Augumented Efficient CBAM with UNet	**93.55**	**78.4**	**86.2**

The Data Augmented Efficient CBAM with UNet model works best for segmentation because it learns from a variety of rotated, stretched, and brightened images, also with Efficient CBAM's attention which integrates Efficient Channel Attention and spatial Attention mechanism that enhances important spatial and channel features helping it detect shapes and boundaries more accurately in different conditions.

Result Analysis:

Classification Inference:

- Attention mechanisms improved accuracy from 81.08% to 92.40%.
- EfficientNetB7 boosted performance further to 93.77%.
- Combining CBAM with EfficientNetB7 achieved the best accuracy of 94.25%.

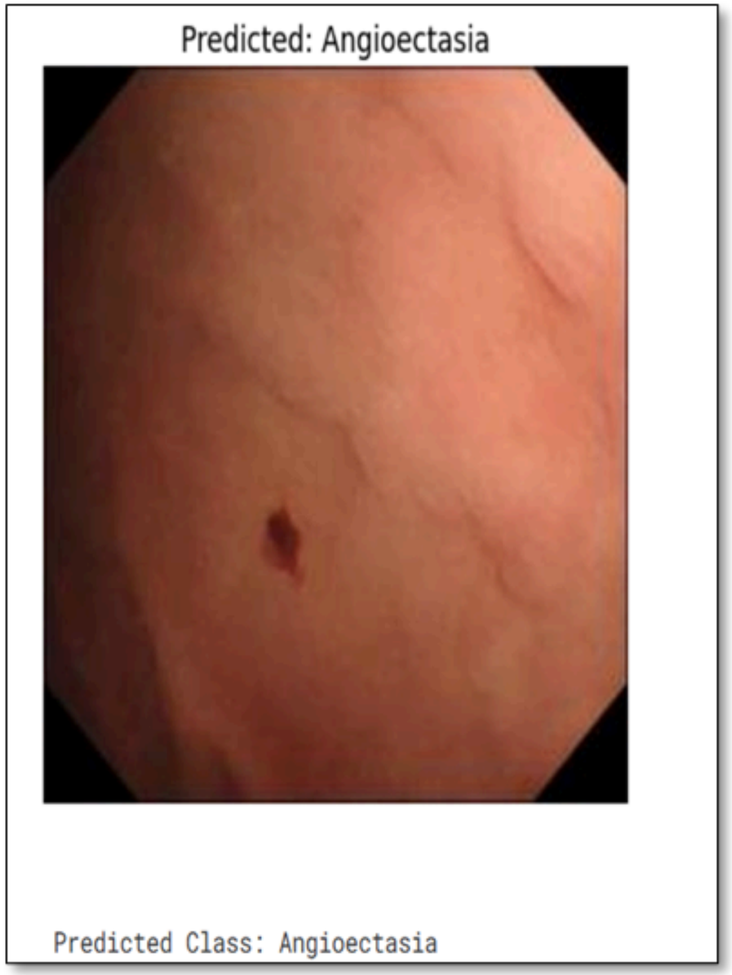
Segmentation Inference:

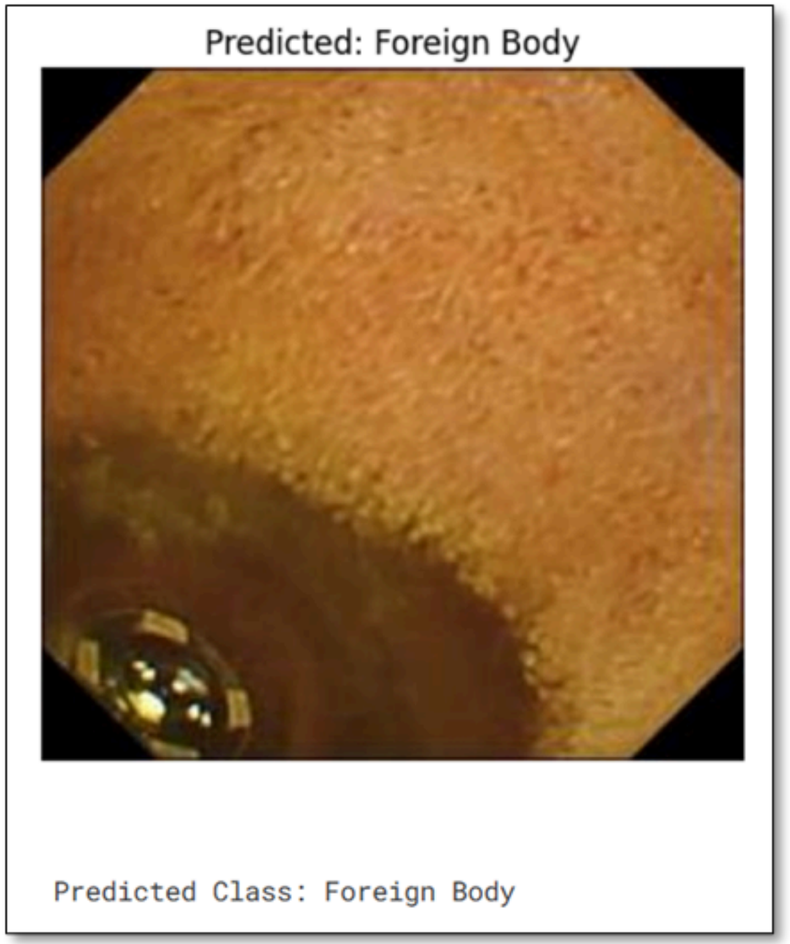
- UNet alone gives decent results with 70.7 IoU and 79.7 mDice.
- Adding attention improves boundary focus, raising mDice to 83.9.
- Best performance (86.2 mDice) achieved with CBAM + UNet + Data Augmentation.

Future Works:

- Integrate classification and segmentation into a unified model for a complete end-to-end diagnostic system.
- Improve accuracy further using advanced architectures and more diverse datasets for better and more reliable predictions.

Qualitative analysis - Classification

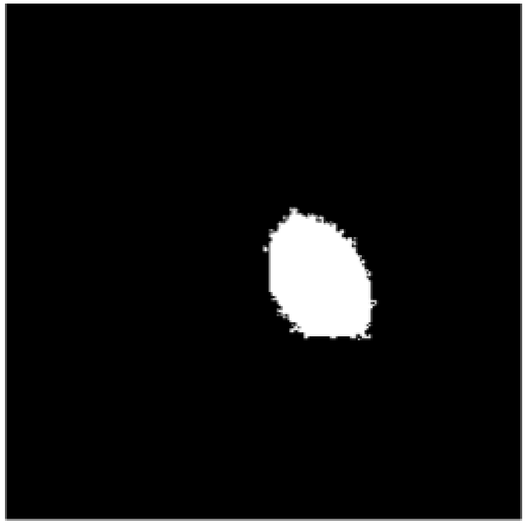




Qualitative analysis - Segmentation

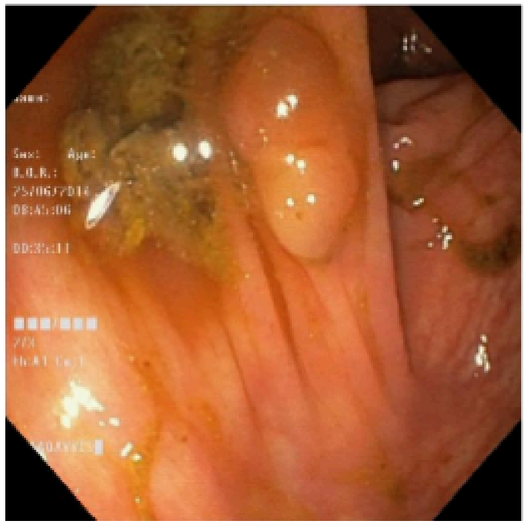
Original Image

Ground Truth Mask



Predicted Mask

Original Image



Ground Truth Mask

Predicted Mask

THANK YOU!