Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

An Interpretable Deep Learning Framework for Distinguishing Esophagitis from Barrett's Esophagus Using Endoscopic Images

Atheequllah Khan M.[1], Krishna J.[2]

[1]Research Scholar, Department of Computer Science and Engineering, Annamacharya University, Rajampet, India.
[2]Associate Professor, Department of Artificial Intelligence and Machine Learning, Annamacharya University, Rajampet, India.

E-mail: [1]cse.atheequlla@ssits.ac.in, [2]jkn@aitsrajampet.ac.in

Orcid ID: [1]0000-0002-7822-005X, [2]0000-0001-9283-4595

Abstract

The accurate distinction between esophagitis and Barrett's Esophagus is important for prompt diagnosis and clinical management, since Barrett's Esophagus is known to be a precursor of esophageal adenocarcinoma. However, the endoscopic appearance is difficult to interpret, as the overall attachment and appearance of the mucosa are similar and can be interpreted differently among clinicians. The aim of this work is to present an interpretable deep learning framework to automatically classify endoscopic images of esophagitis or Barrett's Esophagus into either binary class. The framework was validated on a balanced set, which has been built from the publicly available Kvasir and HyperKvasir databases. To gain insight into the model transparency, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to display the regions of the image that were more relevant for the classification decision. Three-fold stratified cross-validation (CV) was used to evaluate the experimental results and the fine-tuned CNNs consistently outperformed the feature extraction based methods. The most recent model that was evaluated was fine-tuned InceptionV3 with an accuracy of 99.80%, a precision of 99.60%, a recall of 100.00%, an F1-score of 99.80%, and an area under the receiver operating characteristic curve (AUC) of 1.000. The model was found to be more interpretable, as shown by Grad-CAM, which indicated that the model learned about clinically relevant mucosal abnormalities. The results of this study support the potential of the proposed framework as an accurate and interpretable computer-aided diagnostic tool for assessment of esophageal diseases in endoscopy.

Keywords: Esophagitis; Barrett's Esophagus; Endoscopic Image Analysis; Deep Learning; Convolutional Neural Network; Transfer Learning; End-to-End Fine-Tuning; Explainable AI; Grad-CAM; Computer-Aided Diagnosis.

1. Introduction

Esophageal disease, especially when associated with esophagitis and Barrett's Esophagus, constitutes significant and pathologically discernible disease that has prognostic

DOI: https://doi.org/10.36548/jiip.2026.3.020
Received: 13.05.2026, received in revised form: 19.06.2026, accepted: 10.07.2026, published: 21.07.2026

significance. This is now a growing concern worldwide due to the burden of disease and healthcare expenditure. Barrett's Esophagus is an accepted precancerous stage that increases the chance of developing adenocarcinoma of the Esophagus, which is regarded as one of the most deadly cancers recorded in the gastrointestinal tract. Traditional knowledge holds that esophagitis is an inflammatory disease process that may often be resolved with early diagnosis and adequate management. In recent years, deep learning, particularly Convolutional Neural Networks (CNNs), which have been shown to learn discriminative features from endoscopy images, have become promising candidates to analyze such images and, in some cases, have outperformed handcrafted feature-based techniques.

The diagnosis of esophagitis and Barrett's Esophagus can be very useful in making the correct clinical decision and taking the best measures to prevent cancer. A lack of or delay in diagnosis may allow Barrett's Esophagus to progress to esophageal adenocarcinoma, so early detection is important [1]. Normal endoscopic diagnosis is still difficult, however, and has been shown to have a high level of inter-observer variability, particularly in early or mild cases [2]. In gastrointestinal endoscopy, computer-aided diagnosis (CAD) systems using artificial intelligence (AI) have been a recent research focus with the promise of automated and objective image analysis. CNNs have been successfully used to discriminate features learned from endoscope images on par with expert clinicians in the task of identifying a variety of esophageal abnormalities [3], while many important issues still remain. These consist of small-scale generalization across datasets collected using different endoscopic equipment, class imbalance that does not reflect the distribution of real-world data, lack of high-quality labeled data, and the difficulty of understanding the predictions of models, commonly known as the "black-box" problem.

In view of these drawbacks, the fields of transfer learning and hybrid algorithms that combine deep feature extraction with classical machine learning classifiers have received much attention in recent years. Transfer learning has been shown to be useful in discriminative feature extraction and in reducing the need for large domain-specific training data, through being able to transfer knowledge from large general and medical image datasets [4]. At the same time, end-to-end fine-tuning of CNNs is shown to effectively improve performance by tuning the hierarchical features in the image to improve the features of esophageal endoscopy [5]. While architectures like ResNet, DenseNet, and Inception have been shown to be more competitive than traditional approaches in various tasks using suitable data augmentation and cross-validation [6], the pros and cons of hybrid feature extraction versus end-to-end fine-tuning for esophageal condition classification have not yet been fully investigated.

Apart from being accurate, explainable predictions have become essential for the use of AI models in the health sector. This can be attributed to the fact that "black box" predictions, which are opaque and lack clear justification or explanation, may impede trust among healthcare professionals in AI models and create difficulties in gaining regulatory agency approval. For the past few years, there has been a trend in XAI techniques for endoscopic image analysis, which includes Grad-CAM [7].

Inspired by such advances, we propose an interpretable deep learning architecture for the binary classification of esophageal conditions using endoscopic images. More specifically, this paper explores in-depth two paradigms of learning: (i) feature extraction via the latest advances in CNN models integrated with traditional machine learning architecture for image classification, and (ii) end-to-end optimization of CNN models to match endoscopic patterns unique to a specific domain.

Our main contributions of this work are summarized below:

- We propose interpretable deep learning models that evaluate two complementary learning paradigms: deep feature extraction combined with classical machine learning classifier models, and end-to-end fine-tuning of CNNs for the binary classification of esophageal conditions (esophagitis versus Barrett's Esophagus) using endoscopic images.
- To assess their discriminability and robustness for endoscopic image analysis in controlled experiments, multiple pretrained CNN backbones, such as ResNet50, ResNet101, InceptionV3, DenseNet201, and AlexNet, are systematically evaluated for comparative analysis.
- The proposed framework achieved solid classification results on the test set, with accuracy reaching up to 99.80% and 100% recall for Barrett's Esophagus after end-to-end fine-tuning. Grad-CAM-based visualizations were integrated to enhance model transparency by highlighting clinically relevant regions.
- We provide a comprehensive evaluation using multiple performance metrics, ROC analysis, and confusion matrices, while explicitly discussing current limitations and the need for external validation on independent datasets.

2. Related Work

While numerous previous studies have used AI methods for the detection of esophageal cancer, histopathological subtyping, and overall analysis, the current study tackles another clinical application: binary classification of esophagitis and Barrett's Esophagus from endoscopic images. A few representative studies are summarized below.

2.1 Histopathology-Based Classification Methods

Aalam et al. [8] addressed that automated binary classification of esophageal cancer using of high-resolution histopathology images. They employed the ResNet101 network, pre-trained on ImageNet, and a feed-forward network for classification via transfer learning. However, no information was provided about the dataset size, which limits the ability to replicate the results.

Zhang et al. [9] combined histopathological images with molecular genomic information (RNA sequencing and mutation analysis) to develop clinically relevant subtypes of esophageal squamous cell carcinoma. They built a DL system to extract features from whole-slide images from multiple organizations in over 300 ESCC patients. However, the method relies on imaging and genomic information, which may not be present in all clinical settings.

To detect neoplastic and precancerous esophageal tissue, including Barrett's Esophagus, directly from H&E stained histopathological slides, Tomita et al. [10] proposed an attention-based deep neural network without requiring any pixel-level annotations. They employed spatial attention modules in the form of 3D convolutional filters to detect diagnostic regions. This approach will decrease the amount of annotation work, but the high-resolution (whole-slide image) processing infrastructure needed for this model may prove to be a limitation for its use in the clinic.

2.2 Endoscopy Based Detection and Classification Systems

Zhang et al. [11] were interested in the automated detection of esophageal cancer from computerised barium esophagrams, which can be beneficial in situations where endoscopic imaging is not possible. They suggested a two-stage deep learning approach using the Faster R-CNN network that can detect and classify tumors and also provide location and spatial features. This is an extension of the general class of object detection methods for radiographic imaging of the Esophagus, but the approach is only applicable to radiographic images and may not be directly applicable to endoscopic practice.

Detection of early-stage (T1) esophageal squamous cell carcinoma from endoscopy clips has been addressed by Shiroma et al. [12] whose video consists of subtle lesions that are not usually detected during a routine examination. They created a system for real-time screening using large-scale endoscopic videos and a convolutional neural network. The model can be used to analyze video in real time, but may not be as effective with different video quality and clinical conditions.

Gao et al. [13] proposed to increase the accuracy of diagnosis and decrease the clinical workload by recognizing early esophageal cancer from endoscopic images. They used the AutoGluon automated machine learning system for automatic model and hyperparameter selection. Ensemble models are, however, black-box models, which makes them hard to interpret and can overfitting to the institution and imaging protocols used for training.

Chen et al. [14] attempted to detect small and blurry esophageal lesions in endoscopic images. They tweaked the Faster R-CNN model by introducing a technique called Online Hard Example Mining (OHEM) for better training on hard examples. The model was examined on the Kvasir-Capsule and ESOPHAGEAL datasets, which include over 5000 annotated images. It is a high-computation training method, and does not provide any means of explainability to show the reasons behind the detections.

2.3 Attention-Enhanced and Interpretable Architectures

Aftab et al. [15] tackled the problem of concurrent classification and delineation of esophageal lesions in endoscopy images. They adopted a MobileNetV2 backbone to enhance attention with modules in a multi-task learning network, with a feature extractor shared by both classification and segmentation, and decoder heads specific to each task. However, the amount of the dataset used for training was not clearly reported.

Xu et al. [16] concentrated on predicting survival and risk stratification of esophageal cancer patients using spatial relationships between tumors and lymph nodes that were modeled from computed tomography (CT) images. They used images of tumor-lymph cross-plane projection to train a ResNet architecture with a Convolutional Block Attention Module (CBAM). The study offers a contribution of spatial biomarkers, but the data are restricted to a single institution, the quality of the tumor and lymph node segmentation is a limitation, and external validation cohorts were not used.

2.4 Vision Transformer and Foundation Model Approaches

The recent development of gastrointestinal endoscopic image analysis has demonstrated the great potential of Vision Transformers and self-supervised foundation models. Takeda et

al. [17] built an AI workflow that uses a Vision Transformer (ViT-B/16-384) with color imaging for the identification of limited segment Barrett's Esophagus (SS-BE). They also accomplished high accuracy and performance on par with expert endoscopists, demonstrating the effectiveness of transformer models in esophageal disease detection. Self-supervised learning approaches have been gaining attention since there is generally no labeled endoscopic information available. Boro et al. [18] suggested using a foundation model, named EndoMAE, that was trained on more than 10 million endoscopic images in a masked autoencoder framework. For downstream tasks like image classification and polyp segmentation, the results outperformed the traditional self-supervised model. Likewise, a high number of unlabeled gastroesophageal endoscopy images were fed into a large-scale DINO pre-trained discriminative self-supervised model to detect esophagitis [19]. They found that their in-domain self-supervised pre-training method outperforms models pre-trained using the natural image dataset ImageNet. These studies suggest the use of these images to further learn more powerful and generalizable representations with Vision Transformers and foundation models. Few studies, have specifically studied these techniques specifically for differentiating esophagitis from Barrett's Esophagus, however. This is looked upon as a pledge to future research investigating the use of transformer like and self-supervised models for this particular classification.

3. Methodology

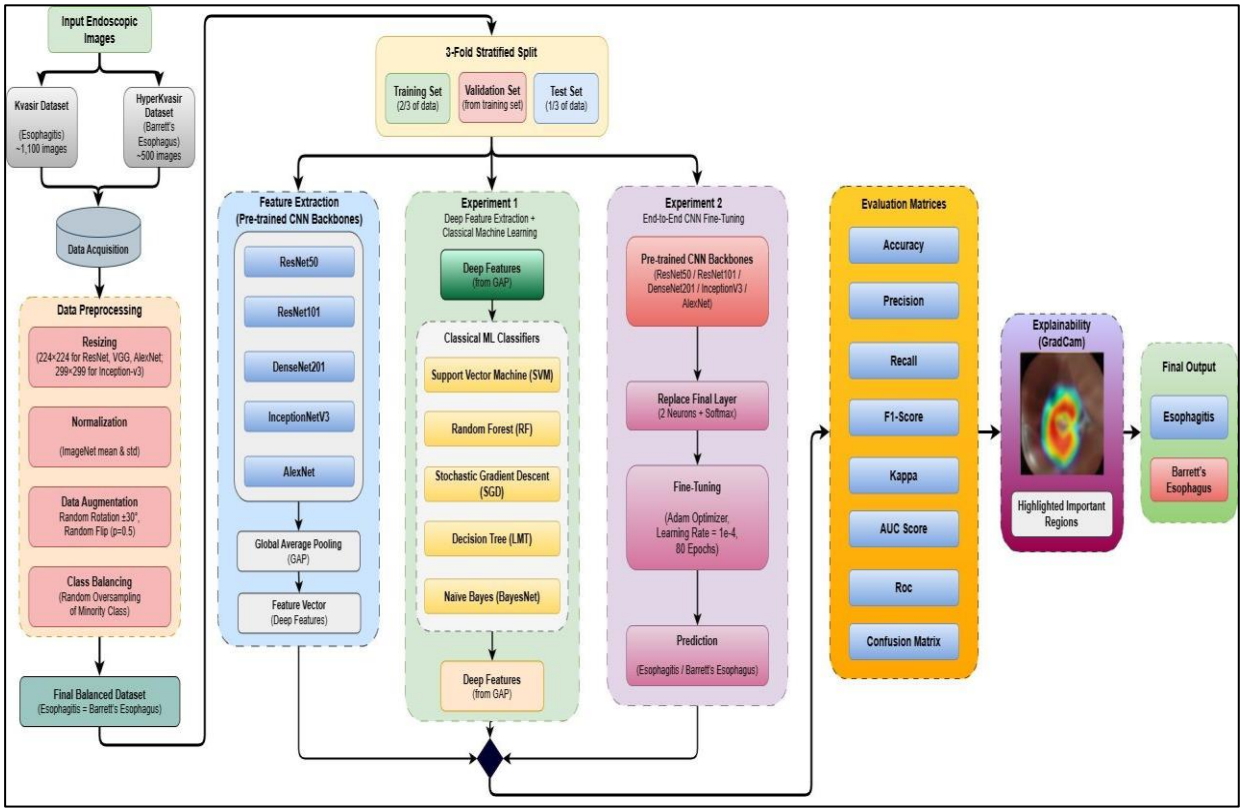

Figure 1. Overall workflow of the proposed interpretable framework

This section describes the two complementary experimental environments designed to assess the proposed interpretable deep learning framework. The entire process, as shown in Figure 1, includes explainability integration, feature learning, model training, and data preprocessing.

3.1 Datasets Description

Two databases with public gastrointestinal endoscopy images, namely the Kvasir and HyperKvasir datasets available on Kaggle, have been used in this research to provide convenience and reproducibility. Figure 2 shows the data sample.

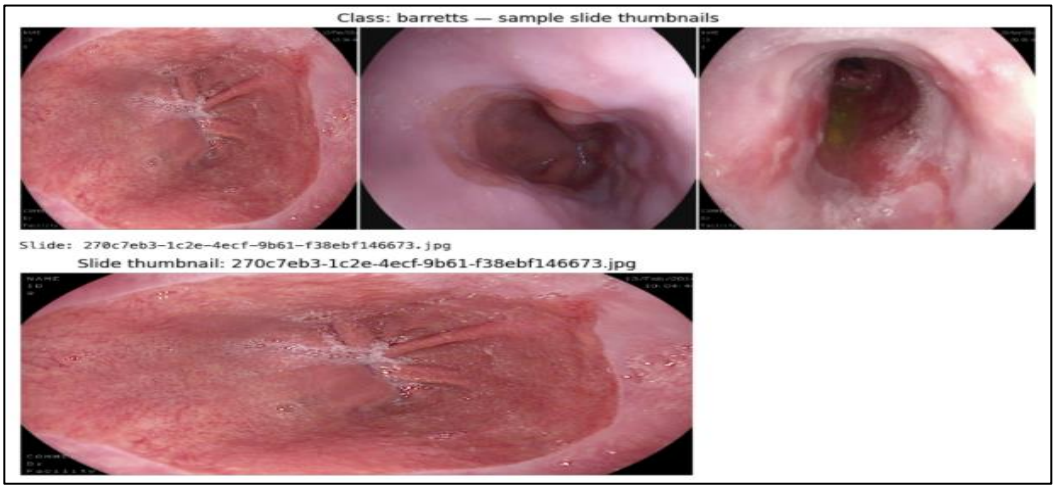

Figure 2. Representative endoscopic images from the Kvasir and HyperKvasir datasets

- **Kvasir Dataset (v1):** The chosen set of endoscopy pictures from Simula Research Laboratory called the Kvasir database on Kaggle consists of 8,000 high-resolution RGB images with equal representation of 8 clinically important classes with diseases of the gastrointestinal tract and morphological structures. The realism of intra-class variation is ensured by the fact that every class includes exactly 1,000 pictures shot from different perspectives and in different lighting conditions with the help of different imaging devices. We deal only with the esophagitis class in our investigation, which means inflammation of the esophageal mucosa. Approximately 1,100 pictures of esophagitis were deleted from the database [20].

- **HyperKvasir Dataset:** HyperKvasir, an expansion of the Kvasir project on a larger scale, is considered to be one of the largest datasets in terms of gastrointestinal endoscopy images. Images from 10,662 expert annotations in 23 categories, which include anatomical landmarks, normal images, and various disease states, were utilized in the present study. The images are annotated by expert gastroenterologists for high clinical accuracy. Images of Barrett's Esophagus from the HyperKvasir database have been classified based on their pathologies. The images had shown Barrett's Esophagus with conventional and short-segment types of this condition and around 500 images, conforming to the specifications of Barrett's Esophagus, have been chosen for the study.

- **Final Dataset Composition and Class Definition:** In this context, the goal is to formulate the problem as a two-class problem; namely, Barrett's Esophagus and Esophagitis. These are two diseases that affect the Esophagus, have visual similarities, but possess different malignant potential.

Class distribution for the original raw classes was as follows:

- **Esophagitis:** Approximately 1,100 images (dominant class obtained from Kvasir)

- **Barrett's Esophagus:** Approximately 500 images (non-dominant class obtained from HyperKvasir)

3.2 Data Preprocessing and Augmentation

To maintain uniformity and optimize feature extraction, all images were resized to 224x224 for models like ResNet and VGG, and 299x299 for the Inception-v3 model. This resizing can be mathematically formulated as follows:

$$x' = Resize\ (x, H, W) \tag{1}$$

where (x) stands for the initial RGB endoscopic image, and (H) and (W) stand for the height and width of the target size required by the corresponding pretrained backbone model. In the proposed architecture, this resize procedure serves two purposes, namely, as an input format and a standardization phase for images acquired at varying resolutions and then fed into ResNet, DenseNet, Inception-v3, and AlexNet. Thus, through transformation of all input images into a specific spatial size using Eq. (1), it becomes possible to perform feature extraction and end-to-end fine-tuning under equal input conditions. Each channel c ∈ {R, G, B} is normalized as follows:

$$x'_c = \frac{x_c - \mu_c}{\sigma_c} \tag{2}$$

where, μ_c and σ_c refer to the ImageNet's mean and standard deviation for channel (c). In this experiment, Eq. (2) adjusts the distribution of intensities of endoscopic images to that predicted by ImageNet-trained CNN backbones. It is vital since the designed architecture applies the concept of transfer learning; hence, normalizing the data allows pre-trained filters to be less affected by pixel-level changes and instead react more reliably to mucosal color and texture and illumination changes. Consequently, the features will be more comparable between images and different CNN backbones.

Furthermore, additional augmentation techniques were implemented during fine-tuning to help improve the performance of models and minimize overfitting due to the small number of samples used. These enhancements include:

Random Rotation: Images were rotated by a random angle θ ∈ [−30°, +30°].

1. Random Flips: The horizontal and vertical axes of the images were randomly flipped with probability p=0.5:

$$x' = Flip(x, p = 0.5) \tag{3}$$

2. Images were normalized according to the ImageNet channel statistics after performing augmentation to maintain consistency with the pretrained backbones.

Augmentation in Eq. (3) is done only during training to ensure that the network is exposed to reasonable variations in the orientation and perspective of the endoscopic images. As different endoscopic frames can vary due to changes in the scope, rotation, and position of frame acquisition, random flipping and rotations ensure that the model learns about mucosal patterns related to the disease rather than the image orientation. In case of binary classification, it is especially important since we have fewer samples of Barrett's Esophagus images in compared esophagitis images, and thus we want to avoid overfitting.

To avoid bias toward the most common class in the validation set, class balancing was performed prior to model training. Particularly, to balance with the amount of esophagitis images, the Barrett's Esophagus class was randomly oversampled with replacement:

$$D_{balanced} = D_{majority} \cup RandomSample\,(D_{minority}, n = \mid D_{majority} \mid) \quad (4)$$

Equation (4) shows the process implemented within the framework. Because the original dataset contains more images of esophagitis than Barrett's Esophagus, class balancing is necessary to prevent majority class domination during training. From a clinical perspective, this is significant because Barrett's Esophagus is the more critical premalignancy condition, and a class bias towards esophagitis could lead to misdiagnoses. To avoid data leakage, balancing was applied within the training folds, while validation and test folds were kept separate for objective evaluation.

The final experimental dataset consisted of 1,000 images (500 Esophagitis and 500 Barrett's Esophagus). To address class imbalance, random oversampling with replacement was applied exclusively to the training folds. No synthetic image generation methods, such as SMOTE, GAN-based augmentation, or diffusion-based synthesis, were used.

3.3 Dataset Splitting Strategy

The original dataset contained about 1,100 images of Esophagitis and 500 images of Barrett's Esophagus, leading to an imbalanced class distribution. Since one diagnosis is Esophagitis and the other is Barrett's Esophagus, the data were randomly shuffled, and an experimental dataset of 500 Esophagitis images mixed with 500 Barrett's Esophagus images was created for model training and fair evaluation. This yielded a balanced experimental dataset of 1,000 images in total, consisting of 500 esophagitis images and 500 Barrett's Esophagus images. A balanced dataset was tested using 3-fold stratified cross-validation. Stratification was used to ensure an equal representation of classes within each fold. Two-thirds of the dataset was randomly selected for model development, and one-third was used as an independent test fold. The model was then optimized by splitting the development set into training and validation subsets and early stopping if necessary. As patient information was missing in the public Kvasir and HyperKvasir datasets, image splitting was used instead of patient splitting. Thus, the results must be set within the context of image-level classification performance. However, patient-level, external, and multicenter validation is necessary prior to clinical deployment.

3.4 Experiment 1: Classical Machine Learning on Deep Features

In this experiment, pretrained CNNs are treated as fixed feature extractors, with no weight updates. The extracted feature vectors are utilized for training classicalh_i ML classifiers. Figure 3 presents the overall pipeline of Experiment 1.

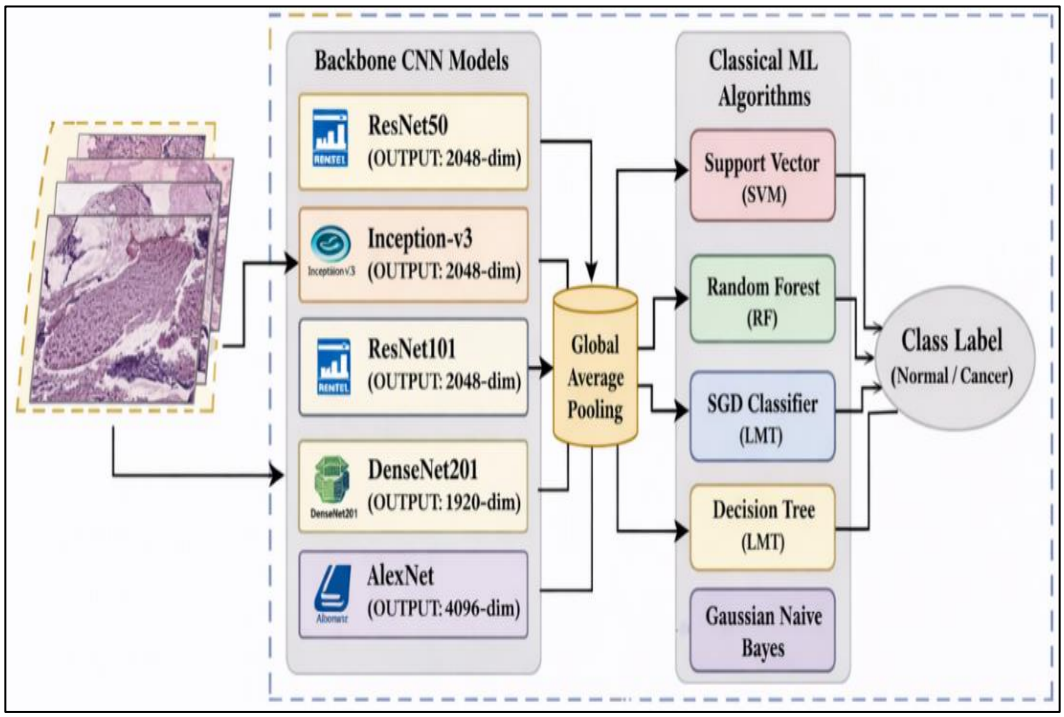

Figure 3. An overview of the experiment 1 pipeline

Feature Extraction Using Pretrained Deep Models: As backbone feature extractors, a collection of pretrained convolutional neural networks (CNNs) is used, including ResNet50, ResNet101, DenseNet201, Inception-v3, and AlexNet. For every input $x_i^{(t)}$, deep features are extracted as:

$$f(x) = CNN_{backbone}(x) \tag{5}$$

Fixed-length representations are obtained by applying global average pooling (GAP):

$$h = GAP(f(x)) \tag{6}$$

The resulting feature dimensionality can be 4096 (AlexNet), 1920 (DenseNet201), or 2048 (ResNet50/101, Inception-v3), depending on the backbone. To create a final image-level descriptor, tile-level features are combined using mean pooling:

$$h_{image} = \frac{1}{N}\sum_{i=1}^{N} h_i \tag{7}$$

The equations (5) to (7) describe the process by which the first experimental branch converts endoscopic images into a discriminative feature representation prior to classification by classical classifiers. Equation (5) learns visual patterns hierarchically from a pretrained CNN backbone, where low-level patterns include colors and edges and high-level patterns include the texture and structure of the mucosa. Equation (6) pools the spatial feature maps through global average pooling to form a fixed-size vector, allowing uniform representation for all images in the resized input format. Finally, equation (7) describes how the feature response is further summarized into an image-level descriptor, which is then passed to the classical classifiers. In this framework, these equations are crucial because they help distinguish between the feature learning and decision-making stages, hence enabling comparison between fixed deep feature extraction and end-to-end CNN fine-tuning.

Classification Models: Four classifiers are evaluated:

- Support Vector Machine (SVM): The decision function is defined as follows using an RBF kernel:

$$f(h) = sign(\sum_{i=1}^{n} y_i \alpha_i K(h, h_i) + b) \quad (8)$$

- Stochastic Gradient Descent (SGD) Classifier:

$$f(h) = w \cdot h + b \quad (9)$$

- Logistic Model Tree (LMT): Splits are chosen to maximize information gain:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \quad (10)$$

- Gaussian Naive Bayes: Conditional independence with class-wise Gaussian likelihood is assumed:

$$p\,(h \mid y = c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} exp\ (- \frac{(h-\mu_c)^2}{2\sigma_c^2} \quad (11)$$

Equations (8) to (11) describe the classifier-level decision mechanisms used to evaluate the separability of the extracted CNN features. The SVM decision rule in Eq. (8) tests whether Barrett's Esophagus and esophagitis can be separated through a margin-based nonlinear boundary in the deep feature space. The SGD classifier in Eq. (9) provides a linear discriminative baseline, helping determine whether the extracted features are already linearly separable. The information-gain criterion in Eq. (10) supports tree-based partitioning of the feature space, while the Gaussian Naive Bayes formulation in Eq. (11) evaluates a probabilistic assumption-based classifier.

Training Protocol: All classifiers are trained using 3-fold stratified cross-validation. Minority Barrett's samples are oversampled in each training fold to reduce class imbalance. To guarantee robustness and statistical reliability, performance is averaged over folds.

3.5 Experiment 2: The End-to-End process of Fine-Tuning for Deep Models

The second experiment assesses complete end-to-end learning, in which classification and feature extraction are adjusted simultaneously.

Network Optimization: The notation $g\phi$ represents the trained CNN model, whose parameters include ϕ and consists of a pretrained backbone along with the newly introduced fully connected classification layer. The output of this model for the image input x_i is the class logits $z_i = g\phi(x_i)$, which are converted to class probabilities by:

$$\hat{y}_{i,c} = \frac{\exp(z_{i,c})}{\sum_{k=1}^{2} \exp(z_{i,k})} \quad (12)$$

where $c \in \{1, 2\}$ refer to the two target categories: esophagitis and Barrett's Esophagus. Here, the significance of the softmax function arises since the function helps to convert the output from the CNN network to probabilities of belonging to certain categories, hence giving the model output a confidence measure too.

The model is optimized by minimizing the categorical cross-entropy loss:

$$L_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2} y_{i,c}\log\left(\hat{y}_{i,c}\right) \tag{13}$$

Where $y_{i,c}$ represents the ground truth value and $\hat{y}_{i,c}$ represents the predicted value of class c. This loss function is critical in the current context of binary classification since it punishes wrong predictions that are highly confident and helps the model predict a high probability for the correct esophageal pathology. The use of such a loss function is more appropriate when considering false-negative predictions of Barrett's Esophagus.

The Inception-v3 architecture uses auxiliary logits for backpropagation of gradients. The total loss was defined as:

$$L_{\text{total}} = L_{\text{main}} + 0.4\,L_{\text{aux}} \tag{14}$$

where L_{main} is the primary classification loss and L_{aux} is the auxiliary classifier loss. The auxiliary loss helps improve optimization stability in deeper networks and supports more reliable convergence during fine-tuning.

The Adam optimizer was utilized to update the network parameters:

$$\phi_{t+1} = \phi_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\varepsilon} \tag{15}$$

where $\eta = 1 \times 10^{-4}$ is the learning rate, and $\hat{m}t$ and $\hat{v}t$ are bias-corrected first- and second-moment estimates of the gradients. In the proposed framework, Adam was selected because it provides adaptive learning-rate updates, which are useful for fine-tuning pretrained CNNs on a relatively small endoscopic image dataset while reducing unstable parameter updates.

Data Augmentation and Validation: During training, data augmentation methods such as rotation, flipping, and intensity variation are used to improve generalization. To ensure consistent evaluation, 3-fold stratified cross-validation was used, consistent with Experiment 1. A batch size of 32 was utilized, and early stopping with a patience of 10 epochs was implemented based on validation loss. Table 1 presents the hyperparameter configurations for the two experiments.

Table 1. Combined hyperparameter settings for experiment 1 and experiment 2

Category	Parameter / Setting	Value / Description
Deep Feature Extraction (Exp-1)	Backbone Architectures	ResNet50, ResNet101, DenseNet201, Inception-v3, AlexNet
	Pretraining	ImageNet
	Feature Extraction	Frozen weights (no fine-tuning)
	Pooling	Global Average Pooling (GAP)
	Feature Dimension	2048 (ResNet/Inception), 1920 (DenseNet), 4096 (AlexNet)
Classical Classifiers (Exp-1)	SVM Kernel	RBF
	SVM Implementation	scikit-learn (default parameters)
	Feature Scaling	StandardScaler
	Decision Tree Criterion	Information Gain
	Naive Bayes	Gaussian Naive Bayes
End-to-End Training (Exp-2)	Backbone Architectures	ResNet50, ResNet101, DenseNet201, Inception-v3, AlexNet
	Pretraining	ImageNet
	Optimizer	Adam
	Learning Rate	1×10^{-4}

	Loss Function	Cross-Entropy
	Batch Size	32
	Number of Epochs	80
	Early Stopping	Patience = 10 epochs (based on validation loss)
	Learning Rate Scheduler	None (fixed learning rate)
	Weight Decay	0 (default)
	Hardware	NVIDIA GPU (CUDA enabled)
Evaluation (Both Experiments)	Metrics	Accuracy, Precision, Recall, F1-score, AUC, Kappa
	Cross-Validation	3-fold stratified
	Final Prediction	Softmax (Exp-2) / Majority vote (Exp-1)

3.6 Explainability and Optimization Integration

After model training, explainable AI (XAI) techniques are integrated to improve clinician trust and transparency. In particular, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to visualize class-discriminative areas in endoscopic images. Grad-CAM locates the spatial areas that significantly influence the model's decision by computing the gradient of the target class score relative to the final convolutional feature maps.

4. Results and Discussion

We applied and tested multiple deep learning classifier models in the binary classification task of esophageal diseases, specifically esophagitis and Barrett's Esophagus. The results from all models tested in both the Feature Extraction + Classical Classifiers and End-to-End Fine-Tuning experiments are presented.

4.1 Experimental Setup

Experiments were performed in a controlled computational setting where stability and reproducibility in training behavior are guaranteed. The deep learning models were implemeted with Python 3.9 and PyTorch 2.1 as the leading deep learning framework. To obtain the pretrained CNN architectures, ResNet50, ResNet101, InceptionV3, DenseNet201, and AlexNet were downloaded from the Torchvision library. The classical machine learning classifiers were implemented with the help of scikit-learn 1.4. Training and testing were conducted on a high-performance workstation that has an NVIDIA RTX 4090 GPU (with 24GB VRAM), CUDA 12.1 and cuDNN 8.9, enabling efficient parallel processing during end-to-end fine-tuning. A multi-core CPU with 32GB of RAM was also used in the process. Data pre-processing and augmentation were done with the help of standard Python libraries including NumPy, OpenCV, and Pillow. Matplotlib and Seaborn were used for all visualizations, such as confusion matrices, ROC Curves and Grad-CAM heat maps. Grad-CAM visualization was made with the help of PyTorch's implementation of gradient back-propagation. A fixed random seed value of 42 was used throughout all experiments.

4.2 Experiment 1: Feature Extraction + Classical Classifiers Performance Analysis

In this study, we present the results of Experiment 1, where we compared the performance of deep learning models for feature extraction followed by machine learning classifiers. For extracting deep features from endoscopic images, deep learning models such as ResNet50, InceptionV3, ResNet101, DenseNet201, and AlexNet were utilized. The extracted deep features were then used to train five different machine learning classifiers: SVM, RF,

SGD, LMT, and GNB. The results obtained using all combinations of models and classifiers are presented in Table 2.

Table 2. Results of experiment-1 (classical ML on deep features)

Backbone Model	Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa	AUC
ResNet50	SVM	98.70	98.03	99.40	98.71	0.974	0.999
	RF	99.30	99.21	99.40	99.30	0.986	1.000
	SGD	98.40	97.45	99.40	98.41	0.968	0.987
	LMT	97.20	95.75	98.80	97.25	0.944	0.972
	GNB	96.80	95.72	98.00	96.84	0.936	0.984
InceptionV3	SVM	98.50	97.09	100.00	98.52	0.970	0.998
	RF	98.60	98.41	98.80	98.60	0.972	1.000
	SGD	98.70	97.84	99.60	98.71	0.974	0.991
	LMT	94.80	90.62	100.00	95.07	0.896	0.948
	GNB	89.10	87.72	91.00	89.32	0.782	0.919
ResNet101	SVM	98.80	98.43	99.20	98.80	0.976	1.000
	RF	99.50	99.01	100.00	99.50	0.990	1.000
	SGD	99.10	98.24	100.00	99.11	0.982	0.995
	LMT	98.10	96.89	99.40	98.13	0.962	0.981
	GNB	93.60	93.27	94.00	93.63	0.872	0.966
DenseNet201	SVM	98.90	97.85	100.00	98.91	0.978	1.000
	RF	99.10	98.24	100.00	99.11	0.982	1.000
	SGD	99.30	98.81	99.80	99.30	0.986	0.995
	LMT	99.20	98.43	100.00	99.21	0.984	0.992
	GNB	98.40	98.42	98.40	98.40	0.968	0.992
AlexNet	SVM	98.70	98.23	99.20	98.71	0.974	0.996
	RF	99.30	98.62	100.00	99.31	0.986	1.000
	SGD	99.40	98.82	100.00	99.40	0.988	0.994
	LMT	97.20	94.78	100.00	97.30	0.944	0.972
	GNB	99.20	99.20	99.20	99.20	0.984	0.992

In general, the performances were high for all techniques, and the RF classifier performed better across all deep learning model workflows. ResNet50, ResNet101, and DenseNet201 exhibited high performance, and AlexNet also performed competitively. Among all classifiers, RF consistently achieved high accuracy, precision, recall, and AUC values in most cases. SVM also performed relatively well, particularly in recall for minority classes (Barrett's Esophagus). The best performance was obtained by RF with ResNet50, achieving 99.30% accuracy, 99.21% precision, 99.40% recall, 99.30% F1-score, and an AUC of 1.000. This was followed by SVM with 98.70% accuracy, 98.03% precision, and 99.40% recall. Poor performance was observed for other classifiers such as SGD, LMT, and GNB. In the case of InceptionV3, RF achieved 98.60% accuracy, 98.41% precision, 98.80% recall, an F1-score of 98.60%, and an AUC value of 1.000. SVM achieved 98.50% accuracy and 100% recall for Barrett's Esophagus, while Naive Bayes yielded a low accuracy of 89.10% and an AUC of 0.919. With ResNet101, RF's performance was strong, with 99.50% accuracy, 99.01% precision, 100% recall, an F1-score of 99.50%, and an AUC of 1.000. For DenseNet201, RF achieved an accuracy of 99.10%, 98.24% precision, and 100% recall. SVM and SGD also yielded similar results. Results for Naive Bayes were comparatively better than those obtained from other backbones. For AlexNet, RF achieved 99.30% accuracy, 98.62% precision, and 100% recall, with F1-score and AUC values of 99.31% and 1.000, respectively, while SGD produced similar results. Overall, the results of the classifiers were fairly consistent with those obtained by some other backbones. Considering the results presented in Table 2, it can be seen that the combination of deep learning feature extraction and classical classifiers was quite

effective in this case. In most backbones, RF performed better than the other classifiers, while SVM showed very good recall for Barrett's Esophagus in a few instances.

4.3 Experiment 2: End-to-End Fine-Tuning Performance Analysis

4.3.1 Performance Evaluation of End-to-End Fine-Tuning with Classical Classifiers

Here, in this section, we discuss our results from Experiment 2, in which pretrained models were fine-tuned end-to-end for the classification of Esophagitis and Barrett's Esophagus (BE), as listed in Table 3. In this experiment, we fine-tuned ResNet50, InceptionV3, ResNet101, DenseNet201, and AlexNet models for 80 epochs using the Adam optimizer with a learning rate of 1e-4. Instead of the last layer's outputs for the model, the output prediction of the two classes was used. All the fine-tuned models produced very good results in terms of performance. Among them, InceptionV3 with an AUC score of 1.000 was the best, with an average accuracy of 99.80%, precision of 99.60%, recall of 100%, and an F1-score of 99.80%. Another model that produced good results was ResNet50, with an accuracy of 99.70%, precision of 99.41%, and recall of 100%.

The fine-tuned ResNet101 had an accuracy of 99.60%, a precision of 99.21%, 100% recall, an F1-Score of 99.60%, and an AUC of 1.000. Similarly, DenseNet201 had 99.50% accuracy, 99.01% precision, 100% recall, an F1-Score of 99.50%, and an AUC of 1.000. As far as the other models were concerned, AlexNet, despite having relatively low precision, was able to obtain 99.40% accuracy, 98.82% precision, 100% recall, an F1-Score of 99.40%, and an AUC of 0.999. 100% Recall was achieved by all the models for Barrett's Esophagus. The findings revealed that the performance of the models increased through the process of end-to-end fine-tuning. InceptionV3 had the best performance overall and in all other categories, while the other models also performed well. It can be seen that the performance of the models post-end-to-end fine-tuning reveals that the pretrained CNN backbones had learned the domain-specific patterns such as color variation, inflammation of the mucosal layer, and epithelial texture. The better performance of InceptionV3 could be due to its multi-scale convolutional structure, which is able to recognize the local and general visual patterns in endoscopic images. The 100% recall of Barrett's cases is clinically important since there is a possibility of having Barrett's cases that could have been overlooked, resulting in a higher risk for disease progression if this is the case. The results need to be seen in the context of the internal balanced dataset, and external validation is required prior to clinical implementation.

Table 3. Results of experiment-2 (end-to-end fine-tuning)

Backbone	Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa	AUC
ResNet50	Fine-tuned ResNet50	99.70	99.41	100.00	99.70	0.994	0.998
InceptionV3	Fine-tuned InceptionV3	99.80	99.60	100.00	99.80	0.996	1.000
ResNet101	Fine-tuned ResNet101	99.60	99.21	100.00	99.60	0.992	1.000
DenseNet201	Fine-tuned DenseNet201	99.50	99.01	100.00	99.50	0.990	1.000
AlexNet	Fine-tuned AlexNet	99.40	98.82	100.00	99.40	0.988	0.999

4.3.2 Confusion Matrix Analysis

For this part, we present the confusion matrix that was derived from the end-to-end fine-tuned models, including ResNet50, InceptionV3, ResNet101, DenseNet201, and AlexNet. The confusion matrix presented in Figure 4 shows a detailed description of both correct and incorrect predictions per class.

Confusion matrices are generated using the best fold from the stratified cross-validation experiment. Each fold includes 333 images, wherein 166 are Esophagitis images while 167 are Barrett's Esophagus images. Therefore, each confusion matrix represents the classification capability of the model for this particular test fold.

The fine-tuned models were able to obtain almost perfect classification results, as seen in Figure 4. The accuracy for all models was 100% for 165/166 Esophagitis images and 99.9% for 166/166 Barrett's Esophagus images. Lastly, all 167 Barrett's Esophagus images were classified correctly, with none being falsely classified as negative. This outcome shows that in the best fold, the models achieved 100% recall for BE. This is of clinical significance because Barrett's Esophagus is a premalignant state, and minimizing false-negative results is crucial to facilitate early diagnosis and subsequent clinical management.

There was one falsely positive case, which was identified as an Esophagitis image and predicted as Barrett's Esophagus. Clinically, this type of error is less detrimental than a failure to diagnose a case with Barrett's Esophagus, since a false positive result might result in further evaluation, while a false negative result might delay diagnosis.

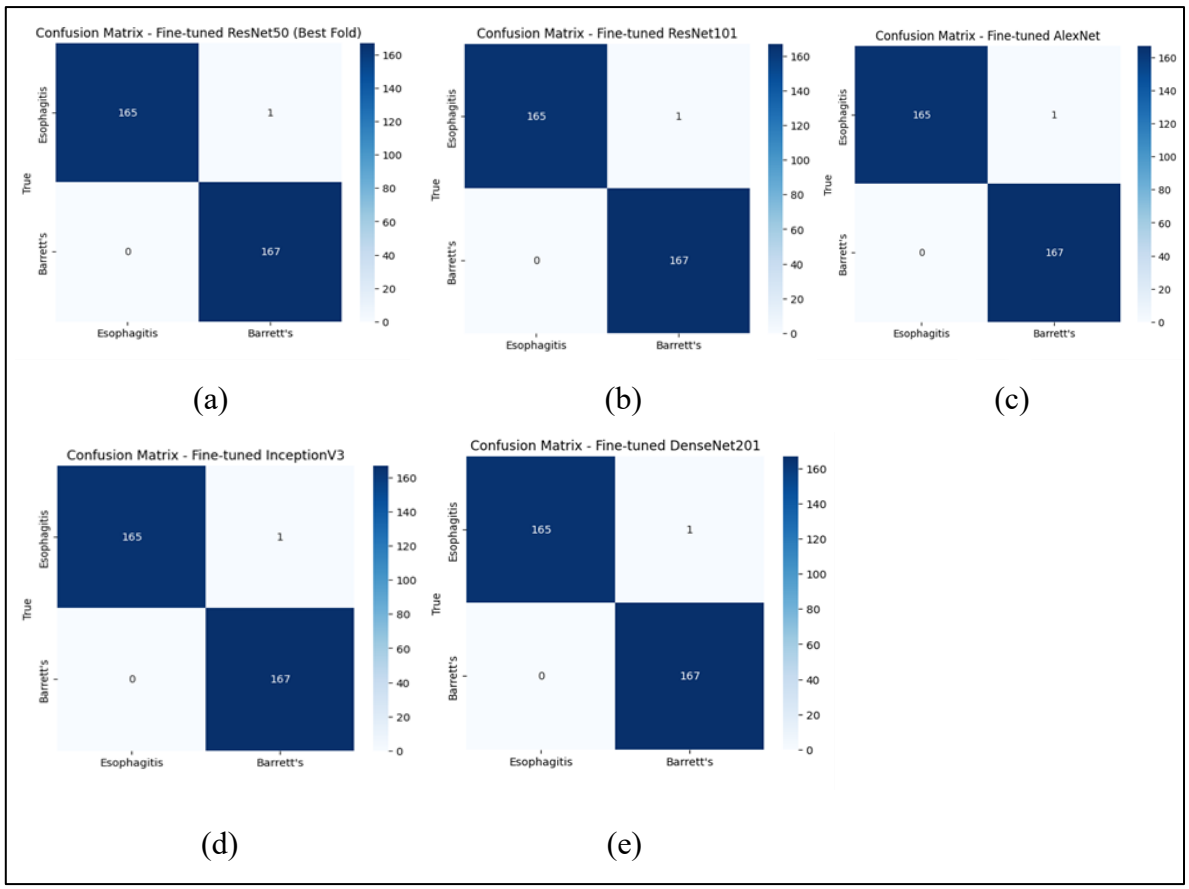

Figure 4. Confusion matrices of the five end-to-end fine-tuned CNN models

The models include (a) ResNet50, (b) Resnet 101, (c) AlexNet, (d) Inception V3, and (e) DenseNet201. Each confusion matrix summarizes the numbers of correctly and incorrectly classified images for the two classes (esophagitis and Barrett's Esophagus), where rows represent the ground-truth labels and columns represent the predicted labels. Diagonal elements indicate correct classifications, whereas off-diagonal elements correspond to misclassifications. Darker colors represent a larger number of samples within each category.

The confusion matrix analysis for each of the fine tuned models further shows that these models are robust and reliable in the classification of esophagitis and Barrett's Esophagus. All models consistently showed zero false negatives cases of Barrett's Esophagus, while being evaluated with minimal false positives and false negatives.

4.3.3 ROC Curves Analysis

The ROC curves of the fine-tuned models (ResNet50, InceptionV3, ResNet101, DenseNet201 and AlexNet) tested for classification of Esophagitis and Barrett's Esophagus are presented in this section. ROC curves are plotted in Figure 5, which shows the True Positive Rate (TPR) and False Positive Rate (FPR) of each model, and the AUC score is also provided. These curves can be used to evaluate the models' classification at different thresholds, and the AUC values are measures of the models' overall discrimination performance.

The fine-tuned models exhibit very high discriminatory power, with almost all models having an AUC close to 1.000, which indicates high accuracy is discriminating between Esophagitis and Barrett's Esophagus. The ROC curve for ResNet50 (Figure 5-a) has an AUC of 0.998, which indicates the model has very good discriminatory power between the two classes. The curve demonstrates that the model performs very well in terms of the True Positive Rate (TPR) while maintaining a low False Positive Rate (FPR), showing the effectiveness of the ResNet50 model when it is fine-tuned for this specific task. AlexNet (Figure 5-b) shows an AUC value of 0.999, very close to perfection. AlexNet has lower accuracy than InceptionV3 and ResNet101, but it still does an excellent job of recalling the information and provides good classification performance, especially in Barrett's Esophagus.

ResNet101 (Figure 5-c) also shows a very good ROC curve (AUC = 1.000), meaning that the model correctly classifies Barrett's Esophagus and Esophagitis in all cases with a recall of 100% for Barrett's Esophagus and a minimal misclassification rate. The DenseNet201 (Figure 5-d) ROC curve's AUC is 1.000, indicating the model can effectively differentiate between Esophagitis and Barrett's Esophagus. The True Positive Rate is at its optimum, indicating the model's strong recall and robustness when fine-tuned.

Lastly, the ROC curve of InceptionV3 (Figure 5-e) shows an AUC of 1.000, the best possible scenario in which the model makes no mistakes in distinguishing Esophagitis and Barrett's Esophagus. The True Positive Rate (TPR) is 1.0 for all False Positive Rates, reinforcing the model has a great classification ability.

We have provided the ROC curves for all five fine-tuned models, proving that end-to-end fine-tuning considerably improves the classification ability of the pretrained models. From the AUC values, we can conclude that these models can distinguish between Esophagitis and Barrett's Esophagus. The almost perfect recall values mean that Barrett's Esophagus will be recognized correctly by these models, which is very important for clinical use.

Perfect or almost perfect AUC values (values vary from 0.998 to 1.000) mean that not only precision but also consistency is ensured by the models. Minimum False Positive Rates and maximum True Positive Rates allow us to say that these models are very appropriate for real-world clinical use because their classification is extremely precise. Thus, ROC curves show that all fine-tuned models have maintained high true-positive rates with very low false-positive rates at decision thresholds. This feature is important for Barrett's Esophagus recognition because sensitivity is clinically more important than the maximization of the overall

accuracy value. Perfect AUC values mean that classes are well-separable; however, this result can also depend on the balanced image-level split.

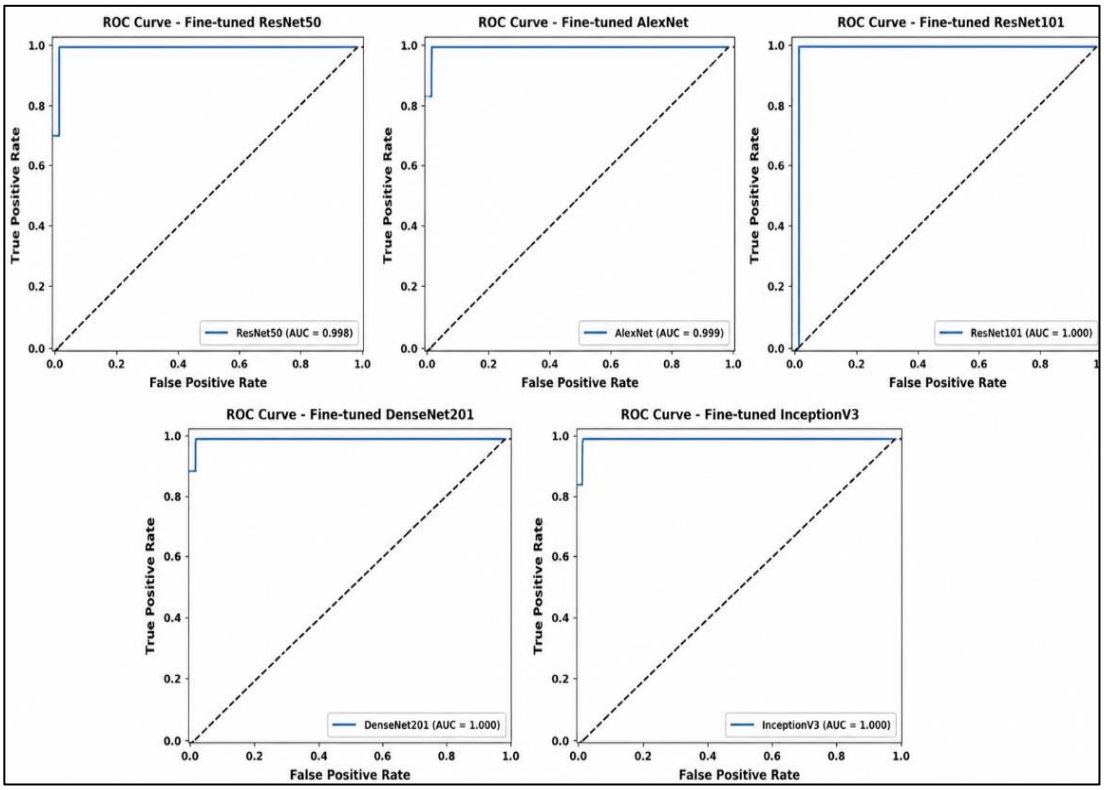

Figure 5. ROC curves of fine-tuned models

4.4 Three-Fold Cross-Validation Performance of End-to-End Fine-Tuned CNN Models

The performance of the end-to-end fine-tuned CNN models is listed in Table 4 using three-fold stratified cross-validation. The performance results are presented as mean ± standard deviation of the three folds to indicate both the average classification performance of each model and its variation across the three different folds. This method of reporting gives a better performance measurement than simply reporting the results of one test run.

Overall, all fine-tuned models achieved strong internal image-level classification performance for distinguishing Esophagitis from Barrett's Esophagus. Among the evaluated models, InceptionV3 achieved the highest mean accuracy of 99.80 ± 0.17%, with a precision of 99.60 ± 0.20%, recall of 100.00 ± 0.00%, F1-score of 99.80 ± 0.17%, and an AUC of 1.000 ± 0.000. ResNet50 also showed highly competitive performance, achieving 99.70 ± 0.30% accuracy and 0.999 ± 0.001 AUC. ResNet101, DenseNet201, and AlexNet also produced consistently high results, with mean accuracies of 99.60 ± 0.35%, 99.50 ± 0.46%, and 99.40 ± 0.52%, respectively.

A notable observation is that all models achieved 100.00 ± 0.00% recall for Barrett's Esophagus across the three folds. This indicates that, within the internal image-level evaluation, the models were highly sensitive in identifying Barrett's Esophagus cases. However, the small standard deviation values and near-perfect performance should be interpreted cautiously, as the experiments were conducted using public datasets with image-level splitting. Since patient-level identifiers were not available, possible patient-level overlap across folds could not be fully excluded.

Table 4. Three-fold cross-validation performance of end-to-end fine-tuned CNN models

Backbone	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa	AUC
ResNet50	99.70 ± 0.30	99.41 ± 0.59	100.00 ± 0.00	99.70 ± 0.30	0.994 ± 0.006	0.999 ± 0.001
InceptionV3	99.80 ± 0.17	99.60 ± 0.20	100.00 ± 0.00	99.80 ± 0.17	0.996 ± 0.003	1.000 ± 0.000
ResNet101	99.60 ± 0.35	99.21 ± 0.68	100.00 ± 0.00	99.60 ± 0.35	0.992 ± 0.007	0.999 ± 0.001
DenseNet201	99.50 ± 0.46	99.01 ± 0.89	100.00 ± 0.00	99.50 ± 0.46	0.990 ± 0.009	0.999 ± 0.001
AlexNet	99.40 ± 0.52	98.82 ± 1.02	100.00 ± 0.00	99.40 ± 0.52	0.988 ± 0.010	0.999 ± 0.001

4.5 Interpretability of the Models

4.5.1 Grad-CAM Visualization - Fine-tuned ResNet50

This section describes the Grad-CAM illustrations for the fine-tuned ResNet50 model, which provide information about the regions in the endoscopic images where the model concentrates while making predictions for Esophagitis and Barrett's Esophagus classification. Figure 6 presents a number of input images along with their respective Grad-CAM heat maps. The heat maps provide visual insights into the regions in the image that are important for the final prediction by highlighting-colored regions that show areas of focus. For Esophagitis, the model identifies areas of inflammation and lesions in the Esophagus region while for Barrett's Esophagus, it identifies abnormalities in mucosa and tissue growth. Grad-CAM visualizations assist in explaining the model's decision mechanism. Through these visualizations, one can better understand the model's decisions and insights into features important for making predictions about the two diseases.

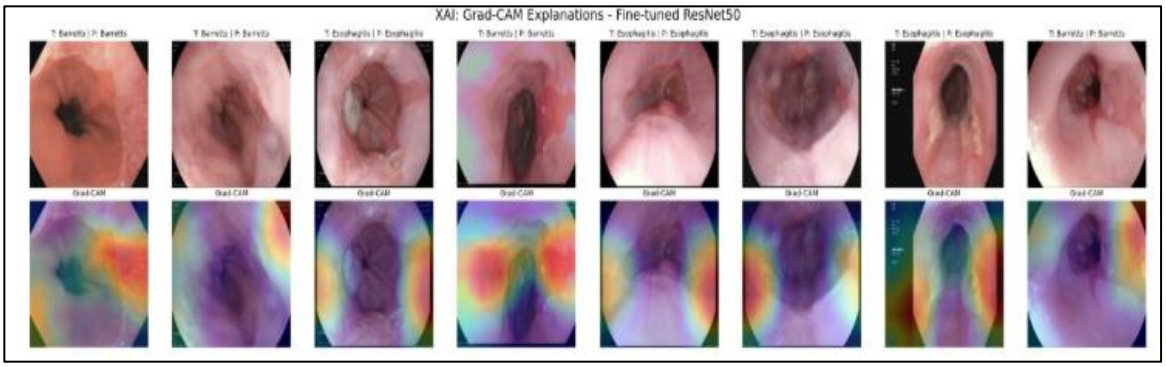

Figure 6. Representative Grad-CAM visualizations produced by the fine-tuned ResNet50 model for the binary classification of Esophagitis and Barrett's Esophagus

In all these examples, the original image captured by the endoscopy camera is displayed on the left side, while the Grad-CAM heatmap is displayed on the right side. The areas that play an important role in the prediction made by the model are those that have warm color tones (yellow and red), while the cooler colored areas (blue) play a lesser part in the model's decision-making.

4.5.2 Grad-CAM Visualization - Fine-tuned ResNet101

This part highlights the Grad-CAM visualization for the fine-tuned ResNet101 model, giving a thorough explanation of how the model arrives at its predictions concerning

Esophagitis and Barrett's Esophagus. The figure below shows the images taken using an endoscope and their respective Grad-CAM heatmaps, which represent the parts of the images most involved in the model's decision-making.

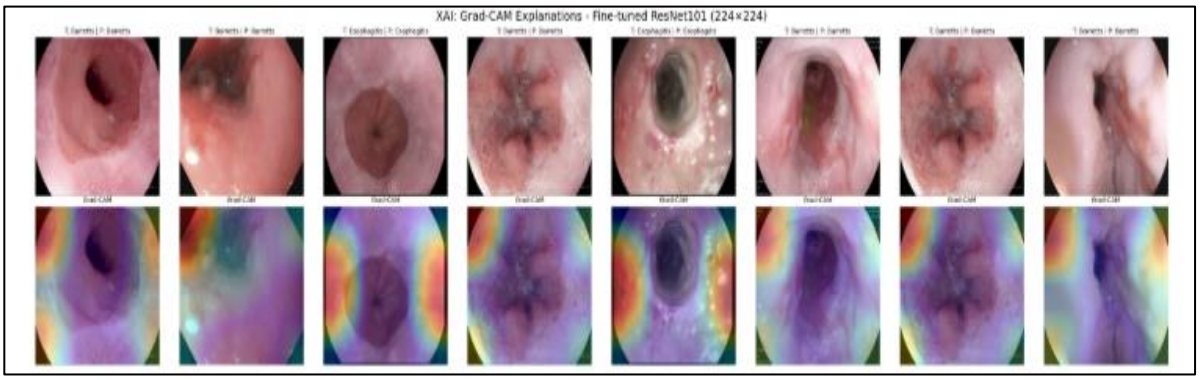

Figure 7. Representative Grad-CAM visualizations produced by the fine-tuned ResNet101 model for the binary classification of Esophagitis and Barrett's Esophagus

Each example shows the original endoscopic image (left) along with the heatmap generated using Grad-CAM technique (right). Warm colors (red/yellow) mean that those parts of the image have higher contributions to the model prediction while cold colors (blue) show lower contribution. It is clear from the heatmaps that the model consistently focuses on areas of the image that contain clinical signs of inflammation and mucosal abnormalities, which are typical for the predicted esophageal disease. Thus, the heatmaps enhance the transparency and interpretability of the classification process.

For Esophagitis, ResNet101 focuses on parts of the image that show inflammation and lesions in the Esophagus (top left quadrants of the images). For Barrett's Esophagus, the model pays attention to the mucosal abnormalities, typical of this condition. Such visualizations help explain how the model makes predictions. Heatmaps provide transparency and insight into how and why the model classifies certain images as Esophagitis and Barrett's Esophagus.

4.5.3 Grad-CAM Visualization - Fine-tuned AlexNet

This section describes the visualizations provided by the Grad-CAM method for the fine-tuned AlexNet architecture. It aims at show the regions of interest in the endoscopic images, which were used as the basis for predictions about Esophagitis and Barrett's Esophagus. Figure 8 provides an example of the input images with Grad-CAM heatmaps illustrating the model's regions of interest. The regions of interest highlighted using the Grad-CAM heatmaps include lesions, abnormalities in tissue growth, and mucosal irregularities in the Esophagus.

Regarding Esophagitis, the model looks for inflammation areas in the Esophagus, and for Barrett's Esophagus, the model identifies mucosal changes typical of the disease. The regions of interest highlighted with the help of Grad-CAM visualizations are medically relevant; therefore, one can conclude that the model's decisions are made on the basis of some interpretable features.

This can be used as an argument for the implementation of AI systems in healthcare because the interpretation of Grad-CAM visualizations makes it possible to understand how the model distinguishes Esophagitis and Barrett's Esophagus.

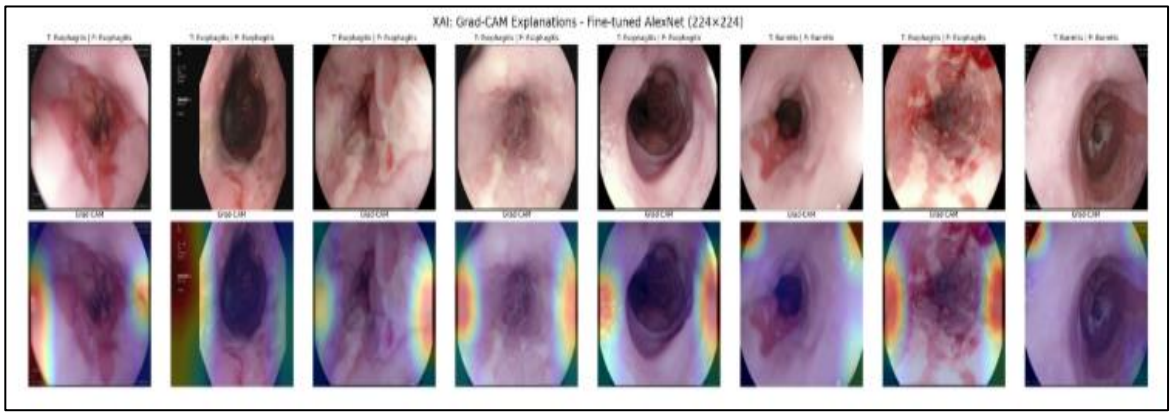

Figure 8. Representative Grad-CAM visualizations produced by the fine-tuned AlexNet model for distinguishing Esophagitis from Barrett's Esophagus.

In each case, the original endoscopic image (left) and the Grad-CAM heatmap (right) are included, where warm colors represent the parts of the image that are more important for the predicted class, while cool colors represent less important parts. The areas in focus are mainly related to the diagnostic mucosal abnormalities, which means that the predictions are based on clinically relevant features.

4.5.4 Grad-CAM Visualization - Fine-tuned InceptionV3

Grad-CAM visualizations for the fine-tuned InceptionV3 model will be discussed in this section, as they provide great insight into how decisions are made in classifying the two diseases, namely Esophagitis and Barrett's Esophagus. The visualizations are shown in Figure 9 below, along with the endoscopic images. In addition, the regions within the Esophagus that helped InceptionV3 make the prediction will be emphasized in the heatmaps.

The model concentrates on lesions and mucosal alterations, as well as the presence of abnormal tissue growth, to classify each case of Esophagitis and Barrett's Esophagus. The model learns to recognize these regions and focuses on them when making its predictions, as can be seen from the Grad-CAM heatmaps. This means that InceptionV3 pays attention to clinically relevant features when making its decisions. It should be noted that the interpretation of the Grad-CAM heatmaps may facilitate the clinical implementation of AI systems.

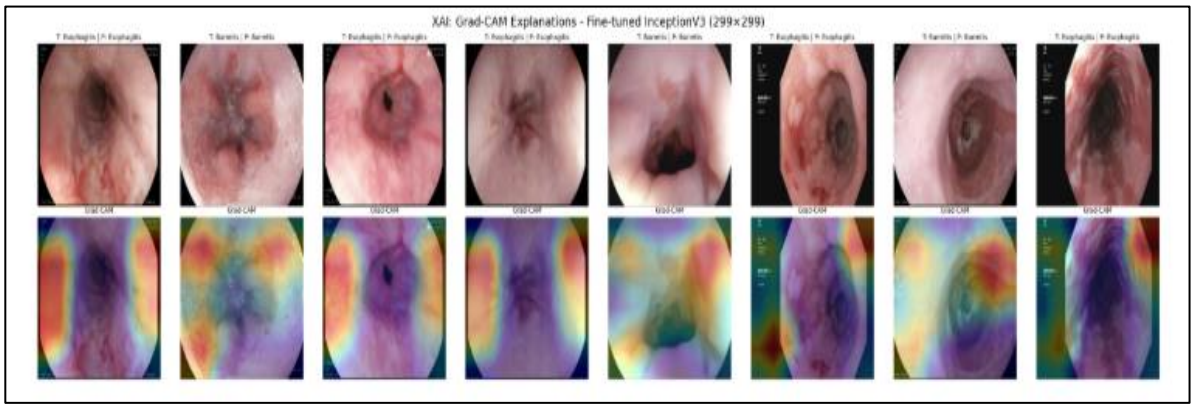

Figure 9. Representative Grad-CAM visualizations generated by the fine-tuned InceptionV3 model for the classification of Esophagitis and Barrett's Esophagus.

The following examples illustrate the endoscopic images (left) along with the heatmaps generated by Grad-CAM (right). The intensity of colors represents the weight of different parts

of the image toward the final decision. It can be observed from the visualization that the model focuses on areas with inflammation, mucosa, and abnormal epithelium, which are critical for distinguishing between the two esophageal diseases.

4.5.5 Grad-CAM Visualization - Fine-tuned DenseNet201

In this section, we discuss the Grad-CAM visualizations of the fine-tuned DenseNet201 architecture, providing insight into how this model classifies Esophagitis and Barrett's Esophagus. Figure 10 presents the original endoscopic images along with their respective Grad-CAM visualizations, highlighting the parts of the Esophagus that the model considered most important for its predictions.

In each of the examples, there is an initial image obtained by the procedure (on the left) along with its corresponding Grad-CAM heatmap (on the right). Warming colors in the heatmaps represent parts of the image that contribute more to the model's decision and colder colors represent those contributing less. Highlighted regions correspond to clinically significant mucosal or inflammatory features, which demonstrates the quality evidence of the model using clinically significant image properties for decision making.

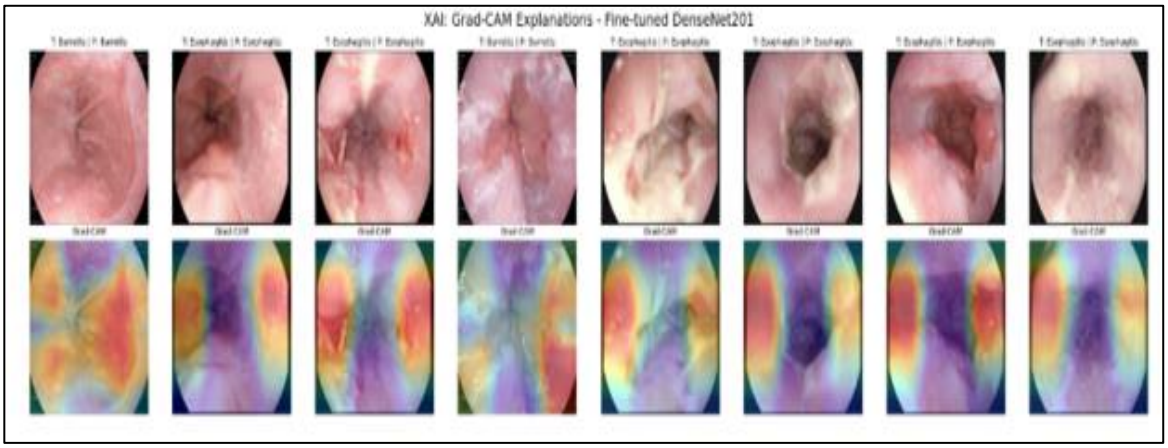

Figure 10. Grad-CAM representation generated by the fine-tuned DenseNet201 model for binary classification of Esophagitis and Barrett's Esophagus

Thus, for Esophagitis, the Grad-CAM heatmap highlights the regions of inflammation and irregularity in the mucosa, whereas for Barrett's Esophagus it highlights the region of abnormal tissue growth in the mucosa. This way, Grad-CAM allows one to identify the areas that the model uses for classification.

The explanation of the model's behavior provided by the Grad-CAM visualization tool increases the interpretability of the model. In turn, it may promote the use of AI-based systems in clinical practice. Using Grad-CAM to visualize the areas of the image that contribute to the model's decision makes the process more comprehensible for medical professionals.

4.6 Clinical Implications

The deep learning models proposed in this research study can be of clinical significance in the diagnosis of Esophagitis and Barrett's Esophagus because of their high accuracy, precision, recall, and AUC. Furthermore, by offering Grad-CAM visualizations, which provide information about the key areas of endoscopic images, the models provide interpretability that can help clinicians understand the rationale of their predictions. This can help create a

trustworthy relationship between clinicians and the AI model. One of the benefits of these deep learning models is their ability to assist in the early diagnosis of Barrett's Esophagus, which is a pre-cancerous disease. Early diagnosis of Barrett's Esophagus can help in taking preventive measures to stop it from becoming esophageal cancer.

4.7 Comparative Analysis

The performance of the proposed framework was evaluated by comparing it with some recent deep learning research on Barrett's Esophagus and the analysis of esophageal diseases, as summarized in Table 5. Massironi et al. [22] developed a narrow-band imaging (NBI) based system for the detection of Barrett's Esophagus, while Bouzid et al. [23] presented a weakly supervised approach for histopathological screening from slides stained with H&E and TFF3. Takeda et al. [17] used a Vision Transformer with linked color imaging for the diagnosis of short-segment Barrett's Esophagus, and Wei et al. [24] explored an approach called Wave-Vision Transformer for pathological classification and staging of esophageal cancer. Together, these studies demonstrate the versatility and performance of deep learning for various imaging modalities and diagnostic applications.

Within the experimental setup used in this study, the proposed end-to-end fine-tuned framework obtained 99.80% accuracy, 99.60% precision, 100.00% recall, a 99.80% F1-score, and a 1.000 AUC for the binary classification of esophagitis and Barrett's Esophagus. The results show that the model has very good discriminative ability, and being interpretable through the visualization of Grad-CAM.

However, the reported performance needs to be taken with a certain dose of caution. The proposed framework, aimed at binary classification of esophagitis vs. Barrett's Esophagus using standard endoscopy images differs from those of several previous studies that concentrated on detecting lesions, histopathological assessment or cancer staging, relying on specialized imaging modalities. As a result, it is important to note that the comparison reported in Table 5 is considered contextual; it does not necessarily imply that the framework is better than the alternatives.

Table 5. Comparative performance analysis with recent studies

Study	Year	Task / Modality	Accuracy	Sensitivity	Specificity	AUC / AUROC
Massironi [22]	2024	BE detection using endoscopic AI / NBI	94.37%	94.29%	94.44%	NR
Bouzid et al. [23]	2024	BE screening using H&E histopathology, weak supervision, external test	91.4%	72.7%	99.2%	0.873
Takeda et al. [17]	2024	SSBE diagnosis using LCI and Vision Transformer	88.6%	88.7%	88.4%	0.937
Wei et al. [24]	2025	Pathological classification/staging using Wave-ViT	88.97%	NR	NR	NR
Proposed Framework	2025	Binary classification of Esophagitis vs Barrett's Esophagus (End-to-End DL)	99.80%	100.00%	99.60%	1.000

4.8 Practical Deployment Considerations

Though the computation of inference time and memory requirements was not officially done as part of this study, all architectures evaluated are well-known CNN architectures whose computational details are well-documented. Among all the evaluated architectures, AlexNet is computationally the least expensive, while DenseNet201 and ResNet101 are computationally expensive. InceptionV3 scored highest in classification results and can be said to provide a good trade-off between performance and computational expense. This is something that shall be explored in future studies.

5. Limitations

The following limitations are associated with this study. Firstly, the experiments were performed on the public Kvasir and HyperKvasir datasets, which are based on image-level splitting because patient ID's information is not provided for this data. Thus, data leakage can occur as a result of using images from the same patients in different folds. Secondly, the dataset includes only two types of disease – Esophagitis and Barrett's Esophagus, so the application of the framework is limited to real practice concerning all kinds of esophageal diseases. Thirdly, despite consistent results obtained during the 3-fold cross-validation, the framework was not tested on external multicenter datasets, and no formal statistical tests were applied due to the comparison-based study design. Additionally, it is difficult to avoid the potential bias of publicly available datasets depending on population demographics and geography; consequently, further regulatory approval and clinical validation are needed in future research. Finally, a particular ablation study was not performed, therefore, the contribution of each module (augmentation, class balancing, feature extraction, fine-tuning, etc.) cannot be estimated. The explainability module was also estimated only qualitatively without consulting with clinicians and without comparing with other methods of explainability, such as Score-CAM and LIME.

6. Conclusion

This study represents a new interpretable deep learning framework proposed to classify esophagitis and Barrett's disease from endoscopy images. A thorough analysis was carried out on two widely popular learning methodologies—deep feature extraction and classical machine learning classifiers—as well as end-to-end fine-tuning of CNNs on a variety of popular CNN architectures. The highest accuracy of 99.80% was achieved by the best end-to-end fine-tuned model (InceptionV3), with 99.60% precision, 100.00% recall, 99.80% F1-score, 0.996 Kappa, and 1.000 AUC, while the best feature extraction-based model yielded up to 99.50% accuracy and 1.000 AUC.The experimental results showed excellent and stable performance for all models, with the end-to-end fine-tuned model demonstrating the best performance in terms of accuracy, recall, F1-score, and AUC. In particular, the 100.00% recall of Barrett's Esophagus is clinically relevant, as, from the best-fold confusion matrix, all 167 images of Barrett's Esophagus were correctly classified with zero false-negative (FN) Barrett's cases. This information suggests the usefulness of the proposed framework as a decision support tool for computer-aided gastrointestinal endoscopy, but external patient-level and multicenter validation is required before its use in clinical practice. This framework is extremely powerful and interpretable when implemented using Grad-CAM. Multi-center patient-level validation, extension to multiple esophageal diseases, tests of significance, computational benchmarking,

and ablation studies with quantification of the contribution of each part of the framework are planned for future research.

References

[1] Ebigbo, Alanna, Helmut Messmann, and Sung Hak Lee. "Artificial Intelligence Applications in Image-Based Diagnosis of Early Esophageal and Gastric Neoplasms." Gastroenterology (2025) 396-415.

[2] Ma, Yikun, Bo Li, Ying Chen, Zijie Yue, Shuchang Xu, Jingyao Li, Lei Ma et al. "Development and Validation of an AI Foundation Model for Endoscopic Diagnosis of Esophagogastric Junction Adenocarcinoma: A Cohort and Deep Learning Study." Eclinicalmedicine 89 (2025).

[3] Tabassum, Sumaiya, Md Faysal Ahamed, Hafsa Binte Kibria, Md Nahiduzzaman, Julfikar Haider, Muhammad EH Chowdhury, and Mohammad Tariqul Islam. "GastroViT: A Vision Transformer Based Ensemble Learning Approach for Gastrointestinal Disease Classification with Grad CAM & SHAP Visualization." arXiv preprint arXiv:2509.26502 (2025).

[4] Zhuang, Huangming, Jixiang Zhang, and Fei Liao. "A Systematic Review on Application of Deep Learning in Digestive System Image Processing." The Visual Computer 39, no. 6 (2023): 2207-2222.

[5] Alharbe, Nawaf R., Raafat M. Munshi, Manal M. Khayyat, Mashael M. Khayyat, Saadia Hassan Abdalaha Hamza, and Abeer A. Aljohani. "Research Article Atom Search Optimization with the Deep Transfer Learning-Driven Esophageal Cancer Classification Model." (2022).

[6] Gehrung, Marcel, Mireia Crispin-Ortuzar, Adam G. Berman, Maria O'Donovan, Rebecca C. Fitzgerald, and Florian Markowetz. "Triage-Driven Diagnosis of Barrett's Esophagus for Early Detection of Esophageal Adenocarcinoma Using Deep Learning." Nature medicine 27, no. 5 (2021): 833-841.

[7] Mascarenhas, Miguel, Francisco Mendes, Miguel Martins, Tiago Ribeiro, Joao Afonso, Pedro Cardoso, João Ferreira, João Fonseca, and Guilherme Macedo. "Explainable AI in Digestive Healthcare and Gastrointestinal Endoscopy." Journal of Clinical Medicine 14, no. 2 (2025): 549.

[8] Aalam, Syed Wajid, Abdul Basit Ahanger, Tariq A. Masoodi, Ajaz A. Bhat, Ammira S. Al-Shabeeb Akil, Meraj Alam Khan, Assif Assad, Muzafar A. Macha, and Muzafar Rasool Bhat. "Deep Learning-Based Identification of Esophageal Cancer Subtypes Through Analysis of High-Resolution Histopathology Images." Frontiers in Molecular Biosciences 11 (2024): 1346242.

[9] Jiang, Guozhong, Zhizhong Wang, Zhenguo Cheng, Weiwei Wang, Shuangshuang Lu, Zifang Zhang, Chinedu A. Anene et al. "The Integrated Molecular and Histological Analysis Defines Subtypes of Esophageal Squamous Cell Carcinoma." Nature communications 15, no. 1 (2024): 8988.

[10] Tomita, Naofumi, Behnaz Abdollahi, Jason Wei, Bing Ren, Arief Suriawinata, and Saeed Hassanpour. "Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides." JAMA network open 2, no. 11 (2019): e1914645.

[11] Zhang, Peipei, Yifei She, Junfeng Gao, Zhaoyan Feng, Qinghai Tan, Xiangde Min, and Shengzhou Xu. "Development of a Deep Learning System to Detect Esophageal Cancer by Barium Esophagram." Frontiers in Oncology 12 (2022): 766243.

[12] Shiroma, Sho, Toshiyuki Yoshio, Yusuke Kato, Yoshimasa Horie, Ken Namikawa, Yoshitaka Tokai, Shoichi Yoshimizu et al. "Ability of Artificial Intelligence to Detect T1 Esophageal Squamous Cell Carcinoma from Endoscopic Videos and the Effects of Real-Time Assistance." Scientific Reports 11, no. 1 (2021): 7759.

[13] Gao, Xin, Jiaxi Lin, Changju Qu, Chao Wang, Airong Wu, Jinzhou Zhu, and Chunfang Xu. "Computer-Aided Diagnostic System with Automated Deep Learning Method Based on the AutoGluon Framework Improved the Diagnostic Accuracy of Early Esophageal Cancer." Journal of gastrointestinal oncology 15, no. 2 (2024): 535-543.

[14] Chen, Kuan-bing, Ying Xuan, Ai-jun Lin, and Shao-hua Guo. "Esophageal Cancer Detection Based on Classification of Gastrointestinal CT Images Using Improved Faster RCNN." Computer Methods and Programs in Biomedicine 207 (2021): 106172.

[15] Aftab, Muhammad, Faisal Mehmood, Kashif Iqbal Sahibzada, Chengjuan Zhang, Yanan Jiang, and Kangdong Liu. "Attention-Enhanced Multi-Task Deep Learning Model for Classification and Segmentation of Esophageal Lesions." ACS omega 10, no. 10 (2025): 10468-10479.

[16] Xu, Jiayang, Chen Huang, Qianshun Chen, Jieyang Wang, Yuyu Lin, Wei Tang, Wei Shen, and Xunyu Xu. "Tumor–Lymph Cross-Plane Projection Reveals Spatial Relationship Features: A ResNet-CBAM Model for Prognostic Prediction in Esophageal Cancer." Frontiers in Oncology 15 (2025): 1567238.

[17] Takeda, Tsutomu, Daisuke Asaoka, Hiroya Ueyama, Daiki Abe, Maiko Suzuki, Yoshihiro Inami, Yasuko Uemura et al. "Development of an Artificial Intelligence Diagnostic System Using Linked Color Imaging for Barrett's Esophagus." Journal of Clinical Medicine 13, no. 7 (2024): 1990.

[18] Boro, Dipika, Qilei Chen, Xiaolong Liang, Yu Cao, and Benyuan Liu. "EndoMAE: a Foundation Model for Endoscopic Image Analysis via Self-Supervised Learning." In Medical Imaging 2026: Computer-Aided Diagnosis, vol. 13926, SPIE, 2026, 625-634.

[19] Friedetzki, Tobias, Naveen Chandraiah, Emil Svoboda, Pavel Pecina, Frank Puppe, and Adrian Krenzer. "Discriminative Self-Supervised Pre-Training for Esophagitis Detection in Upper GI Endoscopy Images." In Medical Imaging with Deep Learning. 2026.

[20] Pogorelov, Konstantin, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato et al. "Kvasir: A multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection." In Proceedings of the 8th ACM on Multimedia Systems Conference, 2017, 164-169.

[21] Borgli, Hanna, Vajira Thambawita, Pia H. Smedsrud, Steven Hicks, Debesh Jha, Sigrun L. Eskeland, Kristin Ranheim Randel et al. "HyperKvasir, a Comprehensive Multi-Class Image and Video Dataset for Gastrointestinal Endoscopy." Scientific data 7, no. 1 (2020): 283.

[22] Massironi, Sara. "Advancements in Barrett's Esophagus Detection: The Role of Artificial Intelligence and Its Implications." World Journal of Gastroenterology 30, no. 11 (2024): 1494.

[23] Bouzid, Kenza, Harshita Sharma, Sarah Killcoyne, Daniel C. Castro, Anton Schwaighofer, Max Ilse, Valentina Salvatelli et al. "Enabling Large-Scale Screening of Barrett's Esophagus Using Weakly Supervised Deep Learning in Histopathology." Nature Communications 15, no. 1 (2024): 2026.

[24] Wei, Wei, Xiao-Lei Zhang, Hong-Zhen Wang, Lin-Lin Wang, Jing-Li Wen, Xin Han, and Qian Liu. "Application of Deep Learning Models in the Pathological Classification and Staging of Esophageal Cancer: A Focus on Wave-Vision Transformer." World Journal of Gastroenterology 31, no. 19 (2025): 104897.