

ARTICLE

GastroNetV4: An Explainable AI-Based Hybrid Framework for Gastrointestinal Diseases Detection Using Endoscopic Images

Areeba Gul¹, Muhammad Ramzan¹, Romana Aziz^{2,*}, Qaiser Abbas³, Ala Saleh Alluhaidan², Summair Raza¹ and Mahwish Ilyas⁴

¹Department of Software Engineering, Faculty of Computing and Information Technology, University of Sargodha, Sargodha, Pakistan

²Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia

³Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia

⁴Department of Computer Science, University of Rasul, Mandi Bahaud Din, Pakistan

*Corresponding Author: Romana Aziz. Email: raabdulaziz@pnu.edu.sa

Received: 18 April 2026; Accepted: 14 August 2026; Published: 28 September 2026

ABSTRACT: Gastrointestinal diseases (GI) are serious diseases that affect people of all ages. Early and accurate diagnosis helps reduce complications and the subsequent impact on patients. Precise diagnoses by endoscopy are important for reducing complications and mortality rates. Manual interpretation of endoscopic images is time-consuming, highly dependent on the specialist's clinical judgment, and subject to variability in multi-class classification tasks. To overcome these limitations, this study proposes GastroNetV4, an explainable hybrid deep learning model for multi-class classification of gastrointestinal diseases and a urinary tract-related class included in the endoscopic dataset used in this study. GastroNetV4 employs an EfficientNet-B4 backbone with a tailored Convolutional Neural Network (CNN) head to extract high-quality discriminative features. Additionally, the proposed approach also integrates Explainable AI (XAI) methods to make AI decisions understandable and transparent. The proposed approach was compared against other publicly available pre-trained models for multi-class classification of GI diseases and a urinary-tract condition. t-Distributed Stochastic Neighbor Embedding (t-SNE) and predictive entropy analyses were used to visualize internal latent features and evaluate the model's uncertainty calibration. Experiments have shown that the proposed approach outperformed other models across various evaluation metrics. The proposed approach achieved the highest accuracy of 98.25% among all evaluated models, showing the effectiveness of the proposed enhancements over the baseline models. The study analyzed uncertainty by showing that correct predictions are associated with low entropy, whereas incorrect predictions are associated with high entropy. By achieving high accuracy, interpretability, and well-calibrated predictions, GastroNetV4 will help develop clinical decision-support systems for the automatic diagnosis of gastrointestinal diseases and other abnormalities.

KEYWORDS: Gastrointestinal; diseases; medical imaging; diagnosis; deep learning; CNN; explainable AI

1 Introduction

Gastrointestinal (GI) diseases are among the most dangerous diseases, having a severe impact on global health. Gastrointestinal diseases affect the stomach, colon, rectum, and small intestine and can cause harm and lead to life-threatening conditions like sepsis and malignancy. Digestive diseases accounted for 2276.27 million prevalent cases (95% UI 2151.23–2398.06) and 2.56 million deaths (2.39–2.72) in 2019, [1].

A number of gastrointestinal diseases (diverticulosis, neoplasms, peritonitis, etc.) are considered high-risk conditions due to the life-threatening complications for the patient. If not detected at an early stage, stomach ulcers and cancer can cause severe health problems. Peritonitis can be a life-threatening condition, causing multiorgan failure. Various ureteric conditions can cause severe pain and renal failure in the human body [2].

Conventional Endoscopy is the method used to diagnose urinary tract and gastrointestinal diseases [3]. Endoscopic diagnosis generally requires bowel preparation and, in some cases, may also require sedation. Diagnosis of gastrointestinal diseases usually depends on the skill and experience of a medical expert. The conventional method of diagnosis is labor-intensive and time-consuming, as endoscopists must continuously examine the real-time video stream to identify different gastrointestinal conditions [4].

Deep learning has emerged as a powerful tool for medical diagnosis; however, it usually demands large labeled datasets. Techniques like transfer learning, pretrained models, and data augmentation are available to alleviate this constraint, especially in gastrointestinal (GI) endoscopy [5].

Convolutional Neural Networks (CNNs), one of the prominent deep learning models, have been a useful approach in the field of medical image analysis. Because of features like textures, shapes, and edges, CNNs are well-suited for image analysis. They are widely used for image classification, detection, segmentation, and other diagnosis-related tasks.

Despite advances in medical image analysis, multi-class gastrointestinal disease detection remains a challenging task. Most public datasets and diagnostic models are focused on common descriptors and conditions. However, in situations where rare subclasses are clinically important, such as diverticulosis, peritonitis, or ureter, standard CNN models may fail to find useful representations. These models may simply overfit to the common feature patterns. To the best of our knowledge, this is the first report of a systematic analysis and performance evaluation of an endoscopic image dataset that is considered best suited for automated GI image classification.

In urinary endoscopy, Liu et al. demonstrated real-time AI-assisted detection and tracking of ureteral orifices (UOs) in cystoscopy, indicating the feasibility of AI navigation-assisted endoscopy in the field of urology [6].

The dataset contains 4000 endoscopic images divided into four classes (Diverticulosis, Neoplasm, Peritonitis, and Ureters-related condition) and does not exhibit any class imbalance. Each class contains 1000 images obtained from a real-life clinical setting, which overcomes the problem of class imbalance.

In this research, an automated deep learning model was developed to classify GI diseases, based on a hybrid CNN architecture named GastroNetV4 that combined a strong pretrained feature extractor as the backbone architecture with a custom convolutional head design to refine feature maps and extract high-level semantics of the local mucosal features in endoscopic images to indicate various GI diseases. This hybrid model attempts to get the generalization benefits of large-scale pretraining along with the flexibility and interpretability of a simple CNN in the GI domain. In wide-ranging experiments, this study benchmarked GastroNetV4 against four pretrained models: InceptionV3, ConvNeXt, MobileNetV3, and Vision Transformer (ViT).

ConvNeXt is an advanced architecture, intended to improve the accuracy and stability of CNNs. MobileNetV3 is another efficient architecture that uses depth-wise separable convolutions and attention mechanisms. InceptionV3 applies multi-scale convolutions so it can capture information at different scales. ViT is a transformer architecture that processes images as a series of patches, self-attending to model long-range dependencies. All models in the research use the same method in order to preprocess and augment

the images, and the same splits for the training, validation, and test sets to allow for a fair comparison of the classifiers.

Along with overall accuracy, the study reports per-class and macro-averaged precision, recall, and F1-scores. The study also includes confusion matrices showing the per-class errors and Receiver Operating Characteristic (ROC) curves to demonstrate the discriminative power across different decision thresholds, providing a multi-metric evaluation protocol for the proposed model. These types of visual explanations are critical for clinicians to trust models, to ensure they focus on medically relevant structures, and to diagnose failure modes.

To explicitly show the novelty and importance of this study, the main contributions of the proposed framework are summarized below in terms of methodological advances and clinical relevance. In this research work:

- Develop an intelligent framework that can classify multiple gastrointestinal (GI) diseases using endoscopic images.
- Proposed architecture combines the strengths of pretrained, CNNs and Transformers.
- Incorporate XAI techniques like Grad-CAM, which highlights important regions that contribute to the disease prediction.
- Use entropy-based density estimation to measure the uncertainty and information of the obtained features.
- Use Distributed Stochastic Neighbor Embedding (t-SNE) to observe the distribution and separability of the learned features.
- For comparison, wide-ranging benchmarking is conducted against state-of-the-art pretrained CNN and transformer models.

In summary, the presented work attempted to reduce the gap between the manual, conventional-based GI disease diagnosis approaches and novel AI-assisted automated ones. For this purpose, the study has proposed a customized hybrid CNN model to effectively diagnose different types of GI and one urinary tract condition. To justify the proposed approach, a thorough performance analysis against recent pretrained architectures was conducted. Results have shown that GastroNetV4 clearly outperformed these models. This clearly suggests that models like this could one day be used as decision-support tools in the endoscopy room, potentially reducing missed diagnoses. The proposed approach also integrates Explainable AI (XAI) methods to reason about decisions.

2 Literature Review

In this study, the application of deep learning is explored for the analysis of endoscopic images. Various existing approaches that help detect GI diseases were thoroughly investigated. According to the recent Global Epidemiology estimates, digestive diseases continue to account for billions of incidents and millions of deaths worldwide each year [7]. The impact of these diseases on the global healthcare system remains substantial across all regions of the world. Furthermore, recent Global Cancer Statistics show that colorectal and gastric cancer remain the leading causes of cancer-related death, highlighting the need for endoscopic screening [8]. Furthermore, observational studies in Asia, Europe, and North America have found increasing incidence rates for inflammatory bowel disease, diverticular disease, and other chronic GI diseases, all of which have a serious impact on healthcare and quality of life [9]. Therefore, there is a strong need for accurate, scalable methods to support the diagnosis of these diseases.

Multiple studies have been conducted on the application of AI techniques to polyp detection and characterization during colonoscopy [10,11]. CNN-based real-time polyp detection systems have been used

to improve adenoma detection rates and reduce adenoma miss rates in several clinical trials [12]. Computer-aided diagnosis (CAD) has been used to considerably reduce withdrawal time during colonoscopy in randomized clinical trials, although the rate of lesions found has not been considerably changed. While CAD may assist physicians in diagnosing a broad range of GI diseases, difficulties remain with respect to dataset coverage, the ability to classify rare disease entities, and the interpretability of image analysis models.

Some systems employ frame-level CNN classifiers, while others can capture temporal patterns by implementing recurrent networks or attention mechanisms that can identify subtle abnormalities across long segments of video [13]. Studies on clinical datasets have shown that the use of AI can considerably reduce reading time without reducing sensitivity for clinically relevant lesions, potentially increasing the scalability of capsule endoscopy in busy endoscopy centers [14,15].

Recent studies have gone beyond the binary classification of “polyp vs. background” and “normal vs. abnormal” by adopting a multi-class GI disease classification task [16]. A few studies have also created datasets of multiple pathologies of the GI tract, such as multiple polyp histologies, colitis subtypes, or different grade dysplasia, and use these to train deep CNNs to distinguish between the different classes [17].

Shoukat et al. proposed machine learning-helped electrochemical SERS to detect multiple proteins in urine, creating an early detection platform for diseases through biochemical urine analysis [18]. On another note, Matov provided a perspective on several urinary biomarkers for systemic diseases such as lung cancer, showing how urine-based detection techniques could provide an understanding of diseases outside the genitourinary system. These studies are part of a broader trend in AI-enabled multiorgan and multimodal disease detection [19].

The EfficientNet or Xception-based architectures have also outperformed older backbones (like VGG or ResNet) on multi-class classification when combined with wide-ranging data augmentation and hyperparameter tuning [20]. Razali et al. proposed that the multi-class classification performance has been improved by class-balanced sampling or cost-sensitive training to address the class imbalance problem [21]. Currently, most of the multi-class datasets still mainly focus on the relatively common pathologies.

Utku proposed a hybrid ConvViT architecture as a combination of CNNs and ViT for the classification of GI diseases in endoscopic images. The combination of local feature extraction in CNNs and the global attention used in a ViT was effective for capturing visual patterns, which, when used to classify the images, achieved an accuracy of 95.87% [22].

Houmaidi et al. used Grad-CAM to explain the outputs of DeepGI, a model that used MobileNetV2, VGG16, and Xception CNN architectures to classify different gastrointestinal diseases based on 4000 endoscopic images. Results showed that VGG16 and MobileNetV2 provided the highest test accuracy score of 96.5%. This work provided a solid baseline performance for automated diagnosis and a proof of principle that deep learning methods can solve complex real-world medical imaging problems [23].

Several studies in the literature have pointed out potential areas for future research. More recent reviews of AI-assisted colonoscopy have noted that, although real-time CAD systems have improved polyp detection rates, most of these systems have been trained and validated on datasets from a single hospital, which limits generalizability [24,25].

Naseem et al. [26] introduced a DeepNeck architecture after modifying NasNetMobile and ResNet50 with bottleneck architectures, feature fusion, and a modified Whale Optimization Algorithm for the elimination of superfluous features and to tackle class imbalance. The framework attained a 96.0% accuracy on Hyper-Kvasir and 98.9% on Kvasir-V2, which indicates good GI disease classification.

Fatima et al. [27] presented dual-branch CNN–Transformer architecture integrating Extended BEiT and EfficientNet-B5 features with model stacking to mitigate misclassification in the diagnosis of GI diseases.

The framework attained 94.12% accuracy/recall/F1-score and 94.15% precision on the Kvasir dataset and hence confirmed enhanced diagnostic performance. In small-bowel endoscopy literature, incomplete expert annotations are referred to as the major bottleneck for reproducible research and clinical translation [28].

Su et al. developed GI-ScreenNetv2, a novel hybrid deep learning approach for the improvement of the diagnosis of gastrointestinal diseases. By combining the strong global modelling power of ViT and the local feature extraction of EfficientNet-B3 with a cross-attention mechanism, the authors achieved a remarkable improvement in representation. The dual-stream method achieved an accuracy rate of 94.87%, which was more accurate than the single-stream architectures [29].

Chou et al. utilized artificial intelligence to analyze hyperspectral imaging in the classification of GI diseases. Using spectral and spatial dimensions beyond the visible RGB, GI diseases were classified, and subtle GI pathologies were detected better [30].

Tsai et al. (2025) proposed a deep learning system using spectral visualization technology to detect early GI disease, improving mucosal and vascular pattern visualization, and addressing the difficulty of detecting early-stage lesions with nonspecific or subtle manifestations [31].

To segment irregular organs with unclear boundaries, Ju et al. [32] proposed UAMSNet, which utilizes boundary-improved attention and adaptive feature extraction modules, specifically for abdominal multi-organ segmentation. CLDSINet, presented by Ju et al. [33], further combines the learning of static and dynamic information for target deformation and medical image segmentation.

Despite the progress summarized above, several gaps in the literature have motivated this study: While the majority of the publicly described datasets/models are focused on common entities such as colorectal polyps or high-level “abnormal” categories, there has been little representation for Diverticulosis, Neoplasms, Peritonitis, or pathology associated with the Ureters. Few studies have been specifically explored these four clinically important conditions. There is limited use of Grad-CAM applied to these diseases. To address these issues, a new hybrid architecture called GastroNetV4 is proposed. The same training protocol for GastroNetV4 is used and compared against ConvNeXt, MobileNetV3, InceptionV3, and ViT.

Overall, in-depth studies of the literature have shown that deep-learning-based CAD tools have already transformed various aspects of GI imaging (notably polyp detection, early upper-GI neoplasia, and Capsule Endoscopy (CE) reading), but the important shortcomings related to gastrointestinal diseases still need to be addressed. An additional key factor is that publicly available datasets are often dominated by high-prevalence diseases, but clinically important and less common conditions remain underrepresented in a real-world multi-class dataset. Furthermore, many studies report data from single-source or proprietary datasets, making it difficult to generalize observations regarding model robustness to other imaging devices, data acquisition protocols, and populations. Some explainable AI methods (e.g., Grad-CAM visualization) have also been considered in GI CAD, but they have mostly been used as illustrative tools rather than systematically using them to study class-specific behavior and failure modes. The above gaps inspired to introduce a study that: (i) addresses a balanced four class problem with under explored GI disease categories, (ii) investigates a hybrid architecture with a rich pretrained feature extractor followed by a task-specific CNN head, (iii) compares the performance with strong CNN and transformer-based baselines under the same preprocessing, augmentation and optimization steps, and (iv) uses rich quantitative metrics with visual explanation techniques. The proposed study is designed to fulfill these requirements and to provide a reliable and reproducible baseline for future studies.

3 Material and Methods

This section provides the details of the proposed methodology and experimental setup.

3.1 Dataset Description

The publicly available endoscopy image dataset (Medical Imaging Dataset) is used [34], which contains colour images taken during a clinical endoscopy procedure. This dataset contains 4000 samples. Three classes belong to gastrointestinal diseases, while the 4th one is related to a urinary tract condition. Each class comprises 1000 images taken from different angles and exposures, with varying mucosal characteristics and the presence of fluids or debris. Representative contrast images for each of the four disease classes are shown in Fig. 1, with a wide variety of anatomical locations and lesion morphologies in each column of the figure.

Table 1 shows the number of samples per class before the split.

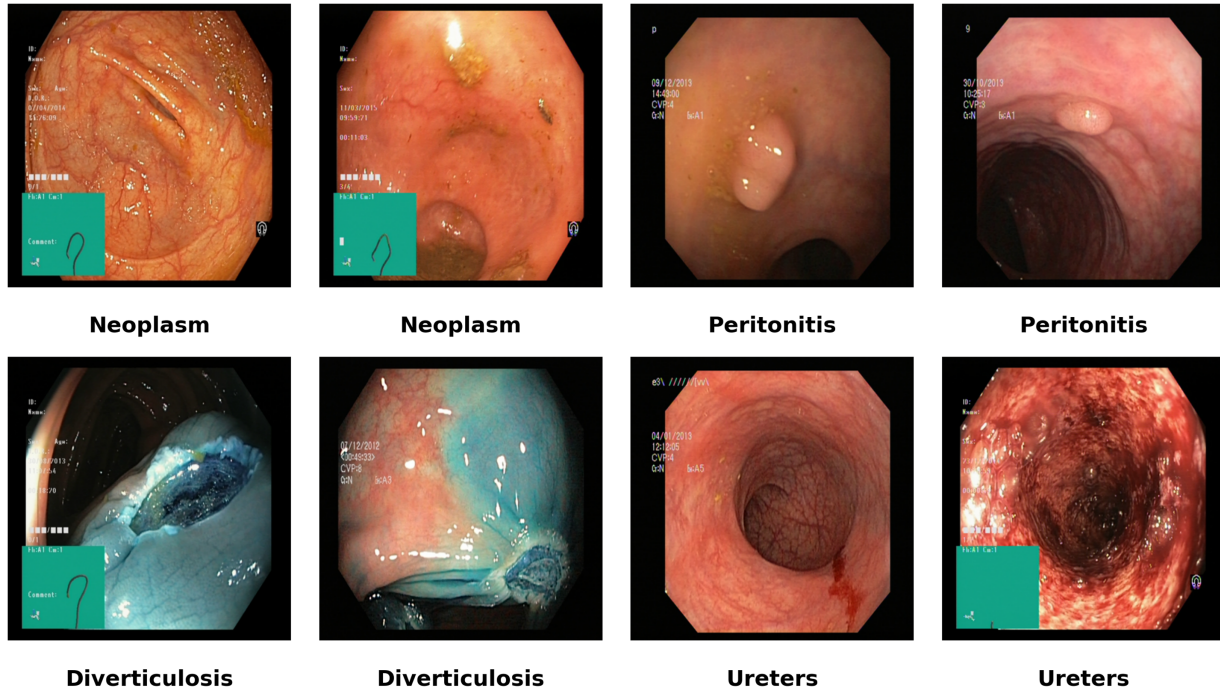


Figure 1: Different images from four classes of the dataset.

Table 1: No. of samples in different classes of the dataset.

Classes	No. of Samples
Diverticulosis	1000
Peritonitis	1000
Neoplasm	1000
Ureters	1000

The data set used in this study is exactly balanced in class sizes. The original data sets contain 1000 images for each class. To maintain such a class distribution, no under-sampling, over-sampling, synthetic augmentation or manual class balancing techniques were used. For the experiments, original distributions of the datasets is used in order to maintain the integrity of the dataset. For experimental purposes, the dataset was divided into training, validation, and testing subsets using an 80:10:10 ratio. A stratified random splitting strategy was applied at the image level using a fixed random seed (SEED = 42) to preserve class distributions consistently across all subsets. The publicly available dataset does not provide patient identifiers; therefore,

splitting at the patient level was not possible. Consequently, the partitioning was performed at the image level, which may allow images from the same patient (if present) to appear in different subsets. This limitation is inherent to the dataset structure and should be considered when interpreting generalization performance.

3.2 Data Preprocessing

All of the endoscopic images in the dataset are resized to the same spatial resolution as required for the input images of the backbone models. In the proposed hybrid EfficientNet-based model and comparison models, the images are resized to 448×448 pixels with three color channels. This experiment uses the same hyperparameters: input resolution measures 448×448 , and pixel intensity normalizes to $[0, 1]$ through division by 255.

$$X_{norm} = \frac{X}{255} \quad (1)$$

here, X represents the original RGB image tensor, whereas X_{norm} is the normalized image. For pretrained models with ImageNet weights, the channels are also standardized with the empirical mean and standard deviation $\mu \in \mathbb{R}^3$ and standard deviation $\sigma \in \mathbb{R}^3$ of the ImageNet dataset:

$$\tilde{X} = \frac{X_{norm} - \mu}{\sigma} \quad (2)$$

This normalization improves numerical stability and aligns the dataset with the statistics of the pretrained backbones. Data augmentation is used only for the training set. It seeks to generalize the model and to reduce overfitting with a stochastic data augmentation operator $T(\cdot)$, which is sampled from a family of medically safe transformations T , and each training image X is transformed as:

$$X' = T(X), T \sim T \quad (3)$$

Fig. 2 shows various augmented images obtained from a single original image.

8 Augmented Samples (Class: Neoplasm)

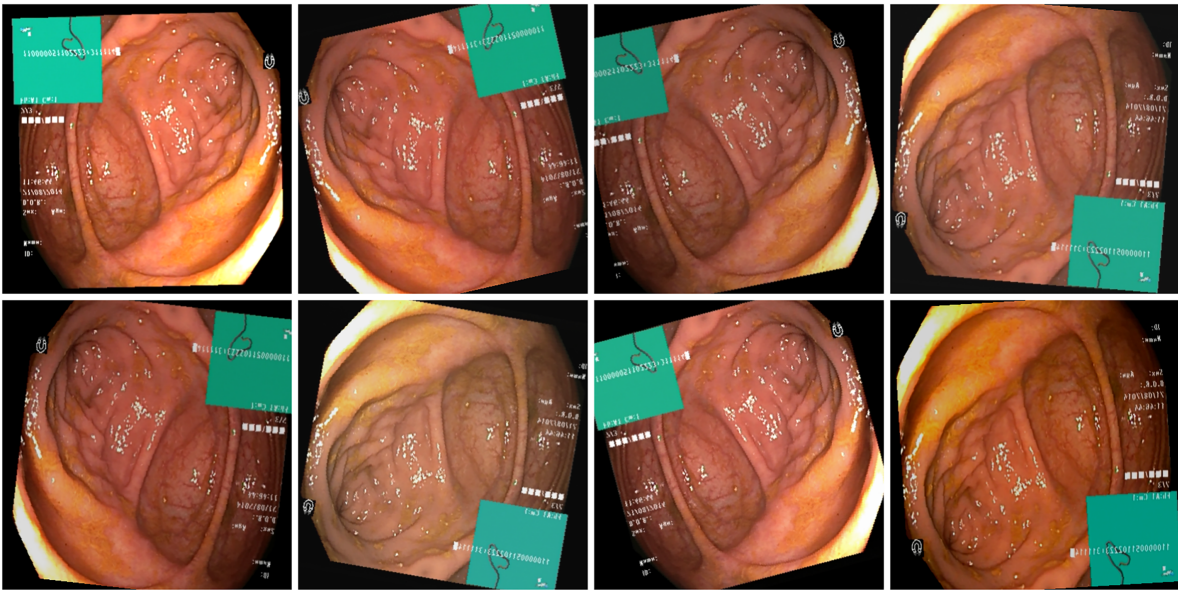


Figure 2: Various augmented images obtained from a single original image.

The augmentation set T includes:

- Horizontal and vertical flipping;
- Random rotations in a range of $\pm 15\%$;
- Random zooming by $\pm 10\%$;
- Random translation of up to 10% of the image width/height;
- Brightness and contrast can vary by $\pm 15\%$.

Unlike the training images, the validation and test images are not augmented; instead, they are resized and normalized.

3.3 Feature Extraction and Classification

The proposed methodology evaluates both the pretrained models and the proposed model under a single training protocol. The model is trained with an image $I = \mathbb{R}^{h \times w \times 3}$ as an input and a target class of interest, $y \in [1, \dots, C]$ is the ground-truth label, where $C = 4$ is the number of categories of GI diseases. Fig. 3 shows the complete workflow of the whole experimentation performed.

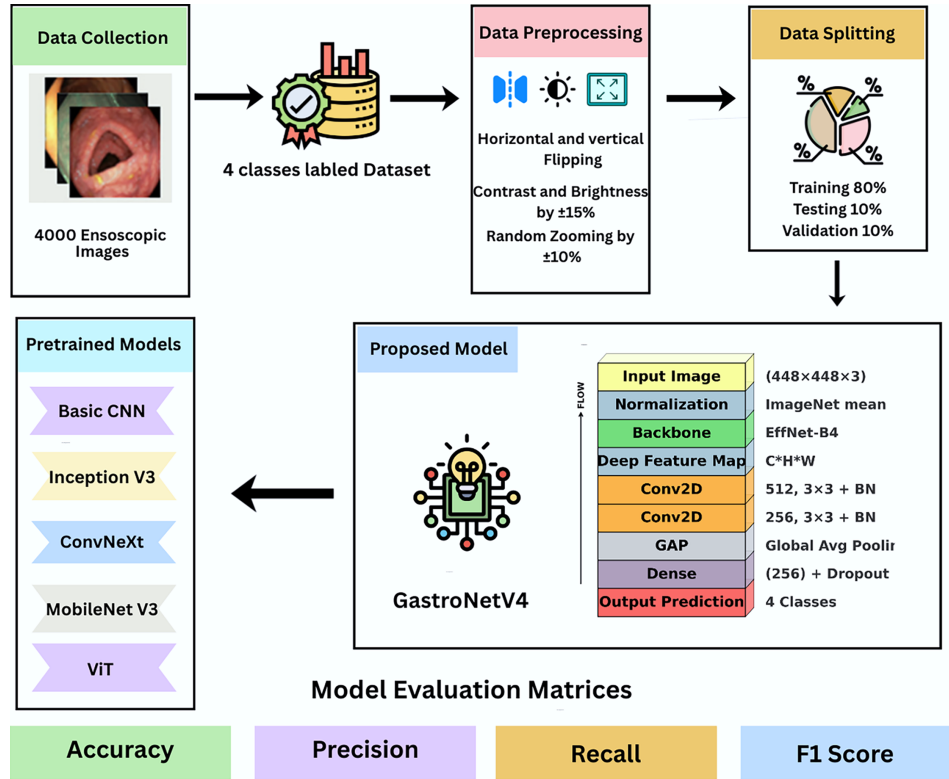


Figure 3: Workflow of the whole experiment performed.

3.3.1 Pretrained Models

In this study, four different pretrained models are evaluated, including InceptionV3 [35], ConvNeXt, MobileNetV3, and ViT. For each model, the final classification layer is replaced with a new fully connected layer of size $C = 4$. The other layers are initialized to ImageNet-pretrained weights, and then fine-tuned on the GI dataset, while the last uses SoftMax. The image I is mapped through a backbone feature extractor $F_{\theta_B} = (\cdot)$ with parameters θ_B to generate a feature tensor:

$$H = F_{\theta_B}(I) \in \mathbb{R}^{h \times w \times d} \quad (4)$$

where h , w , and d are the spatial and channel dimensions of the final feature map. A global average pooling (GAP) operation aggregates the spatial d -dimensional feature vector z :

$$z_k = \frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w H_{i,j,k} \quad (5)$$

The networks are then trained with a standard multi-class cross-entropy loss over a mini-batch of size N .

$$L(\theta_B, \theta_C) = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C 1[y_n = c] \log \hat{y}_{n,c}, \quad (6)$$

Training is performed on a Kaggle instance with a GPU P100. For a fair comparison, all the pre-trained baseline models are run with the same experimental settings and splits, and are trained with a fixed input resolution of 448×448 pixels and a batch size of 12. For the training dataset, the same data augmentation pipeline was applied. The optimizer used was AdamW with a learning rate of 3×10^{-4} and a weight decay of 1×10^{-4} .

3.3.2 Proposed Model

GastroNetV4 is a hybrid architecture comprising an EfficientNet-B4 backbone pretrained on ImageNet and a custom convolutional head. The backbone learns generalized features, while the head is designed specifically to learn GI-related features. Let $F_{\theta_B}^{Eff}$ denote the EfficientNet-B4 feature extractor truncated before its original classification layer. For each image, I , the EfficientNet-B4 backbone outputs a high-level feature map. The feature map is processed by two customized convolution blocks: a 3×3 convolution block consisting of 512 filters with batch normalization and Rectified Linear Unit (ReLU), and a 3×3 convolution block consisting of 256 filters with batch normalization and ReLU. A global average pooling layer is also used. The feature map of the basic CNN is given as:

$$H^{(l)} = \sigma \left(BN^{(l)} \left(Conv^{(l)} \left(H^{(l-1)} \right) \right) \right), H^{(0)} = H \quad (7)$$

where BN denotes batch normalization, and $Conv$ is the learnable convolution. For the EfficientNet-B4 backbone of GastroNetV4, all layers are initialized with initial weights trained on ImageNet. To initialize the EfficientNet-B4 backbone of GastroNetV4, pretrained weights from ImageNet were used. GastroNetV4 was also trained in an end-to-end manner throughout the entire training without stage-wise freezing/unfreezing. The model was trained for 50 epochs with the AdamW optimizer at the initial learning rate 2×10^{-4} and weight decay of 1×10^{-4} . The hyperparameters were chosen empirically based on a combination of performance on validation data and stability of training. Models were trained using the AdamW optimizer with an initial learning rate of 2×10^{-4} and weight decay of 1×10^{-4} using a cosine annealing learning rate schedule, in which the learning rate is gradually reduced. An input image size of 448×448 was chosen to maintain details from the endoscope videos, with a batch size of 12, chosen to allow stable optimization while being compatible with the hardware. The dataset was augmented by random resized cropping, flipping, rotating and jittering of the color to improve generalization and reduce overfitting. These settings have all tended to converge well and lead to good validation results. A cosine annealing schedule is used with a cosine cycle of length E_2 , to further tune the backbone to the endoscopic image classification domain while ensuring limited catastrophic forgetting, leading to a final test accuracy of around 98.25%.

Besides the architecture design, one of the key contributions of GastroNetV4 is the different reliability-aware evaluation tools employed after the transfer learning pipeline. While classical transfer learning pipelines aim for accuracy, there is addressed explainability through Grad-CAM, latent space inseparability through t-SNE, uncertainty estimation through predictive entropy, and model calibration through ECE. The method thus eases multi-class classification with accurate and interpretable predictions as well as safety assessments, which is of utmost importance in real-world clinical decision-support systems. To the best of knowledge, combined classification and safety assessment approach has not been applied systematically to a balanced four-class gastrointestinal dataset before. Fig. 4 shows workflow of the proposed model.

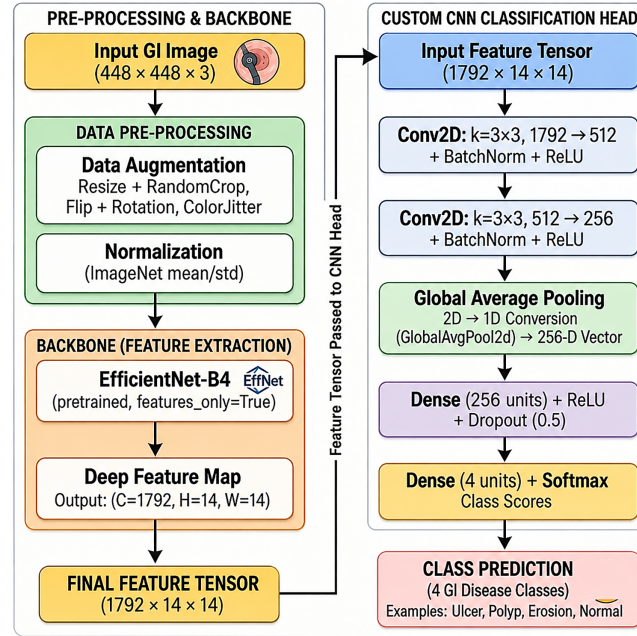


Figure 4: Workflow of the proposed GastroNetV4 model.

3.4 Performance Measure

Performance of the proposed model was evaluated using the test set. The following metrics were taken into consideration:

- Precision ($\frac{TP}{TP+FP}$)
- Recall ($\frac{TP}{TP+FN}$)
- F1 Score ($\frac{2PR}{P+R}$) provides mean of precision and recall
- Overall accuracy ($\frac{TP+TN}{Total}$) shows the proportion of predictions that were classified correctly
- Confusion Matrix
- Multi-class ROC curve with AUC values

These measures were used to evaluate the classification performance of the proposed model both at the overall and class level.

4 Results

This section presents the results obtained using the proposed deep learning framework for multi-class Gastrointestinal disease detection. The performance of the pretrained models was evaluated on test samples using metrics discussed above and compared with the performance of the proposed model.

The InceptionV3 achieved a final test accuracy of 98.00%. As shown in Fig. 5, the InceptionV3 model has given the best results for the Diverticulosis and Neoplasm classes. It did not misclassify any image from diverticulosis, whereas only one image was misclassified from the neoplasm class. This model has the worst results for the Peritonitis class, as it mistakenly classified four images of the Peritonitis class as Ureters, but this model also gave promising accuracy after our proposed model.

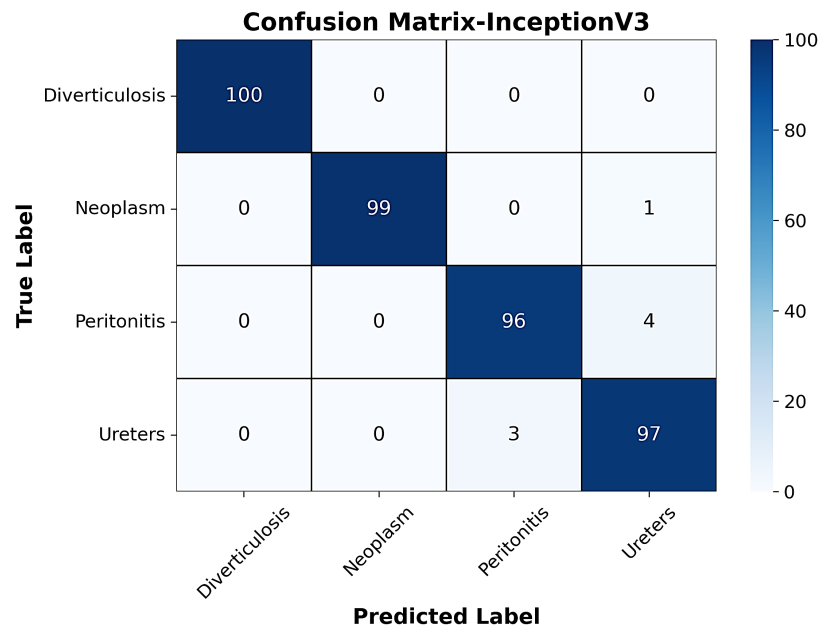


Figure 5: Confusion matrix of InceptionV3 model.

Fig. 6 shows the confusion matrix of the ConvNeXt model. It can be seen that the ConvNeXt architecture correctly predicted all 100 diverticulosis condition images. Further, the best performance was achieved by the Ureter class. However, classification performance was comparatively low for the other three classes. The gap in performance may be due to dataset scale sensitivity, as ConvNeXt architectures tend to benefit from larger, scaled training data and regularization.

Another pre-trained model called MobileNetV3 was also used to compare its performance with the proposed model. This model achieved a final test accuracy of 97.50%. Fig. 7 shows that the MobileNetV3 model performed the best prediction on the diverticulosis class with no misclassified images. The model showed the least good performance for the peritonitis and Ureter classes. Five images of the ureter were incorrectly predicted as Peritonitis. For the Peritonitis class, four images were wrongly predicted as Neoplasm and Diverticulosis, respectively.

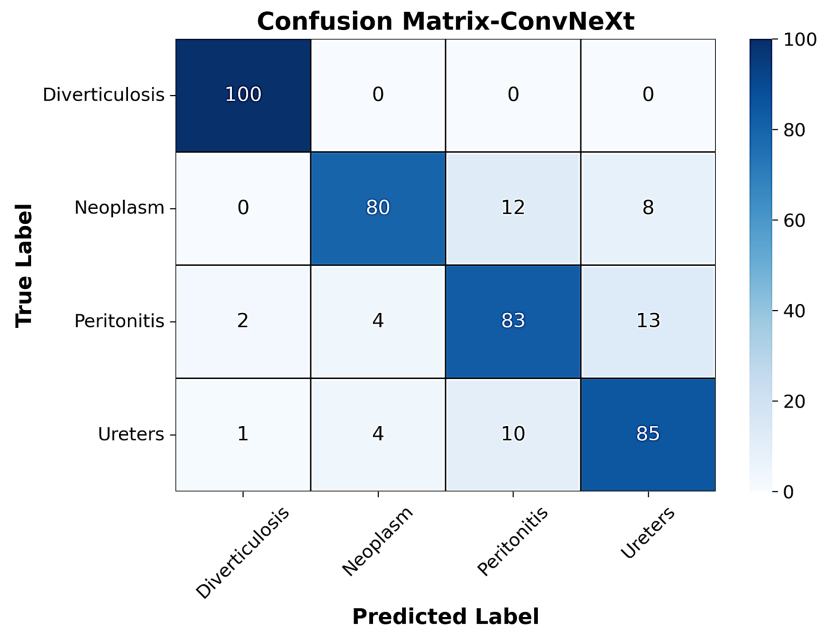


Figure 6: Confusion matrix of ConvNeXt model.

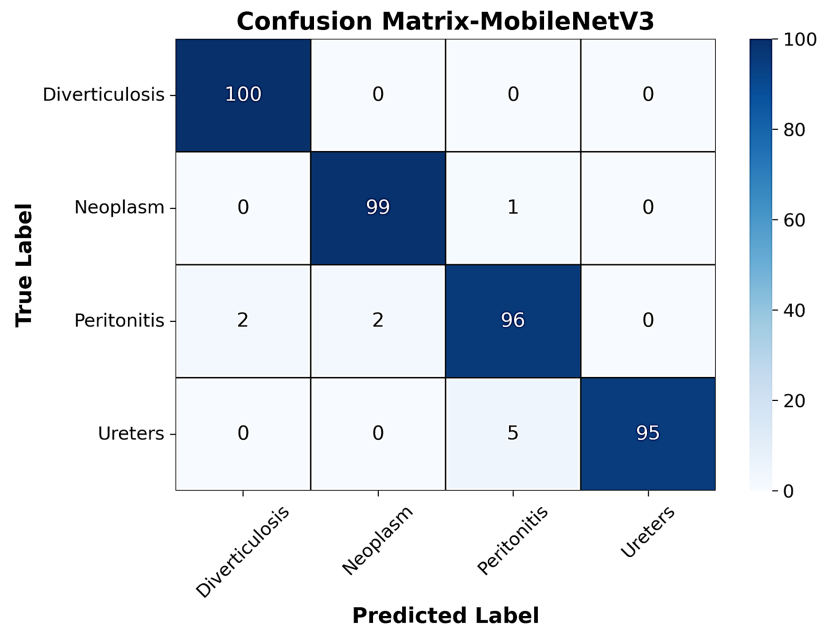


Figure 7: Confusion matrix of MobileNetV3 model.

ViT, on the other hand, achieved a test accuracy of 97.00%. [Fig. 8](#) shows the performance of ViT. Six images of the Peritonitis class were wrongly classified as Neoplasm and Ureters, whereas four images from the Ureters class were wrongly classified as Neoplasm and Peritonitis, respectively, by this model.

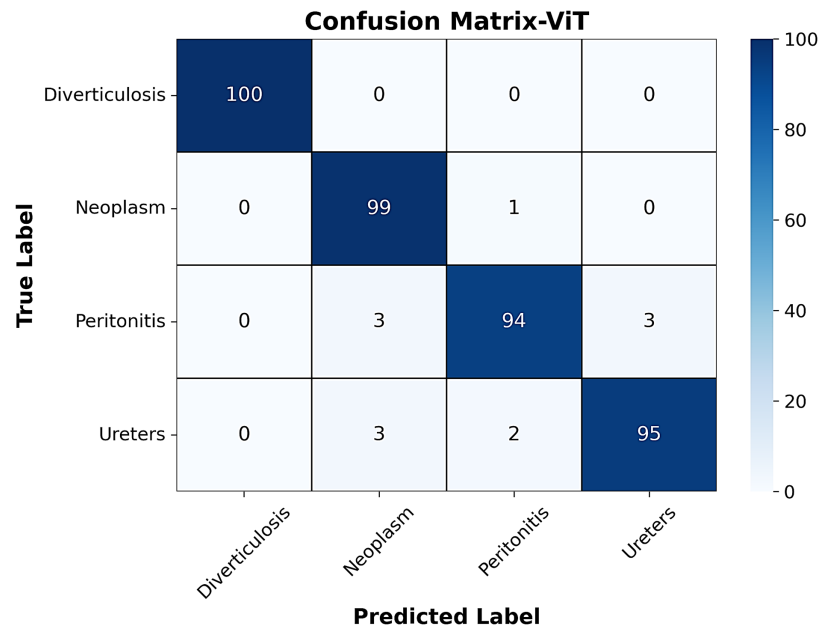


Figure 8: Confusion matrix of ViT model.

The proposed GastroNetV4 model achieved a final test accuracy of 98.25% across all four classes. The test sample on which the evaluation is conducted was not used during training or validation. Table 2 presents the detailed class-wise performance of the model in terms of all metrics.

Table 2: Classification report of the proposed GastroNetV4 model.

	Precision	Recall	F1 Score	Support
Diverticulosis	1.000	1.0000	1.0000	100
Neoplasm	0.9706	0.9900	0.9802	100
Peritonitis	0.9899	0.9800	0.9849	100
Ureters	0.9697	0.9600	0.9648	100
Accuracy			0.9825	400

The most interesting behavior was shown in the case of Diverticulosis, Neoplasm, and Peritonitis classes, where the model achieved the highest precision of 100%, 97.06%, and 98.99%, respectively. These prominent results confirm that the fewest misclassifications were made for these classes in the complexity matrix. This demonstrates that the model has perfectly learned the distinguishing features of these three diseases. The F1 Score of 100% was obtained for Diverticulosis. The class where the proposed GastroNetV4 model performed least well among others was the Ureter class, yet it gave a better result. The Ureter class had an F1 score of 96.97%, while the Peritonitis class achieved an F1 score of 98.49%.

After analysis of the test dataset, the confusion matrix obtained is shown in Fig. 9. The diagonals represent the correctly classified samples, and non-diagonals represent the misclassified samples. The result shows that the majority of the samples were correctly predicted by the proposed model.

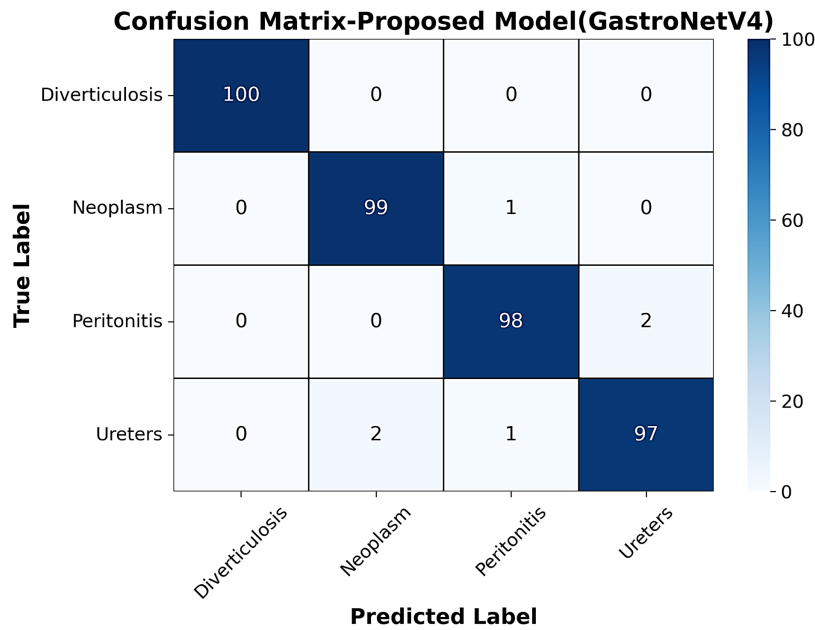


Figure 9: Confusion matrix of proposed GastroNetV4 model.

Fig. 10 below shows the ROC Curve for four classes. All classes had high Area Under the Curve (AUC) values under the ROC Curve, indicating effective classification performance. The AUC was 1.000 for Diverticulosis, 0.998 for Neoplasm, 1.000 for Peritonitis, and 0.997 for Ureters for the proposed GastroNetV4 model. Macro-average AUC and micro-average AUC were determined to be 0.999, indicating that the four classes were well separated.

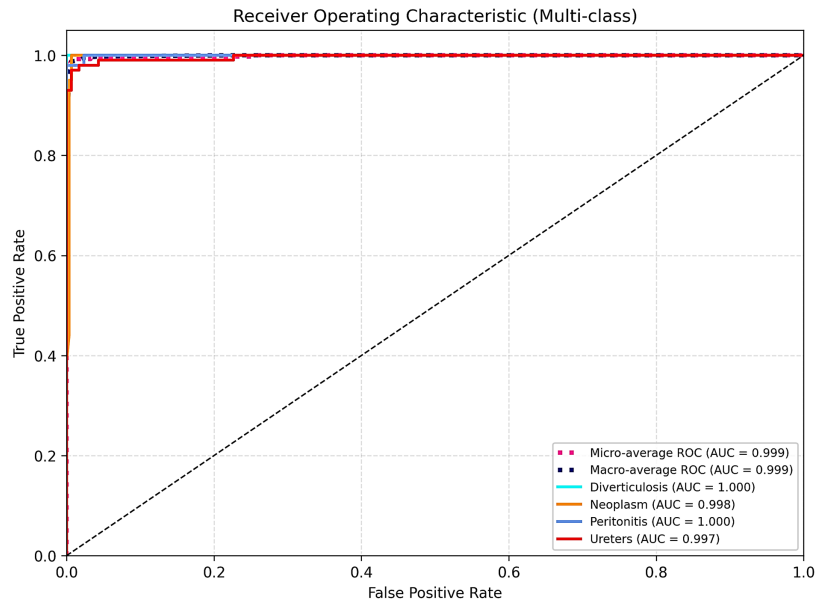


Figure 10: Multi-class ROC curve of the proposed GastroNetV4 model.

Table 3 presents the performance comparison of the proposed model with the other four models. It can be seen that the proposed model outperformed the other models in terms of metrics such as precision, recall, accuracy, and F1 score.

Table 3: Comparative performances of different models.

Models	Accuracy	Precision	Recall	F1 Score
InceptionV3	98.00	98.02	98.00	98.00
ConvNeXt	87.00	87.29	87.00	87.00
MobileNetV3	97.50	97.54	97.50	97.50
ViT	97.00	97.03	97.00	97.00
Proposed Model	98.25	98.25	98.25	98.25

The results show that the proposed GastroNetV4 model performs better than all the other architectures. The model produced the best results overall, with 98.25% Accuracy, 98.25% Precision, 98.25% Recall, and 98.25% F1-score, respectively.

These findings demonstrate that the proposed model provides the most effective method for reducing false positives (improved precision) and false negatives, while also producing generally accurate predictions. With an F1-score of 98.00%, InceptionV3 performed the closest among the models. ConvNeXt performed the worst, achieving an F1 score of 87.03%. Fig. 11 shows the comparative performance of the proposed GastroNetV4 model with other models across different metrics.

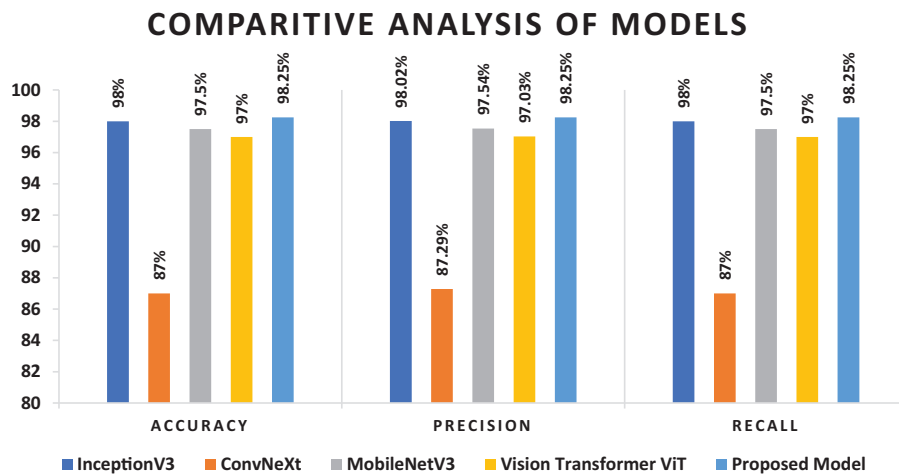


Figure 11: Comparative performance of different models across metrics.

Some models achieved high performance in selective classes, yet they failed to provide promising results overall. Hence, the proposed model has outperformed all models and can be considered a reliable and promising architecture for gastrointestinal disease detection.

Table 4 compares the number of parameters across all models. GastroNetV4 has a relatively small additional number of parameters compared to the base EfficientNet-B4 provided by the additional convolutional refinement head, and is comparatively less expensive than the large transformer and ConvNeXt models. The number of parameters quantifies the performance of a model, given its size, representing a trade-off between accuracy and computing cost.

Table 4: Parameter scale of all models.

Models	Parameter (Millions)
InceptionV3	21.8
ConvNeXt	28.6
MobileNetV3	5.4
ViT	22.1
EfficientNet-B4	19.3
Proposed Model	21.2

Ablation study on the CNN Refinement Head

To assess the contribution of the convolutional refinement head, an ablation was performed by fine-tuning EfficientNet-B4 directly with the default global average pooling and linear classification layer, removing the additional convolutional head, and training both models using the same hyperparameter configurations. EfficientNet-B4 achieved an accuracy of 97.25% (ECE = 0.0249), whereas GastroNetV4's accuracy was 98.25% (ECE = 0.0171). Overall accuracy of GastroNetV4 is higher as compared to EfficientNet-B4, also hybrid configuration had a more balanced class-wise recall distribution for clinically difficult classes (Peritonitis and Ureters) and maintained stable feature adaptation, without sacrificing calibration performance.

Next, The study performed three additional ablation experiments to validate the success of the proposed CNN-head architecture. The experiment replaced the 2-layer CNN-head with a 1-layer and 3-layer convolutional head while other settings remained the same. Table 5 shows that the test accuracy of the 1-layer CNN-head-based model was 97.50%, while that of the model with the proposed CNN-head architecture was 98.25%. The best test accuracy for the 3-layer CNN head was found to be 97.75%. Since the proposed 2-layer CNN head can lead to the best feature extraction performance while keeping the model complexity lower than the 3-layer CNN head, it is sufficient for endoscopic image classification. As such, the final GastroNetV4 architecture retained the 2 layer CNN-head.

Table 5: Comparative performance of 1-layer, 2-layer and 3-layer CNN-head configurations.

CNN-Head Configuration	Accuracy
1-Layer CNN-head	97.50%
2-Layer CNN-head	98.25%
3-Layer CNN-head	97.75%

t-SNE Visualization

To show that the model is not simply memorizing the training set, but learns features of biological importance, the study projected the high-dimensional latent space to the two-dimensional space using t-distributed stochastic neighbor embedding (t-SNE), as shown in Fig. 12. A clustering behaviour was observed in the projection of the images with the same pathology.

It can be seen that the Diverticulosis and Neoplasm classes have large decision margins, showing that they generate very compact clusters, which is in agreement with the high sensitivity scores of these classes in quantitative tests. The somewhat large overlap of the Peritonitis and Ureters classes can be explained by the similarity of their tissue structures. This compact intra-class variance and large inter-class

distance indicate that the proposed GastroNetV4 backbone extracts strong and generalizable gastrointestinal feature representations.

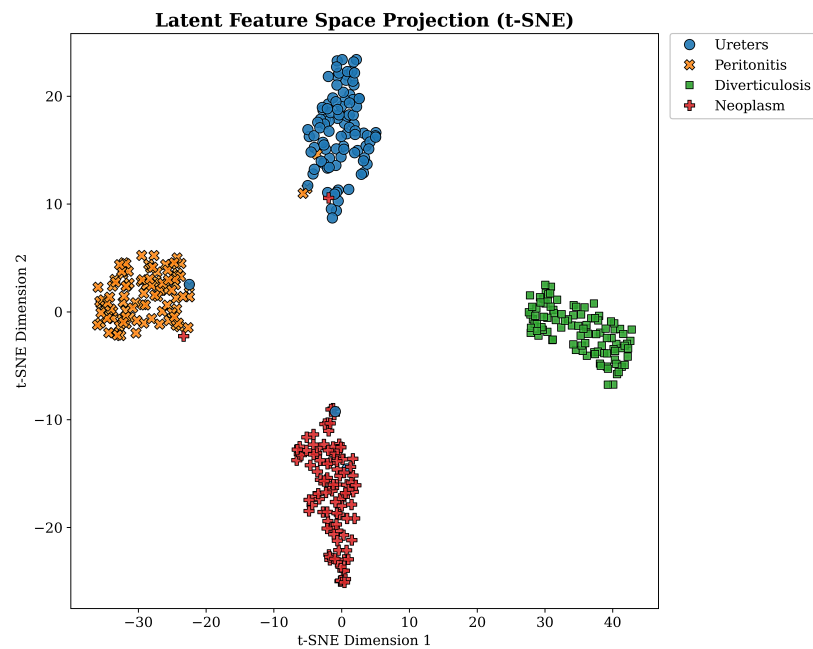


Figure 12: Distributed stochastic neighbor embedding (t-SNE) visualization of latent feature space on the test set.

Uncertainty (Entropy) Analysis

In clinical practice, a model should not only be accurate but also safe because a ‘confidently wrong’ model is not safe. Therefore, the study quantified the uncertainty in the model’s prediction on the test set using predictive entropy to assess the safety of the model, as shown in [Fig. 13](#).

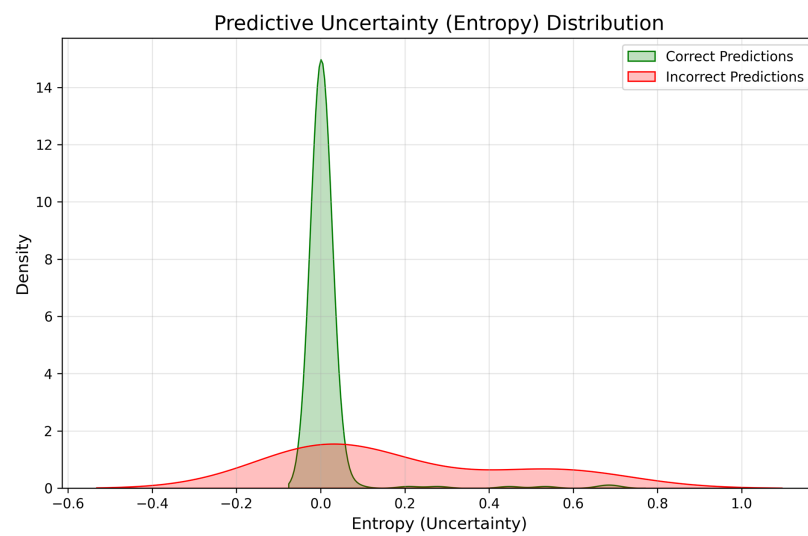


Figure 13: Density estimation of entropy for correct (Green) vs. incorrect (Red) classifications.

Statistical analysis indicates an important difference between correct and incorrect predictions. Correct diagnoses (Green curve) cluster around zero entropy value, indicating high confidence. In contrast, misclassified samples (Red curve) have a much flatter distribution and a much higher entropy. Although these findings show the model's ability to separate certain and uncertain predictions, simply comparing entropy is not sufficient to determine whether the model's probabilities are well calibrated.

The use of an entropy threshold will allow uncertain cases to be flagged for review by the gastroenterologists. While an operational entropy threshold was not used in this analysis, the separation of entropy distributions for correct and incorrect predictions suggests that it may be clinically useful. In future clinical implementations, a threshold could be selected from validation data to ensure an appropriate balance between diagnostic sensitivity and the need for manual review. Such images can be flagged for review by a gastroenterologist in a human-in-the-loop decision support system if predictive entropy is above some threshold, helping to reduce confidently incorrect automated predictions.

Explainable AI Analysis (Grad-CAM)

To better understand the decision-making of the proposed model and avoid predictions from a “black box,” the study conducted an Explainable AI (XAI) analysis. Grad-CAM was applied to the last convolutional layer of the CNN head of GastroNetV4, a hybrid EfficientNet-B4 + CNN model. Given an input endoscopic image, Grad-CAM computes the gradient of the predicted class score with respect to the feature maps of the last convolutional layer of a CNN. The gradients are averaged spatially to get the class-specific feature map importance weights. A heatmap is obtained by taking a weighted sum of the feature maps and applying ReLU. The resulting heatmap can be up-sampled to the size of the original image and overlaid on the RGB image to see which structure contributed more to the predicted class.

To visualize whether the network was focusing on the right areas to classify the cases, Grad-CAM was computed on correctly predicted high-confidence cases from all four classes and representative incorrect test images. Grad-CAM visualizations were generated for correctly classified, high-confidence test images from all four classes, and for a random sample of misclassified test images. These were inspected visually to verify that the model was focusing on the correct areas of the lesions, and to gain perception into the model's failure modes.

Fig. 14 shows the Grad-CAM overlays for true positive, high-probability images of the four GI diseases. For Diverticulosis, the strongest model activations are around the diverticular outpouchings along the colonic wall. In Neoplasm, the mass lesion is highlighted by a heatmap indicating the mass lesion and the irregular mucosal pattern. Likewise, in Peritonitis, the inflamed or exudative area is highlighted. In Ureters cases, the activations usually occur on or near the ureteric orifice.

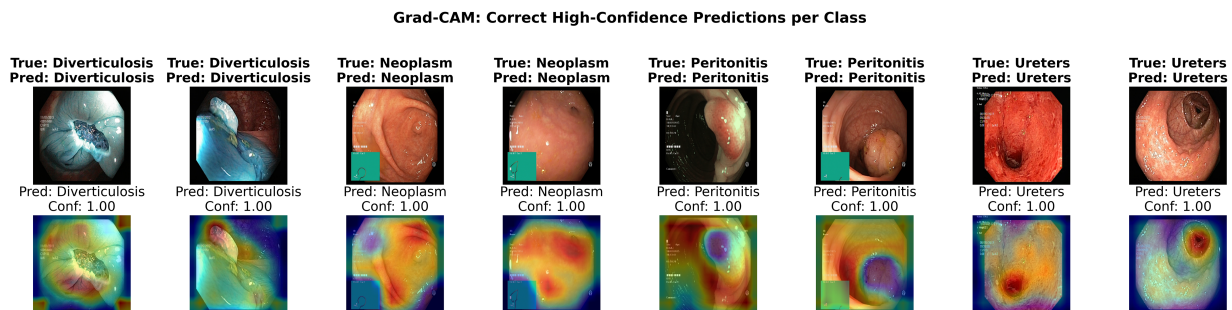


Figure 14: Grad-CAM overlays for correctly classified samples.

Fig. 15 shows examples where the focus of the model was on visually ambiguous regions, e.g., minor inflammatory changes next to a small neoplastic area; inflamed diverticula with fluid and folds. These are examples where humans also find it difficult to diagnose. The mistakes are not made because the model has ignored the lesion completely, but rather because the image is intrinsically difficult.

Grad-CAM: Misclassified Examples

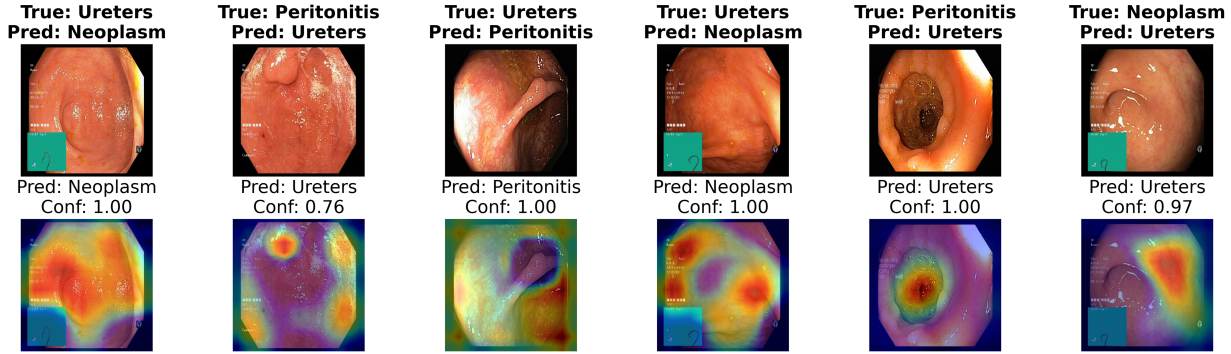


Figure 15: Grad-CAM overlays for misclassified test samples.

Grad-CAM analysis in this paper is strictly qualitative. The publicly available dataset does not provide pixel-level lesion annotations, segmentation masks, or expert-defined bounding boxes, which would have allowed for quantitative assessment of localization using Intersection-over-Union (IoU) and pointing game metrics. Thus, Grad-CAM results should be taken as further evidence that the model attends to, and can prioritize, clinically meaningful areas rather than accurate boundaries of the lesions. Visual inspection shows that the model performs salient activations over pathological structures for positive classes, consistent with the model encoding meaningful morphological features, rather than incidental patterns from the background. Future work will use clinically annotated datasets to quantitatively evaluate explainability, using lesion-overlap measures and expert agreement analysis to consolidate the assessment of interpretability.

Calibration Analysis

In addition to the classification performance, the calibration of predicted probabilities was evaluated on the proposed model in terms of the Expected Calibration Error (ECE). In clinical decision-support workflows, calibration is important because the confidence in a prediction could affect whether additional imaging, biopsy, and follow-up is performed. The ECE was computed from the softmax probabilities on the test partition. Let $\hat{p}_i$ denote the predicted confidence (maximum probability) for sample i and let $\hat{y}_i$ and y_i correspond to the predicted and true labels. These predictions were classified based on their confidence level into $M = 10$ equidistant bins within the interval $[0, 1]$. In the context of a confidence level bin B_m , the accuracy and average confidence levels were computed as follows:

$$accuracy(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i) \quad (8)$$

$$confidence(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \quad (9)$$

Expected Calibration Error (ECE) is defined as:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{N} |accuracy(B_m) - confidence(B_m)| \quad (10)$$

where N denotes the number of test samples.

The ECE measure was computed using a customized NumPy function for implementing the traditional confidence binning technique. With 10 equidistant bins, the obtained ECE value of GastroNetV4 is 0.0171 (1.71%). This demonstrates that the confidence levels of the predictions made by the algorithm show high correspondence with the empirical accuracies. The results are also shown in Fig. 16 through the reliability plot, where the calibration curve is very close to the ideal line.

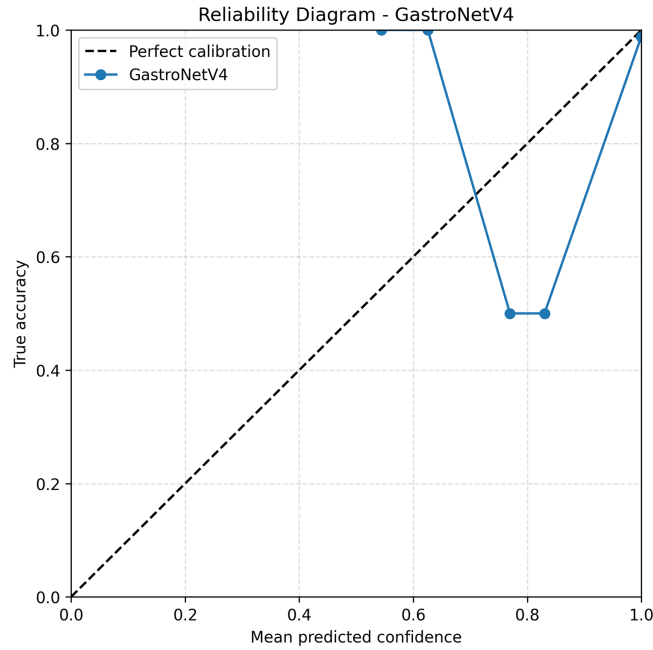


Figure 16: Reliability diagram of the proposed GastroNetV4 model on test samples.

In conclusion, the proposed model presents a promising performance for automated multi-class endoscopic disease classification for four classes: three gastrointestinal and one ureter-related disease. The performance is competitive with the state-of-the-art (SOTA) architectures/competing models evaluated. The proposed model has the potential to be applied to automated endoscopic image analysis. Further validation in larger, more heterogeneous and clinically representative datasets and in potential clinical studies will be required prior to integrating this model into clinical decision support systems.

5 Discussion

The study proposed GastroNetV4, a deep learning-based model for detecting gastrointestinal diseases in endoscopic images. The hybrid architecture with an EfficientNet-B4 backbone and a lightweight CNN head yielded an overall test accuracy of 98.25% among the four classes of interest (Diverticulosis, Neoplasm, Peritonitis, and Ureters). Class-wise accuracy can be observed on the diagonal of the confusion matrices. The multi-class ROC analysis and multi-class AUC analysis show high discriminative power across a range of

decision thresholds. Together, the results of the analysis indicate the proposed model's ability to learn highly discriminative features for multi-class GI disease detection.

The results reported are consistent with findings in other medical imaging studies that have demonstrated the power of transfer learning and pretrained feature extractors. By tuning the upper layers of the EfficientNet-B4 backbone and adding a task-specific CNN classifier layer, GastroNetV4 successfully adapts generic visual features to the specific domain of endoscopy. Compared with strong pretrained architectures (ConvNeXt, MobileNetV3, InceptionV3, and a pre-trained ViT) for the same preprocessing and data augmentation, the hybrid architecture has better accuracy, precision, recall, and F1 scores. Part of the success can be attributed to using a balanced dataset, meaning that the same number of samples is taken for each disease, preventing the model from being biased in favour of one over the other. Augmentation operations that are well defined and medically safe (rotation, flips, zoom, translations, colour jittering, etc.) were also applied to the data, which helped the models learn better generalizations of realistic variations in viewpoint and illumination.

Another important aspect of this work is the explicit evaluation of calibration and explainability, which are not typically considered in GI CAD. Reliability analysis shows that model confidence scores of GastroNetV4 are well-calibrated, having a low ECE. Our model's reliability curve is also close to the ideal diagonal. This is important if the predicted probabilities are to be used to decide the diagnosis or to trigger a follow-up intervention. The visualizations show that the network fixates on clinically important abnormalities across all classes, while the misclassified cases are associated with intrinsically ambiguous or poorly imaged cases.

Despite these positive findings, the experiments were all conducted with the same curated dataset. A limitation of this study is that a single-source publicly available dataset was used. Although the dataset was class balanced and contains clinically relevant endoscopic images, it may not reflect the variability in hospitals, endoscopy equipment, imaging protocol, patient population, and image acquisition conditions present in routine clinical practice. External validation using independently curated datasets and data from multiple testing centers and sites will need to be performed to assess the generalizability of GastroNetV4 against possible domain shifts in the clinical space and thus to safely introduce the model into clinical decision-support systems. But well defined and medically valid (rotation, flips, zoom, translations, colour jittering, etc.) augmentation of the images can give the models more varied training data so that they can more easily generalize to cases with variability in viewpoint or illumination. Overall, these findings suggest that CNN-based approaches, particularly hybrid CNN architectures, offer promising decision support systems for automated GI disease detection.

6 Conclusion

In this research, a hybrid architecture is proposed. The goal of GastroNetV4 was to transfer generic visual features from pretrained image classification models to endoscopy images by fine-tuning the top layers of an EfficientNet-B4 backbone, using an image-based CNN classifier. The hybrid model outperforms the strong pretrained models (ConvNeXt-Tiny, MobileNetV3, InceptionV3, and a pre-trained ViT) in accuracy, precision, recall, and F1 scores using the same preprocessing and data augmentation methods. In addition to this, the datasets were balanced, in that each class contained the same number of observations, to reduce the influence of class imbalance on the model. Image augmentation was also applied using safe and medically interpretable image transformations (including rotation, flipping, zoom, translation, and color jittering) to facilitate learning by allowing these models to generalize learned representations to variance in viewpoint, illumination, and other attributes in the real world.

Although transfer learning of EfficientNet backbones in the medical domain was already well explored, the main scientific contribution of GastroNetV4 is its domain-specific hybrid refinement strategy. Instead of using a linear classifier head for fine-tuning the pretrained backbone, GastroNetV4 employs a lightweight convolutional refinement head to equip the high-level ImageNet features to the more fine-grained structures in gastrointestinal endoscopy. This allows local pathological patterns, texture abnormalities, and subtle inflammatory features to be more effectively highlighted before being consolidated and classified in a global manner, and still maintains a reasonable computational burden while focusing features for the four clinically relevant GI classes.

However, because all of the experiments were run on the same curated dataset, it is possible that the images captured in this dataset do not capture the full range of a disease's clinical manifestations seen in a clinical environment with different centers, endoscope vendors, and patient populations, making it unclear if results from the datasets studied generalize to other environments. Validation of GastroNetV4 using multi-source datasets is an important area for future work. Overall, these results indicate the promise of CNN-based methods, in particular hybrid CNN architectures, as potential decision support systems for automated GI disease detection.

Acknowledgement: This work was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R765), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding Statement: This work was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R765), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: Conceptualization was done by Romana Aziz, Muhammad Ramzan and Areeba Gul; methodology, Muhammad Ramzan and Areeba Gul; software, Mahwish Ilyas; validation, Ala Saleh Alluhaidan and Mahwish Ilyas; formal analysis, Qaiser Abbas; investigation, Summair Raza; data curation, Areeba Gul; writing 1st draft preparation, Areeba Gul; writing—review and editing, Romana Aziz and Ala Saleh Alluhaidan; visualization, Areeba Gul; supervision, Summair Raza; funding acquisition, Romana Aziz. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: In this research, a publicly available dataset has been used, named Medical imaging [34].

Ethics Approval: The study was based on an open-access, totally anonymized endoscopic image dataset. The authors did not engage in the direct recruitment of human subjects or collect identifiable patient data. Thus, no ethical approval or informed consent was required in accordance with institutional guidelines.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang R, Li Z, Liu S, Zhang D. Global, regional, and national burden of 10 digestive diseases in 204 countries and territories from 1990 to 2019. *Front Public Health*. 2023;11:1061453. doi:10.3389/fpubh.2023.1061453.
2. Ananthakrishnan AN, Xavier RJ. Gastrointestinal diseases. In: Ryan ET, Hill DR, Solomon T, Aronson NE, Endy TP, editors. *Hunter's tropical medicine and emerging infectious diseases*. Amsterdam, The Netherlands: Elsevier; 2020. p. 16–26. doi:10.1016/b978-0-323-55512-8.00003-x.
3. Zhu Y, Wang QC, Xu MD, Zhang Z, Cheng J, Zhong YS, et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. *Gastrointest Endosc*. 2019;89(4):806–15.e1. doi:10.1016/j.gie.2018.11.011.
4. Li H, Hou X, Lin R, Fan M, Pang S, Jiang L, et al. Advanced endoscopic methods in gastrointestinal diseases: a systematic review. *Quant Imaging Med Surg*. 2019;9(5):905–20. doi:10.21037/qims.2019.05.16.

5. Wang J, Wang S, Zhang Y. Deep learning on medical image analysis. *CAAI Trans Intell Technol.* 2025;10(1):1–35. doi:10.1049/cit2.12356.
6. Liu D, Peng X, Liu X, Li Y, Bao Y, Xu J, et al. A real-time system using deep learning to detect and track ureteral orifices during urinary endoscopy. *Comput Biol Med.* 2021;128(4):104104. doi:10.1016/j.compbiomed.2020.104104.
7. Wang Y, Huang Y, Chase RC, Li T, Ramai D, Li S, et al. Global burden of digestive diseases: a systematic analysis of the global burden of diseases study, 1990 to 2019. *Gastroenterology.* 2023;165(3):773–83. doi:10.1053/j.gastro.2023.05.050.
8. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209–49. doi:10.3322/caac.21660.
9. Heydari K, Rahnavard M, Ghahramani S, Hoseini A, Alizadeh-Navaei R, Rafati S, et al. Global prevalence and incidence of inflammatory bowel disease: a systematic review and meta-analysis of population-based studies. *Gastroenterol Hepatol Bed Bench.* 2025;18(2):132–46. doi:10.22037/ghfbb.v18i2.3105.
10. Hoerter N, Gross SA, Liang PS. Artificial intelligence and polyp detection. *Curr Treat Options Gastroenterol.* 2020;18(1):120–36. doi:10.1007/s11938-020-00274-2.
11. Misawa M, Kudo SE, Mori Y, Cho T, Kataoka S, Yamauchi A, et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. *Gastroenterology.* 2018;154(8):2027–9.e3. doi:10.1053/j.gastro.2018.04.003.
12. Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, et al. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology.* 2018;155(4):1069–78.e8. doi:10.1053/j.gastro.2018.06.037.
13. Jayababu M, Eswar B, Afrid D, Sekhar AC, Prasad KK. Classification of gastrointestinal disease images through residual learning algorithm and comparative analysis of deep learning architectures. In: *Proceedings of the 2025 8th International Conference on Trends in Electronics and Informatics (ICOEI)*; 2025 Apr 24–26; Tirunelveli, India. p. 1388–94. doi:10.1109/ICOEI65986.2025.11013720.
14. Bektaş M, Tan C, Burchell GL, Daams F, van der Peet DL. Artificial intelligence-powered clinical decision making within gastrointestinal surgery: a systematic review. *Eur J Surg Oncol.* 2025;51(1):108385. doi:10.1016/j.ejso.2024.108385.
15. Barua I, Vinsard DG, Jodal HC, Løberg M, Kalager M, Holme Ø, et al. Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis. *Endoscopy.* 2021;53(3):277–84. doi:10.1055/a-1201-7165.
16. Hosny M, Elgendy IA, Ahmad Albashrawi M. Beyond transfer learning: attention-enhanced deep learning framework for multiclass gastrointestinal disease classification. *Expert Syst Appl.* 2026;295(23):128852. doi:10.1016/j.eswa.2025.128852.
17. Gao Y, Wen P, Liu Y, Sun Y, Qian H, Zhang X, et al. Application of artificial intelligence in the diagnosis of malignant digestive tract tumors: focusing on opportunities and challenges in endoscopy and pathology. *J Transl Med.* 2025;23(1):412. doi:10.1186/s12967-025-06428-z.
18. Shoukat N, Jeon JH, Mun C, Lee JY, Lee SH, Yang JY, et al. Machine learning-assisted electrochemical SERS for sensitive detection of multiple urinary proteins. *Sens Actuators B Chem.* 2026;454:139606. doi:10.1016/j.snb.2026.139606.
19. Matov A. Urinary biomarkers for lung cancer detection. *J Liq Biopsy.* 2026;12:100456. doi:10.1016/j.jlb.2026.100456.
20. Pessoa ACP, Quintanilha DBP, de Almeida JDS, Junior GB, de Paiva AC, Cunha A. Evaluating EfficientNet architectures for pathology detection in endoscopic gastrointestinal tract images. *SN Comput Sci.* 2025;6(5):431. doi:10.1007/s42979-025-03986-3.
21. Razali MN, Arbaiy N, Lin PC, Ismail S. Optimizing multiclass classification using convolutional neural networks with class weights and early stopping for imbalanced datasets. *Electronics.* 2025;14(4):705. doi:10.3390/electronics14040705.
22. Utku A. Enhanced gastrointestinal disease classification using a ConvViT hybrid model on endoscopic images. *Phys Eng Sci Med.* 2025;48(4):1539–54. doi:10.1007/s13246-025-01600-7.

23. Houmaidi W, Hadadi M, Sabiri Y, Chtouki Y. DeepGI: explainable deep learning for gastrointestinal image classification. arXiv:2511.21959. 2025. doi:10.48550/arxiv.2511.21959.
24. Attallah O, Sharkas M. GASTRO-CADx: a three stages framework for diagnosing gastrointestinal diseases. PeerJ Comput Sci. 2021;7(12):e423. doi:10.7717/peerj-cs.423.
25. Jiang Q, Yu Y, Ren Y, Li S, He X. A review of deep learning methods for gastrointestinal diseases classification applied in computer-aided diagnosis system. Med Biol Eng Comput. 2025;63(2):293–320. doi:10.1007/s11517-024-03203-y.
26. Naseem S, Jahangir R, Alturki N, Shehzad F, Ullah MS. DeepNeck: bottleneck assisted customized deep convolutional neural networks for diagnosing gastrointestinal tract disease. Comput Model Eng Sci. 2025;145(2):2481–501. doi:10.32604/cmes.2025.072575.
27. Fatima S, Dahan F, Shah JH, Almohamedh R, Aloqaily M, Riaz S. A multimodal learning framework to reduce misclassification in GI tract disease diagnosis. Comput Model Eng Sci. 2025;145(1):971–94. doi:10.32604/cmes.2025.070272.
28. Spada C, McNamara D, Despott EJ, Adler S, Cash BD, Fernández-Urién I, et al. Performance measures for small-bowel endoscopy: a European society of gastrointestinal endoscopy (ESGE) quality improvement initiative. Endoscopy. 2019;51(6):574–98. doi:10.1055/a-0889-9586.
29. Su C, Lin S, Su Q, Liu X, Wei L, Yang C. GI-ScreenNet v2: a modular framework for gastrointestinal disease detection based on an integrated transfer learning. Int J Med Robot Comput Assist Surg. 2026;22(1):e70128. doi:10.1002/rcs.70128.
30. Chou CK, Lee KH, Karmakar R, Mukundan A, Chen TH, Kumar A, et al. Integrating AI with advanced hyperspectral imaging for enhanced classification of selected gastrointestinal diseases. Bioengineering. 2025;12(8):852. doi:10.3390/bioengineering12080852.
31. Tsai TJ, Lee KH, Chou CK, Karmakar R, Mukundan A, Chen TH, et al. Enhancing early GI disease detection with spectral visualization and deep learning. Bioengineering. 2025;12(8):828. doi:10.3390/bioengineering12080828.
32. Ju J, Liu M, Song W, Zhang T, Liu J, Xu P, et al. A boundary-enhanced and target-driven deformable convolutional network for abdominal multi-organ segmentation. Pattern Recognit. 2026;172(2):112386. doi:10.1016/j.patcog.2025.112386.
33. Ju J, Zhang T, Song W, Xiao Z, Tu H, Guan Z, et al. Collaborative learning of dynamic and static information for medical image segmentation. Inf Fusion. 2026;125(1):103509. doi:10.1016/j.inffus.2025.103509.
34. Pandey S. Medical imaging [Internet]. Kaggle. [cited 2026 Aug 14]. Available from: <https://www.kaggle.com/datasets/heartzhacker/medical-imaging>.
35. Nawaz F, Ramzan M, Mehmood K, Ullah Khan H, Hayat Khan S, Raheel Bhutta M. Early detection of diabetic retinopathy using machine intelligence through deep transfer and representational learning. Comput Mater Contin. 2021;66(2):1631–45. doi:10.32604/cmc.2020.012887.