

ARTICLE

Bidirectional Motion-Temporal Deep Learning for Explainable Multi-Class Classification of Gastrointestinal Lesions in Wireless Capsule Endoscopy

Sarfaraz Natha^{1,*}, Mohammad Siraj^{2,*}, Mohammed Muflih Alamer³, Aaqid Syed⁴, Ayesha Shafique⁵ and Kashan Memon⁶

¹Department of Software Engineering, Sir Syed University of Engineering & Technology, Karachi, Pakistan

²Department of Electrical Engineering, College of Engineering, King Saud University, Riyadh, Saudi Arabia

³Department of Curriculum and Instruction, College of Education, King Saud University, Riyadh, Saudi Arabia

⁴Resident, Internal Medicine, Mobile Infirmary Medical Center, Mobile, AL, USA

⁵School of IoT Engineering, Wuxi Taihu University, Jiangsu Key (Construction) Laboratory of Intelligent IoT Technology and Applications in Universities, Wuxi, China

⁶Department of Electronic Engineering and Information Sciences, University of Science and Technology of China, Hefei, China

*Corresponding Authors: Sarfaraz Natha. Email: sarfaraz.a.s.natha@gmail.com; Mohammad Siraj. Email: siraj@ksu.edu.sa

Received: 12 June 2026; Accepted: 27 August 2026; Published: 28 September 2026

ABSTRACT: Gastrointestinal (GI) tract cancers are a serious health concern worldwide due to their high mortality rates. Wireless Capsule Endoscopy (WCE) provides a valuable non-invasive approach for detecting gastrointestinal abnormalities that may be associated with cancer. Despite WCE examinations generating many images, manual assessment is time-consuming. Therefore, automated methods capable of accurate and efficient lesion classification are highly desirable. Deep Learning (DL) techniques have demonstrated considerable potential for medical image analysis. However, many existing deep learning methods struggle to capture both broader contextual relationships and suitable patterns at the same time. While many are limited to binary classification. To address this limitation, this study proposed Bidirectional Motion Temporal (BiMT) for multi-class classification of gastrointestinal images by exploiting spatial and temporal information from image sequences. BiMT architecture effectively integrates spatial and temporal information to reduce the limitation of existing approaches in medical image analysis. In the first stage, pretrained CNN models, including InceptionNetV3 and DenseNet201, are employed to extract discriminative feature representations from the WCE dataset. This feature extraction process captures rich spatial characteristics from WCE images, providing a robust foundation for subsequent analysis. In the second stage, a BiLSTM model is used to capture the temporal dependencies between consecutive frames and is combined with Multi-Head Self-Attention (MHSA) to effectively model short-term contextual relationships. Furthermore, a custom Transformer encoder incorporating relative positional embeddings to model sequential information and increase gastrointestinal disease detection. During training, Categorical Focal Loss (CFL) is employed to emphasize difficult-to-classify samples and clinically relevant features. We combine the publicly available Kvasir-Capsule-v1 and Kvasir-v2 datasets to construct a unified dataset comprising four clinically relevant classes: ulcerative colitis, polyps, dyed-lifted polyps, and normal. The more sophisticated ConvNeXt-BiMT model achieved an average accuracy of 98.41%.

KEYWORDS: Convolutional neural networks; bidirectional motion temporal model; wireless capsule endoscopy; gastrointestinal disorders

1 Introduction

Gastrointestinal disorders (GDs) have a substantial impact on many people. The World Health Organization estimates that gastrointestinal disorders harm and ultimately kill over 1.9 million people each year [1]. WCE is a modern noninvasive diagnostic technique primarily used to visualize and evaluate the small intestine, which is difficult to examine using conventional endoscopy because of its anatomical location between the stomach and colon [2,3]. The patient swallows a small capsule that typically measures approximately 26 mm × 11 mm with a miniature camera in WCE. The capsule naturally passes through the gastrointestinal tract. It captures and transmits images of the small intestine to detect and assess abnormalities. The method generates many images. The manual review is time-consuming and theoretically prone to human error. Consequently, accurately identifying abnormalities in every frame can be challenging for physicians, strengthening the need for a computer-aided diagnosis (CAD) system specifically designed for gastrointestinal image analysis. Such a system can assist clinicians by automatically detecting and classifying normal frames with high accuracy [4,5]. However, variations in lesion appearance, image texture, scale, and visual characteristics can make automated analysis challenging. Therefore, robust automated technologies are needed to support accurate lesion categorization. In recent years, machine learning (ML) and deep learning (DL) techniques have demonstrated considerable potential for medical image analysis. These techniques have been applied to diagnostic tasks involving modalities such as X-ray, ultrasound, and magnetic resonance imaging (MRI) [6,7]. Machine learning (ML) techniques can analyze features extracted from images to differentiate between normal and abnormal conditions in the gastrointestinal tract. However, conventional ML methods often require extensive manual effort and expert knowledge to design handcrafted image features [8]. This feature engineering stage is not only labor-intensive but also inefficient, as it depends strongly on human expertise and careful selection of parameters [9].

A deep learning approach is well-suited to extracting the best spatial and temporal features from images with varying textures and sizes, which are more difficult to extract manually. Deep learning models such as Convolutional neural networks (CNNs) have made great strides in medical image processing in recent years because of their powerful capacity to automatically learn and extract pertinent characteristics from data [10,11]. These techniques are quite successful in spotting intricate patterns and aiding in the process of making diagnostic decisions [12]. CNN-based methods have trouble identifying items at various sizes within pictures, which might hinder their ability to accurately localize and diagnose diseases [13] despite their effectiveness. This integrated method aims to improve multi-classification of gastrointestinal images while also enhancing accurate localization of regions of interest (RoI) [14]. CNNs can identify features in images through convolutional layers that apply filters and kernels, and through activation layers that recognize patterns at different scales. CNNs are generally used to analyze WCE images. CNNs can efficiently capture spatial features, supporting tasks such as image recognition and generation. They face several challenges due to their dependence on broad datasets. Researchers use transfer learning (TL) for generalization, the capability of a technique to maintain reliable performance when applied to new and unseen datasets [15]. CNN models focus on learning the local features of individual samples and inputs. Another type of deep learning algorithm that can handle temporal features in data is the Recurrent Neural Network (RNN), which can use past information to make decisions at the current step, but this past information may become weaker or less clear the farther back it is. Therefore, a Long Short-Term Memory (LSTM) network [16] was introduced to address this issue. Additionally, the attention mechanisms can enhance memory for relevant information over long sequences [17]. They are effective on some computer vision tasks, such as object detection and image processing [18] and object classification [19]. Various attention mechanisms have been explored, such as Self-Attention (SA) [20], Bidirectional Long Short-Term Memory (BiLSTM), and Multi-Head Self-Attention (MHSA) [21]. The WCE dataset comprises images of Ulcerative colitis, Polyps, and

dyed-lifted polyps. This study proposes an approach that accurately detects and classifies gastrointestinal conditions with highly optimal features from WCE data by leveraging prominent patterns. The Bi-LSTM network processes sequential data to output more accurate recognition by extracting clinically significant features from a series of frames and adaptively adjusting attention weights, together with the convolutional operations of CNNs that extract spatial features from endoscopic images. The MHSA technique was also presented to detect associations among short-term frames.

We introduce the Bidirectional Motion Temporal (BiMT) model for gastrointestinal disease detection and classification using WCE images extracted from videos. The proposed BiMT architecture effectively integrates spatial and temporal information to reduce the limitations of existing approaches in medical image analysis. In the first stage, pretrained CNN models including InceptionNetV3 and DenseNet201 are employed to extract discriminative feature representations from the WCE dataset. This feature extraction process captures rich spatial characteristics from WCE images, providing a robust foundation for subsequent analysis. In the second stage, a BiLSTM model is used to capture the temporal dependencies between consecutive frames and is combined with Multi-Head Self-Attention (MHSA) to effectively model short-term contextual relationships. Furthermore, a custom Transformer encoder incorporating relative positional embeddings to model sequential information to increase gastrointestinal disease detection. During training, Categorical Focal Loss (CFL) is employed to emphasize difficult-to-classify samples and clinically relevant features. Thereby improving classification performance. We use multiple publicly available WCE datasets, which are merged to form a comprehensive dataset comprising four clinically relevant classes: ulcerative colitis, polyps, dyed-lifted polyps, and normal. Finally, we perform an extensive evaluation of the proposed model using different performance metrics. The more sophisticated ConvNeXt-BiMT model achieved an average accuracy of 98.41%. The experimental results demonstrate that the proposed framework achieves strong performance in accuracy in identifying and classifying gastrointestinal disease.

The paper is organized into five sections: in [Section 2](#), the literature review; the proposed methodology in [Section 3](#); the evaluation and experimental results in [Section 4](#); and, finally, the conclusion, limitations, and future directions in [Section 5](#).

2 Literature Review

There are many state-of-the-art methods proposed by various investigators covered in this section. The researchers state that three main feature-extraction techniques are commonly employed for the classification of WCE gastrointestinal images. In this research, Khan et al. [22] suggested a hybrid deep learning model that uses depth-wise concatenation to combine CNN-GRU and SC-DSAN for efficient classification of gastrointestinal diseases. Bayesian Optimization and EMPA are used to correct hyperparameters, which are used to improve model performance. Although it requires significant computational resources, validation on smaller datasets indicates further optimization such as quantization and trimming. Bajhaiya and Unni [23] proposed WCE images to identify gastrointestinal disorders like lymphangiectasis, bleeding, and ulcers by using a 13-layer CNN network with GINet. The model achieved good accuracy with sensitivity, but its performance can be improved. It needs to be tested on a larger dataset. Its ability to reduce bias and improve interpretability is one of the main benefits. However, it relies on a single dataset that cannot generalize to different datasets. Zhang et al. [24] suggested the AFR model as a dependable method for categorizing colonoscopy images of ulcerative colitis (UC), attaining high accuracy in all Mayo score categories. The model locates subtle lesion details and reduces noise and irrelevant features using attention mechanisms and feature improvement. The self-supervised learning and modular structure help the model generalize better, as evidenced by testing on independent datasets, although patient-specific characteristics, lighting conditions, and image quality variations pose challenges. To enhance its ability to detect small pathogenic

changes, more advanced feature extraction and better optimization systems will be required. In addition, due to the computing cost, the model may not be practical for real-time clinical applications. Medical image administering has seen significant progress with deep learning. This technology is already in use to automatically interpret images from colonoscopy. Recent studies show that these models are efficient at diagnosing and classifying gastrointestinal disorders such as Xie et al. [25] employed the EfficientNet-B5 architecture to detect and grade ulcerative lesions associated with small-bowel Crohn's disease, an inflammatory bowel disease affecting the small intestine. Their approach achieved an accuracy of greater than 96% in recognizing ulcers. These findings highlight the potential of deep learning-based methods to support more accurate and objective diagnosis. Murugesan et al. [26] used the YOLOv3 MSF deep learning framework for the analysis of colonoscopy pictures for the detection and classification of the stages of colon cancer. Their method combines deep learning-based analysis with measurements of polyp size features such as breadth and length. This provides a technology-based solution that can enhance the treatment planning and early detection of colon cancer. Karthikha et al. [27] proposed a deep learning model based on Dilated-U-Net-Seg to segment polyps in GI images, which produced good performance for accurate polyp identification according to different assessment metrics. The key advantage of the proposed UViT-Seg framework is its capability to simultaneously capture global context and fine local details with the combination of a vision transformer-based encoder and an enhanced decoder [28]. The dual attention devices and residual connections help retain more features, improving the accuracy of segmentation, and the model shows good generalization on multiple datasets with relatively low computational demands, making it suitable for applied clinical applications. The design is, however, somewhat complicated, which may make it harder to implement and understand. It is also sensitive to the classification of polyp and normal tissue because they share similar feature patterns. Limitations of the model include accuracy reduction in deeper polyp regions and ill-defined boundaries. It may also not perform as well on datasets with different characteristics. However, larger polyps are often not visible. While the model is largely effective, more work needs to be done to refine the model for consistency and reliability across different types of gastrointestinal imaging scenarios. In Zhang et al. [29] proposed a multimodal method called MSGC was proposed to tackle the task of grading the severity of gastric cancer using endoscopic images and diagnostic textual records by jointly learning from them via BERT to extract semantic information from text and an improved ResNeXt model to extract visual characteristics, and a contrastive learning method is utilized to map the text and image representations into a shared feature space to enhance intra-class similarity and cross-modal alignment. A multi-head attention module can be added to the model to dynamically weight feature components so that the model can focus on the most relevant diagnostic cues. The method has several advantages: it combines complementary information from visual and textual modalities, which results in more accurate diagnoses, but the model trained on a private dataset has limited generalizability to larger-scale clinical settings and a wider range of patient populations, and its efficacy is highly dependent on the quality and completeness of both image and text inputs. Prakash and Krishnamurthy [30] introduced a deep learning-based framework to analyze the WCE dataset to aid in the non-invasive diagnosis of small-intestinal disorders. It develops a hybrid model, TU-MNetv2, to enhance depth estimation from endoscopic frames by combining TransUNet and MobileNetV2. It employs a unique IF-AZOA optimization method that considers entropy, edge density, key points, and image moments. It uses preprocessing to reduce noise and emphasize important visual features, which improve image quality, and ensemble learning to improve lesion-detection performance in difficult scenarios. While these advantages are beneficial, the proposed framework has a high computational cost due to its deep learning and optimization components, and its performance is sensitive to preprocessing and the availability of well-annotated datasets. However, it may not generalize well to other capsule devices or clinical environments, and it still has difficulty with subtle or low-contrast lesions. In this study, Balasubramanian et al. [31] developed the HARU-Net framework, which

combines histogram equalization with an attention-based, residual-gated U-Net architecture, to improve classification of gastrointestinal (GI) disorders. The proposed method offers the advantage of preprocessing to enhance image quality, making feature extraction from video capsule endoscopy images easier. The residual gating used in Attention U-Net enhances feature representation and strengthens the learning process. The classification accuracy of 98.21% is higher than that of other traditional deep learning models, such as ResNet and the standard Attention U-Net. This indicates that it is well-suited to identify small patterns in medical images; however, the dataset is small and is derived from a single database, which could limit its effectiveness in other clinical settings, and the model is reliant on multiple preprocessing steps, which can be computationally complex and challenging for real-time applications. Furthermore, it is challenging to optimize the encoder–decoder architecture with various activation functions or loss strategies, and it needs to be validated on large-scale and multi-center datasets to verify its generalizability. The limitations of the most advanced techniques currently in use are listed in Table 1. By tackling several significant issues, this study seeks to increase the model's capacity to generalize while maintaining its robustness. Combining several datasets has lessened class imbalance, one significant problem. As a result, the data space is covered more widely, and the diversity of the data is increased. The model's objectivity, robustness, and dependability are improved because of the more evenly distributed classes. To support this technique, two publicly available WCE datasets were sequentially combined to create a larger collection of training and testing images. Although the initial datasets included eight gastrointestinal classifications, this study concentrates on four therapeutically significant categories: ulcerative colitis, polyps, dyed-lifted polyps, and normal.

Table 1: The methodology, strength, and limitations of prior studies.

Authors, Year, Reference	Methodology	Strength	Limitations
Malik et al., 2024 [32]	Gated Recurrent Unit (GRU) + ResNet152V2, two different types of gastric ulcer using WCE images.	This model achieved higher accuracy with a hybrid approach, reducing overfitting.	The detection of disease in the proposed model lacks promising outcomes.
Li et al., 2025 [33]	The attention approach enhances module precision and improves lesion-sensitive compactness.	The model has 96.6% precision using the Kvasir-Capsule-v1 dataset.	More data is required to train the model to improve the accuracy.
Singh et al., 2026 [34]	To propose a lightweight transformer-based framework	Achieves strong and consistent performance (up to 97% accuracy and 97% F1-score) with only 6.42M trainable parameters	Limited evidence of clinical validation and generalization testing.
Li et al., 2025 [35]	Multi-task learning (MTL) to classify bleeding, polyps, and ulcers	The model achieved 96.80% accuracy.	The method lacks extensive validation on large, diverse, real-world clinical WCE datasets, limiting confidence in its robustness

(Continued)

Table 1 (continued)

Authors, Year, Reference	Methodology	Strength	Limitations
Yakar et al., 2025 [36]	Hybrid deep learning approach	The model achieved 92.04% precision.	The frames are partially degraded due to poor illumination.
Attallah et al., 2025 [37]	Hybrid deep learning architecture.	The Kvasir v2 and Hyper Kvasir datasets had accuracy rates of 97% and 98%, respectively.	It suffers from high latency, and the tuning process is lacking for large datasets.
Chou et al., 2025 [38]	This proposed model integrates the Spectrum-Aided Vision Enhancer mechanism.	This model achieved an F1-score of 82% and a precision of 82%.	The tuning process is lacking for larger datasets.

The datasets used in previous studies are relatively small and often derived from a single database, which may limit the generalization of the developed models across diverse clinical environments and patient populations. In addition, some previous state-of-the-art approaches rely on proprietary datasets and are sensitive to the completeness and quality of the available data, potentially reducing their reliability when relevant information is missing. To these issues, we propose the Bidirectional Motion Temporal (BiMT) model for gastrointestinal disease detection and classification using WCE images extracted from videos. The proposed BiMT architecture effectively integrates spatial and temporal information to reduce the limitations of existing approaches in medical image analysis. In the first stage, pretrained CNN models including InceptionNetV3 and DenseNet201 are employed to extract discriminative feature representations from the WCE dataset. This feature extraction process captures rich spatial characteristics from WCE images, providing a robust foundation for subsequent analysis. In the second stage, a BiLSTM model is used to capture the temporal dependencies between consecutive frames and is combined with Multi-Head Self-Attention (MHSA) to effectively model short-term contextual relationships. Furthermore, we introduce a custom Transformer encoder incorporating relative positional embeddings to model sequential information to increase gastrointestinal disease detection. During training, Categorical Focal Loss (CFL) is employed to emphasize difficult-to-classify samples and clinically relevant features. Thereby improving classification performance. We use multiple publicly available WCE datasets, which are merged to form a comprehensive dataset comprising four clinically relevant classes: ulcerative colitis, polyps, dyed-lifted polyps, and normal.

3 Methodology

The suggested model, the Bidirectional Motion Temporal (BiMT) model, combines transformer-based classification, MHSA-BiLSTM, and CNNs and RNNs as illustrated in Fig. 1. WCE images are collected from videos and used as input to the proposed model. We use different publicly available datasets and combine them. These images are divided into train, test, and validation folders. WCE images are passed to a CNN with an RNN model to learn the best spatial and temporal features. The MHSA layer processes the BiLSTM output and enables the model to focus on the critical features needed for anomaly detection. To capture temporal relationships within the anomaly frames, an additional BiLSTM layer is employed to process the features. The data goes to the transformer encoder block, and positional embeddings are applied. The transformer encoder

block's output is passed through a series of max-pooling layers, and SoftMax layers are used for final GDs classification, such as ulcerative colitis, polyps, and dye-lifted polyps. GDs tract WCE images constitute the input dataset. Our proposed model integrates CNN, RNN, BiLSTM-MHSA, and transformer layers to extract the best temporal and spatial features that enable precise detection and classification of GI tract abnormalities from the WCE dataset. The proposed model architecture is presented in Fig. 2.

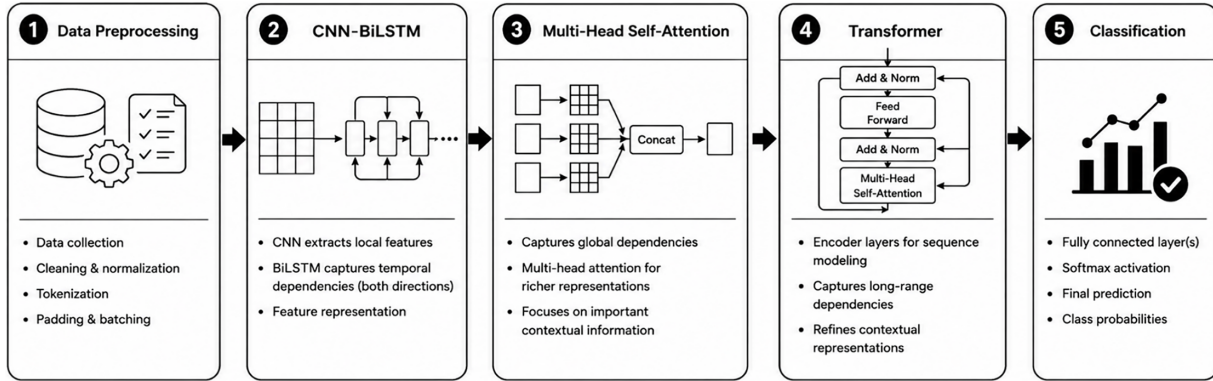


Figure 1: The flowchart of the proposed model.

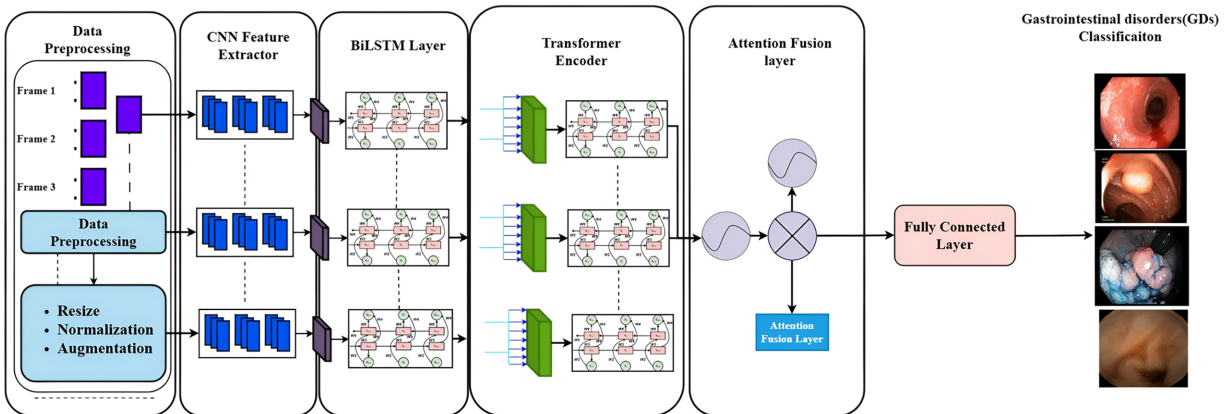


Figure 2: The proposed base model BiMT architecture.

3.1 Bidirectional Long Short-Term Memory (Bi-LSTM) Model

We employed a BiLSTM model [39,40], a type of Recurrent Neural Network (RNN), to capture temporal dependencies in sequential data. LSTM networks consist of input, forget, and output gates together with a cell state that functions as the network memory unit [41]. The input gate controls the extent to which newly generated information is incorporated into the cell state, whereas the forget gate determines which information from the previous cell state should be retained or discarded [42]. The output gate regulates the information passed from the cell state to the hidden state. LSTM can effectively preserve relevant information over long sequences and mitigate the vanishing gradient problem commonly encountered in conventional RNNs. Fig. 3 illustrates the base structure of the LSTM model. A BiLSTM network extends the LSTM architecture by processing the input sequence in both forward and backward directions using two separate LSTM layers. Each direction maintains its own hidden and cell states, allowing the network to capture information from both preceding and succeeding time steps. The outputs of the forward and backwards

LSTM layers are subsequently combined. Typically, through concatenation to generate both past and future contextual information. In Fig. 4 $x(t-1)$, x_t and $x(t+1)$ represent the input data at time $(t-1)$, (t) and $(t+1)$ respectively while $h(t-1)$, h_t and $h(t+1)$ denote the corresponding hidden states. Similarly, $o(t-1)$, o_t and $o(t+1)$ representing the corresponding outputs. The weights $w_1, w_2, \dots, w_6$ represent the layer-specific parameters associated with the LSTM computations. During the forward and backward passes, these parameters are used to update the respective hidden and cell states. The final Bi-LSTM output is obtained by combining the representation generated from both temporal directions. Thereby enabling the model to exploit information from the complete sequence.

Eqs. (1)–(3) describe it.

$$h_t = f_1(w_1x_t + w_2h_{t-1}) \quad (1)$$

$$h'_t = f_2(w_3x_t + w_4h'_{t+1}) \quad (2)$$

$$o_t = f_3(w_5h_t + w_6h'_t) \quad (3)$$

where f_1, f_2 , and f_3 are activation functions between different layers. To determine the input state f_t, v_t, i_t in Eqs. (4)–(6).

$$f_t = \sigma(w_f[h_{t-1}x_t] + b_f) \quad (4)$$

$$v_t = \tanh(w_c[h_{t-1}x_t] + b_c) \quad (5)$$

$$i_t = \sigma(w_j[h_{t-1}x_t] + b_j) \quad (6)$$

Calculate the cell state C_t in Eq. (7).

$$C_t = f_t.C_{t-1} + i_t.v_t \quad (7)$$

Compute the output gate o_t and the hidden state h_t at the current state as present in Eqs. (8) and (9).

$$o_t = \sigma(W_o.[h_{t-1}x_t] + b_o) \quad (8)$$

$$h_t = o_t \tanh(C_t) \quad (9)$$

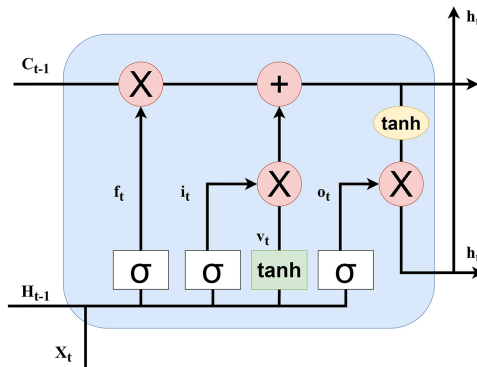


Figure 3: The LSTM model architecture.

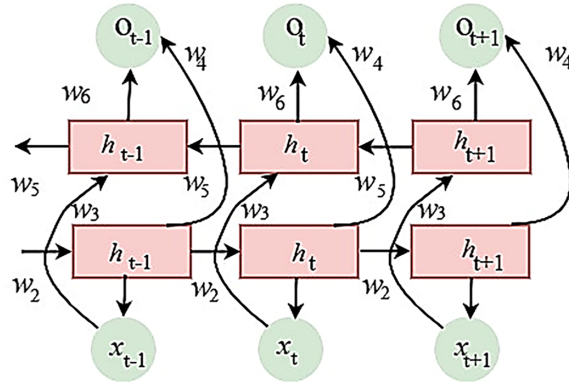


Figure 4: BiLSTM architecture.

Here W_o and b_o represent the weights and biases of the training matrix. The sigma σ denotes a non-linear activation function that outputs values within the range $[0, 1]$. The hidden layer unit is denoted by h_t and f_t present the forgotten gate unit. The cell state unit is shown by v_t . The input gate is denoted by i_t . The output gate is represented by o_t and C_t is represented as the cell state to synchronize and output information from the previous unit. BiLSTM techniques support sequence knowledge in both directions and integrate this substantial information into the current output layer. By leveraging relationships from both past and present data. BiLSTM improves its prediction capability. The complete BiLSTM hidden layer consists of cascading vectors that merge the outputs from the forward and reverse processes. The proposed spatiotemporal feature extraction module generates input for the attention module. This module displays stacking multiple Residual Attention BiLSTM blocks connected in sequence, with the output from each being passed to the initial BiLSTM of the following block. Let $c_t^{(i)}$ where i th the temporal features vector is made by the BiLSTM with t time. The attention layer produces a background vector v_t for $c_t^{(i)}$ at time t .

The multiple stacked Residual Attention BiLSTM blocks are connected in sequence, with the output from each block being passed to the initial BiLSTM of the following block. An activation function is applied to the hidden state h_t of the first BiLSTM layer to compute the relevance score $r_t^{(i)}$ in Eq. (10).

$$r_t^{(i)} = \tanh(WH_t + b) \quad (10)$$

Here $r_t^{(i)}$ is the relevance feature i with respect to time t . W and b represent the weight and bias. The activation function is hyperbolic tangent $\tanh(\cdot)$. The attention module computes the attention weight $A_t^{(i)}$ with respect to time t present in Eq. (11).

$$A_t^{(i)} = \frac{\exp(w_t^{(i)} r_t^{(i)})}{\sum_{j=1}^N \exp(w_t^{(i)} r_t^{(i)})} \quad (11)$$

Here $w_t^{(i)}$ shows the model weight learned feature i at time t and collects high-level sequence information in our method by using stacked RNN with multi-LSTM layers. The data in an LSTM is often processed through one layer before being sent to the output. On the other hand, it examines data at multiple levels while solving temporal sequences. In this configuration, the hidden state from the preceding layer is sent into each LSTM layer, forming a hierarchy of layers that jointly process the input as shown in Eq. (12).

$$y = f(x) + x \quad (12)$$

Here y is the output and x is the input of the BiLSTM block.

3.2 Multi-Head Self-Attention (MHSA) Method

A form of attention mechanism called MHSA is incorporated into the proposed model to enhance feature representation and enable the model to assign greater importance to regions that are more likely to contain disorders. In our approach, the input to the attention module is the output of the first BiLSTM layer denoted by Z [43]. The feature representation Z is reshaped into dimensions as shown in Eq. (13).

$$Z_{\text{reshape}} = \text{Reshape}(Z) \in \mathbb{R}^{(n_{\text{rows}} \times n_{\text{cols}}) \times n_{\text{channels}}} \quad (13)$$

The query, key, and value representations are obtained through three learnable projections as present in Eqs. (14)–(16).

$$Q = ZW_Q, K = ZW_K, V = ZW_V \quad (14)$$

$$Z \in \mathbb{R}^{(n_{\text{rows}} n_{\text{cols}}) \times n_{\text{channels}}} \quad (15)$$

$$W_Q \in \mathbb{R}^{n_{\text{channels}} \times d_k}, W_K \in \mathbb{R}^{n_{\text{channels}} \times d_k}, W_V \in \mathbb{R}^{n_{\text{channels}} \times d_v} \quad (16)$$

The scaled dot product attention is then calculated as shown in Eq. (17).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (17)$$

In the MHSA mechanism, the same input representation (z) is projected into (h) independent attention heads. Each head learns separate query, key, and value projections and captures different relationships within the feature representation. The i th attention head is defined as Eqs. (18) and (19).

$$\text{head}_i = \text{Attention}(ZW_i^Q, ZW_i^K, ZW_i^V), i = 1, \dots, h \quad (18)$$

$$W_i^Q \in \mathbb{R}^{n_{\text{channels}} \times d_k}, W_i^K \in \mathbb{R}^{n_{\text{channels}} \times d_k}, W_i^V \in \mathbb{R}^{n_{\text{channels}} \times d_v} \quad (19)$$

The outputs of all multi-heads are concatenated and linearly projected using the output weight matrix W^O as present in Eq. (20).

$$\text{MultiHead}(Z) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \quad (20)$$

where $W^O \in \mathbb{R}^{(hd_v) \times d_{\text{model}}}$ is the product of the output weight of MultiHead(Q, K, V). The MHSA receives the output from the first BiLSTM layer, which allows the model to focus on many objects simultaneously. The facility of models to focus on multiple items at once is resolved by the many attention heads in the MHSA layers. The BiLSTM layer processes each frame feature map, understanding the features of the data of these attention heads. This mechanism is improved by the second BiLSTM layer. MHSA layer is between the two BiLSTM layers in the proposed model. This arrangement is expected to organize situations in which many objects are concurrently involved in different irregular occurrences and enable the model to investigate the association between the most significant elements with later and earlier frames. This arrangement gives better performance and accuracy.

3.3 Transformer Encoder

The modified transformer encoder is designed to improve the processing and representation of image data by effectively capturing complex spatial relationships among different regions and features within an image. The conventional transformer architecture is enhanced through the integration of dense projecting layers, multi-head self-attention, and normalization layers in a carefully structured arrangement. This configuration enables the network to learn more meaningful and discriminative spatial feature representations, as illustrated in Fig. 5, thereby improving the overall performance of image classification. The first dense block consists of four fully connected layers with three dropout layers positioned between them. The dropout layers enhance the model's capacity for generalization and lessen the possibility of overfitting. The retrieved visual characteristics are further refined and transformed by the subsequent thick layers. The network can easily identify complicated patterns, connections, and discriminative attributes in the input images for the feature learning process, which produces a strong feature description for precise image classification and identification.

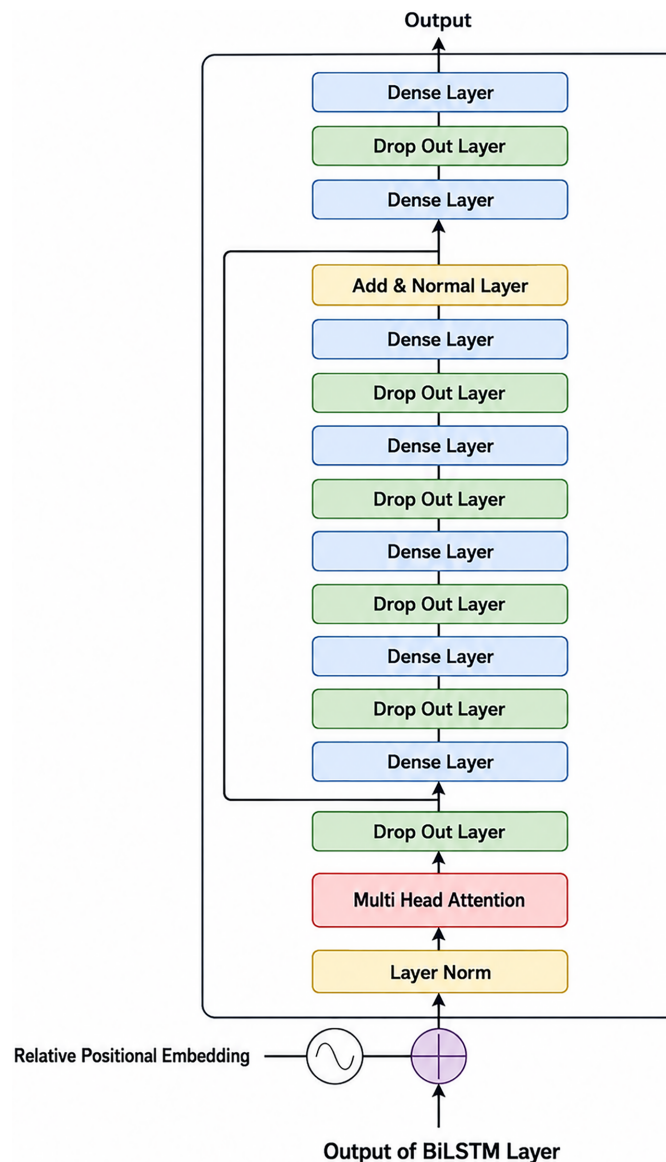


Figure 5: Transformer encoder architecture.

3.4 Categorical Focal Loss (CFL)

This approach deals better with imbalanced class distributions and allows for the detection of less common classes; however, because the mainstream class has more weight in the loss calculation, it can sometimes exacerbate class imbalance by training the model to predict that class [44], thus focusing less on the underrepresented class. Two main mechanisms are used to address this bias in the Class-Balanced Focal Loss (CFL): the focusing parameter (γ). This parameter lowers the loss contribution from samples that are already correctly classified, allowing the model to focus more on instances that are more difficult to classify. Weighting factor (α): As shown in Eq. (21), this factor ensures that the samples of the underrepresented class have a greater impact on the learning process of the model by giving them a higher weight than the majority class.

$$CFL = \sum_{i=1}^{i=r} \alpha(i(probability)_i)^r \times \log((probability)_i) \quad (21)$$

These changes adjust the loss for correctly classified instances while placing greater importance on misclassified instances. This ensures the model's attention shifts from categories with an abundance of instances to those that are under-represented. The output from BiLSTM and MHSA layers are passed into the transformer encoder block, which takes in positional embedding for handling temporal information; absolute position encoding is used traditionally but, in this study, relative position embedding has been employed, which is advantageous with the WCE data since it captures the evolving context of the WCE data by considering their temporal distance to other frames and paying attention to their relative relationships.

This enables the model to capture both short- and long-term dependencies in gastrointestinal tract examinations, which is particularly beneficial when processing WCE data with varying sequence durations. The output from the Transformer encoder block is passed to a global max pooling layer, which captures the most prominent features from the feature maps while reducing their dimensionality and computational complexity. This is followed by a dropout layer to mitigate overfitting and improve model generalization, and a fully connected layer with SoftMax activation for final classification. The fully connected layer transforms the extracted high-level features into class probabilities, enabling accurate and efficient detection and classification of gastrointestinal abnormalities from WCE data.

4 Dataset and Preprocessing

The performance and generalization capacity of the classification framework were assessed using two publicly accessible gastrointestinal endoscopic image datasets, such as Kvasir-Capsule-v1 [45] and Kvasir-v2 [46]. Kvasir-Capsule-v1 is a WCE dataset, while Kvasir-v2 is a traditional gastrointestinal endoscopy dataset that includes images from colonoscopies. We employed an image dataset for this investigation that included four classes: Ulcerative Colitis, Polyps, Dyed-Lifted Polyps, and Normal, as seen in Fig. 6.

The consolidated dataset contains a total of 6510 images, including 1590 Dyed-Lifted Polyps images, 1630 Normal images, 1640 Polyps images, and 1650 Ulcerative Colitis images present in Table 2. The number of images across the four classes is nearly equal, resulting in a well-balanced dataset with minimal class imbalance. A balanced class distribution helps reduce the risk of classifier bias toward majority classes and enables the model to learn discriminative features from each gastrointestinal condition more effectively.

4.1 Data Augmentation

Data augmentation techniques were employed to improve the model's generalization capability by artificially expanding the dataset. The applied augmentation techniques included 90° rotation in both clockwise and counterclockwise directions, 180° rotation (upside-down), horizontal and vertical flipping

with a 50% probability, and small random rotations within the range of -12° to $+12^{\circ}$, as shown in Fig. 7. Furthermore, horizontal shearing was applied within the range of -5° to $+5^{\circ}$. Exposure and brightness levels were varied within the range of -11% to $+11\%$ to simulate realistic variations in illumination and image intensity. Gaussian blurring was applied by randomly selecting the standard deviation from -11% to $+11\%$, while an appropriate odd-valued kernel size was used for the Gaussian filter.

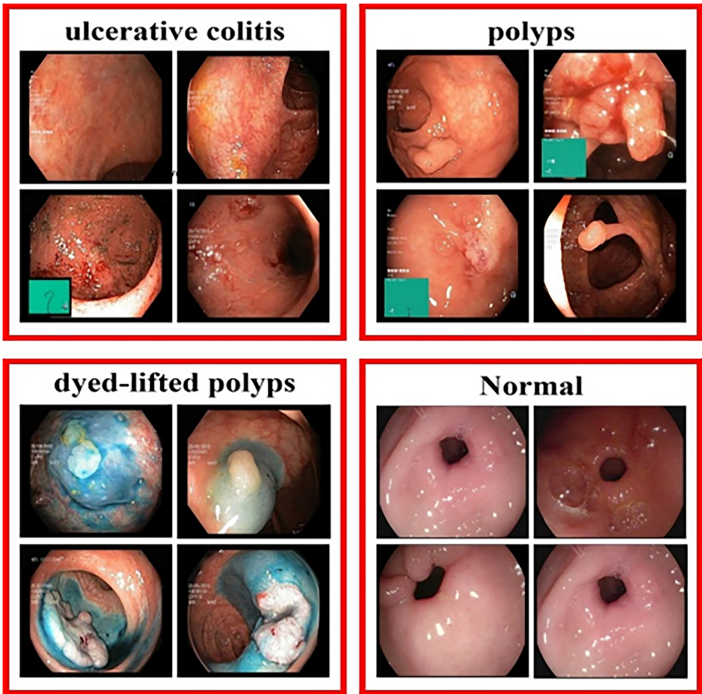


Figure 6: Sample WCE images present various ulcers.

Table 2: Total number of WCE images.

Class	Images
Dyed-lifted polyps	1590
Normal	1630
Polyps ulcer	1640
Ulcerative Colitis ulcer	1650

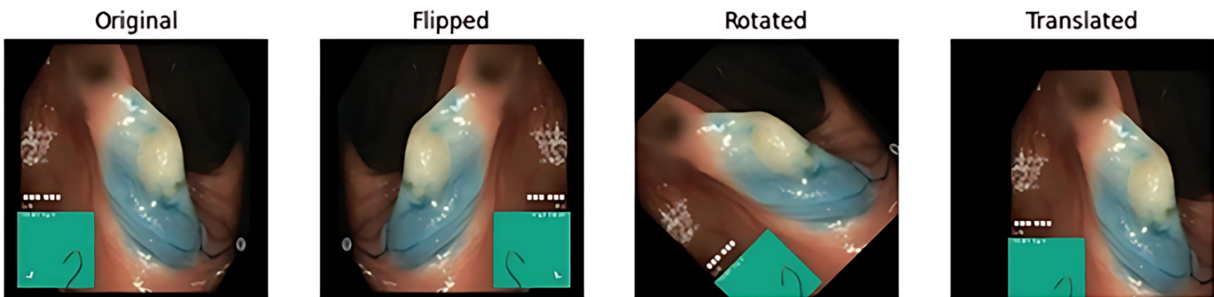


Figure 7: Sample image after applying the augmentation techniques.

4.2 Configuration of Proposed Model

In the proposed BiMT framework, each input where each frame represents extracted spatial features from the CNN backbone. The dataset was organized into four clinically relevant classes: ulcerative colitis, polyps, dyed-lifted polyps, and normal cases. The dataset was utilized to assess the performance of the proposed model using a holdout validation strategy, where the images were divided into training, validation, and testing subsets with a ratio of 70%, 20%, and 10%, respectively. The resolution of each frame is set to $256 \times 256 \times 3$ pixels to ensure efficient operation with the pre-trained CNN. To provide the model a variety of sample configurations and help it acquire more robust feature representation, the dataset was randomly shuffled after each epoch during training. The Stochastic Gradient Descent (SGD) optimizer is used to train the model for 50 epochs with a batch size of 32 and a learning rate of 0.0001. By lowering reliance on neurons, a dropout rate of 0.5 was included to reduce overfitting. Categorical cross-entropy was used as the loss function for multi-class classification, as described in Algorithm 1. Table 3 summarizes the computational complexity of the proposed model. The experiments were conducted using a workstation equipped with an Intel Core i5 11th Generation processor, 8 GB of system RAM, and an NVIDIA GeForce RTX 3080 Ti GPU with 12 GB of GDDR6X VRAM.

Table 3: Description of the trainable parameter measurement.

Description	Measurement
Trainable Parameters	Approximately 5.2 million
Model Training Duration	Roughly between 4 and 6 h
Typical Time per Prediction	Around 11.4 ms per sample
GPU Memory Utilization	Close to 790 megabytes

There are different deep learning models that were evaluated based on their performance measures. We experimented with different CNN architectures to find which models have better accuracy, recall, precision, F1-Score, specificity, and MCC present in Table 4.

Table 4: The classification outcome for various models.

Model	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)	Specificity (%)	MCC (%)
Inception V3	79.87	79.93	79.91	79.80	79.83	79.90
DenseNet-201	78.23	78.30	78.19	78.28	78.20	78.12
ResNet152	73.12	74.01	73.80	73.02	73.01	73.01
VGG-19	71.98	72.88	71.99	71.98	72.10	72.10
VGG-16	72.52	73.75	72.91	72.90	73.90	72.55
EfficientNetB6	75.53	76.52	75.50	75.30	75.78	76.10
ResNetV2	76.25	73.42	77.61	75.46	76.12	73.80

Table 4 presents the performance comparison of different pretrained CNN architectures on the WCE dataset in terms of classification accuracy. Among the evaluated architectures, Inception V3 achieved the highest classification accuracy of 79.87%, followed by DenseNet-201 with an accuracy of 78.23%.

Algorithm 1: Bidirectional motion temporal (BiMT) model for gastrointestinal disease classification

```

BEGIN
1. Dataset Integration
   DS ← Load(Kvasir-Capsule-v1, Kvasir-v2)
   Merge DS into four classes:
   C1 ← Ulcerative Colitis
   C2 ← Polyps
   C3 ← Dyed-Lifted Polyps
   C4 ← Normal
2. Image and Frame Preprocessing
   FOR each sample s ∈ DS DO
       Extract required frames from WCE sequences
   END FOR
   FOR each image I ∈ DS DO
       Resize(I, 256 × 256 × 3)
       Normalize(I)
   END FOR
3. Dataset Partitioning
   Training Set ← 70% of DS
   Validation Set ← 20% of DS
   Test Set ← 10% of DS
4. Feature Extraction
   FOR each input sequence X ∈ TrainingSet DO
       F ← CNN(X)
       Extract spatial features using pretrained CNN backbone
   END FOR
5. Temporal Feature Modeling
   T ← BiLSTM(F)
   A ← MHSA(T)
   R ← Transformer Encoder(A, Relative Positional Embedding)
6. Hyperparameter Configuration
   Learning Rate ← 0.0001
   Optimizer ← SGD
   Loss Function ← Categorical Focal Loss
   Batch Size ← 32
   Epochs ← 50
   Dropout Rate ← 0.5
   Early Stopping ← Enabled
7. Model Training
   FOR epoch = 1 TO 50 DO
       Shuffle(Training Set)
       Ŷ ← BiMT(Training Set)
       Loss ← Categorical Focal Loss(Y, Ŷ)
       Update model parameters using SGD

```

(Continued)

Algorithm 1 (continued)

```

    Validation Loss ← Validate(Model, Validation Set)
    IF Early Stopping(Validation Loss) = TRUE THEN
        BREAK
    END IF
END FOR
8. Model Evaluation
    Predictions ← BiMT(Test Set)
    Accuracy ← Calculate Accuracy(Y, Predictions)
    Recall ← Calculate Recall(Y, Predictions)
    Precision ← Calculate Precision(Y, Predictions)
    F1Score ← CalculateF1Score(Y, Predictions)
    Specificity ← Calculate Specificity(Y, Predictions)
    MCC ← Calculate MCC(Y, Predictions)
9. Classification Output
    Results ← {Accuracy, Precision, Recall, F1Score, AUC}
    RETURN Results
END

```

4.3 Model Performance

The classification models for ulcerative colitis, polyps, and dyed-lifted polyps were evaluated using several standard metrics, including accuracy, sensitivity, specificity, precision, F1-score, and Matthews correlation coefficient (MCC). To comprehensively evaluate the classification performance across all categories. Additionally, confusion matrices were generated for each of the four models to provide detailed analysis of classification outcomes. Accuracy represents the proportion of correctly classified images relative to the total number of images evaluated [47]. This metric is derived from the confusion matrix, which consists of four fundamental components: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The most used evaluation metric, accuracy, provides an overall measure of the model's classification performance. The percentage of real positive samples that were correctly classified as positive (true positives) is known as recall, also called sensitivity or the true positive rate.

$$Accuracy(\%) = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (22)$$

$$Precision = \frac{TP}{TP + FP} \quad (23)$$

$$Sensitivity/Recall = \frac{TP}{TP + FN} \quad (24)$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (25)$$

$$Specificity = \frac{TN}{TN + FP} \quad (26)$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (27)$$

4.4 Proposed Model

A Bidirectional Long Short-Term Memory (BiLSTM) system, Multi-Head Self-Attention (MHSA), and a modified Transformer module are integrated in the suggested model framework, BiMT. Based on the BiMT base model architecture, three models have been created and assessed. Each is designed for different experimental objectives.

InceptionV3-BiMT: This model serves as the lightweight model that uses InceptionV3 as the feature extractor, achieving a trade-off between accuracy and computational efficiency.

DenseNet201-BiMT: Uses DenseNet201 as the feature extractor to capture the essential information from all previous layers, improving feature reuse and representation quality.

ConvNeXt-BiMT: ConvNeXt improves upon conventional convolutional neural networks by incorporating design principles inspired by Transformer models while maintaining computational efficiency.

4.5 Results

The results show that the ConNeXt-BiMT model performance is better than the other versions of BiMT models such as InceptionV3-BiMT and DenseNet201-BiMT. The performance metrics of the BiMT base model on the WCE dataset are shown in Table 5. The second model, InceptionV3-BiMT, has moderate performance measures, as shown in Table 6. The third DenseNet201-BiMT model performance measure is shown in Table 7. The ConvNext-BiMT model performance measure is shown in Table 8. The standard BiMT model performs moderately yet consistently, averaging around 94.23% across all criteria. When InceptionV3 and BiMT are used together, that achieved the average accuracy of 96.33%, which means that they can better extract useful visual features from WCE images. The DenseNet201-BiMT variant shows a further improvement, with scores of about 97.48% on the same measures. This indicates that information flows deeper between layers and that features can be recycled better. The Precision, Recall, F1-Score, Specificity, MCC, and Accuracy of the ConvNeXt-BiMT model are about 98.41%, indicating that the ConvNeXt backbone can better learn and represent fine-grained spatial patterns and texture changes in WCE images for reliable classification of ulcerative colitis, polyps, dyed-lifted polyps, and normal.

Table 5: The BiMT model performance using the WCE dataset.

Class	Precision	Recall	F1-Score	Specificity	MCC	Accuracy
Ulcerative colitis	94.18	94.20	94.18	94.12	94.23	94.21
Polyps	94.05	94.08	94.04	94.09	94.05	94.04
Dyed-lifted polyps	94.21	94.28	94.24	94.24	94.27	94.28
Normal	94.38	94.37	94.38	94.37	94.38	94.37

Table 6: The InceptionV3-BiMT model performance using the WCE dataset.

Class	Precision	Recall	F1-Score	Specificity	MCC	Accuracy
Ulcerative colitis	96.42	96.41	96.40	96.43	96.42	96.44
Polyps	96.25	96.20	96.29	96.24	96.26	96.24
Dyed-lifted polyps	96.31	96.29	96.32	96.32	96.31	96.30
Normal	96.39	96.35	96.35	96.37	96.39	96.33

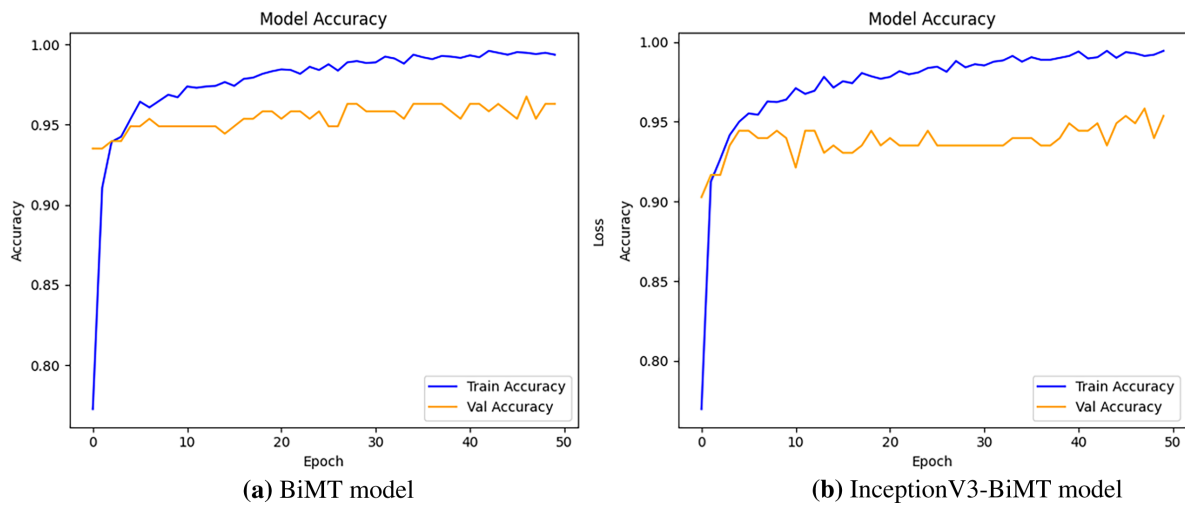
Table 7: The performance measure of the DenseNet201-BiMT model using the WCE dataset.

Class	Precision	Recall	F1-Score	Specificity	MCC	Accuracy
Ulcerative colitis	97.79	97.80	97.78	97.83	97.88	97.82
Polyps	97.21	97.18	97.25	97.24	97.18	97.18
Dyed-lifted polyps	97.48	97.41	97.41	97.44	97.45	97.41
Normal	97.55	97.48	97.48	97.51	97.52	97.52

Table 8: The performance measure of the ConvNeXt-BiMT model using the WCE dataset.

Class	Precision	Recall	F1-Score	Specificity	MCC	Accuracy
Ulcerative colitis	98.01	98.30	98.30	98.31	98.30	98.31
Polyps	98.25	98.30	98.31	98.29	98.28	98.26
Dyed-lifted polyps	98.61	98.64	98.61	98.65	98.60	98.61
Normal	98.45	98.39	98.41	98.43	98.47	98.45

The accuracy learning curves of each of the four models are shown in Fig. 8. The ConvNeXt-BiMT achieves the highest and most stable accuracy during training, showing good convergence and generalization. The lowest total loss is due to the consistent reduction in losses. The original BiMT converges to lower accuracy, whereas the DenseNet201-BiMT and InceptionV3-BiMT models exhibit steadily increasing performance. Fig. 9 presents the corresponding loss curves. ConvNeXt-BiMT exhibits a rapid and stable decrease in loss, reaching the lowest final loss among the evaluated models, which indicates effective optimization and robust feature learning. In Fig. 10, ConvNeXt-BiMT achieves the highest observed AUC values, ranging from 0.98 to 0.99, across the four classes on the test set, demonstrating strong discriminative capability between normal and abnormal images. These results indicate that the model maintains high classification performance while providing a favorable balance between sensitivity and specificity. In Fig. 11 presents the confusion matrix of various proposed models.

**Figure 8:** (Continued)

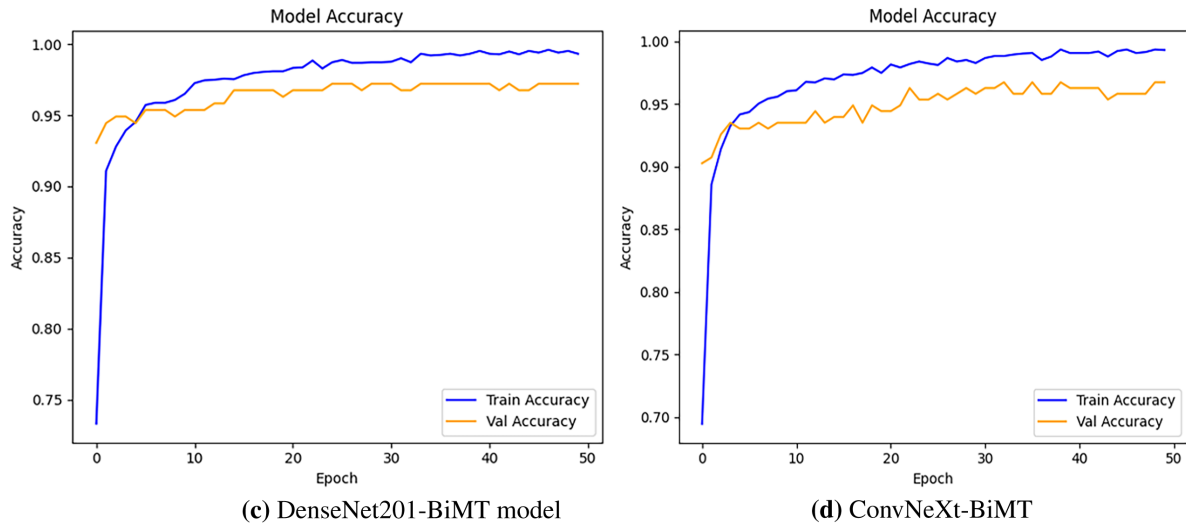


Figure 8: The various proposed models training and validation accuracy.

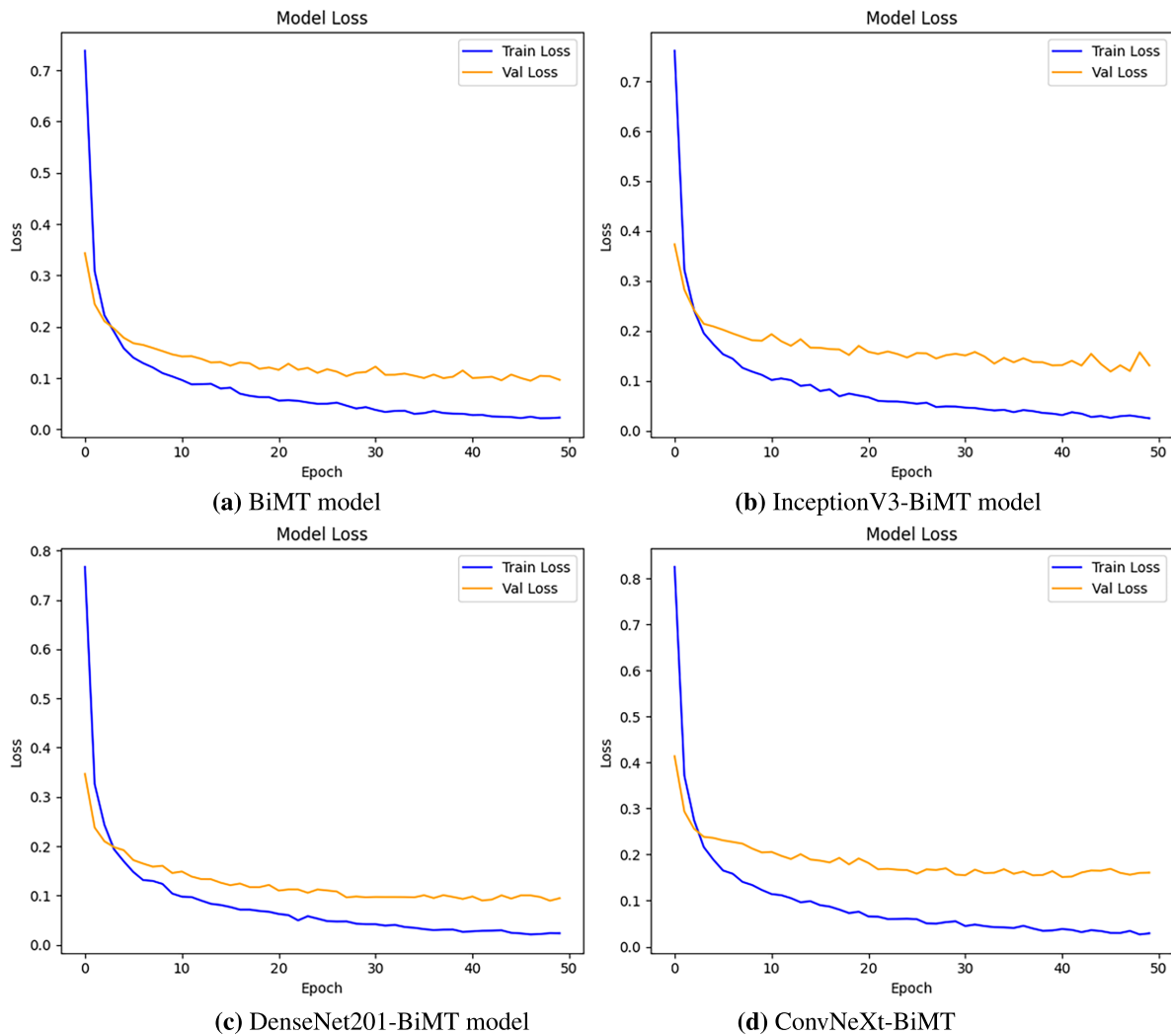


Figure 9: The four proposed models training and validation loss.

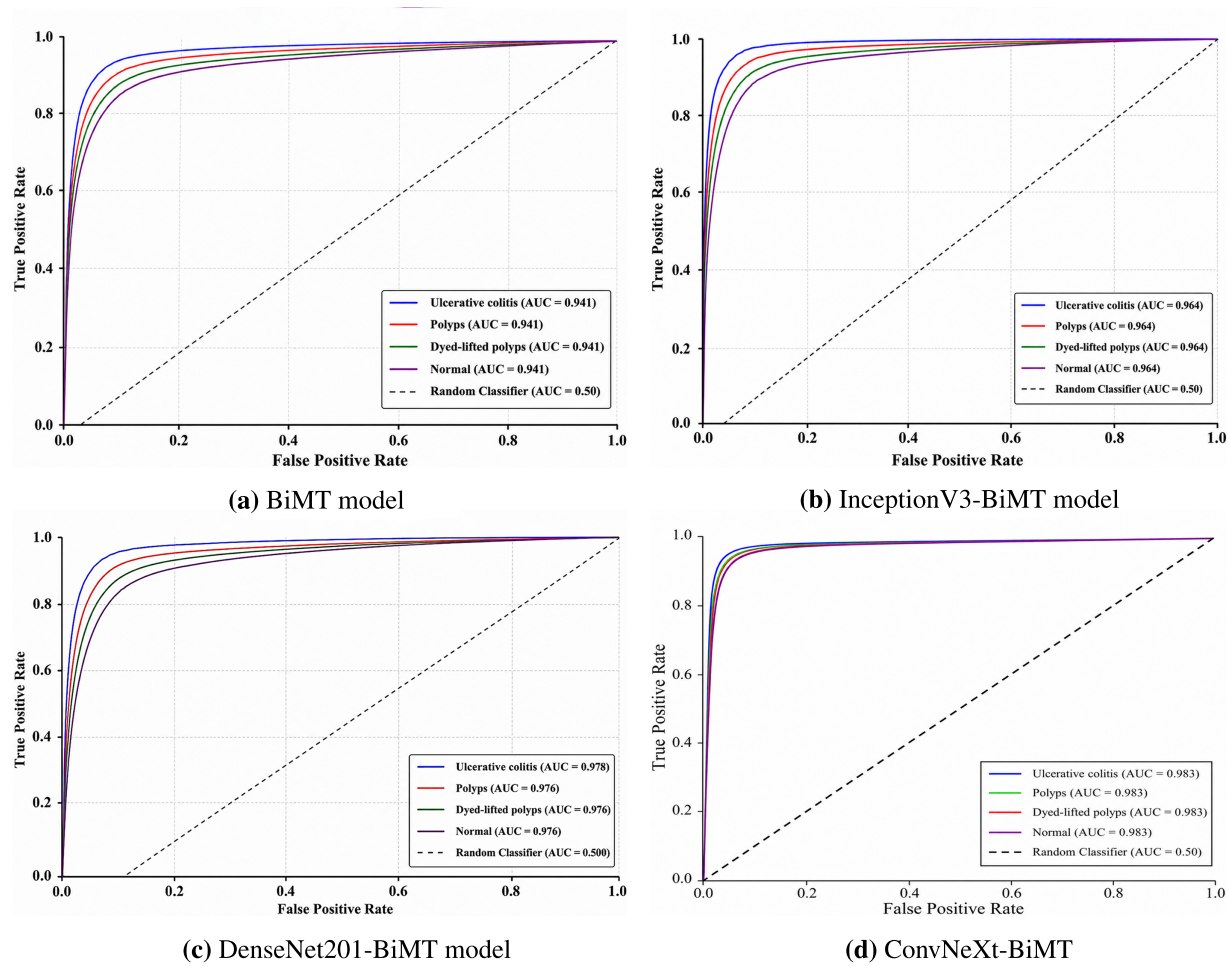


Figure 10: The ROC (AUC) curves of four proposed model.

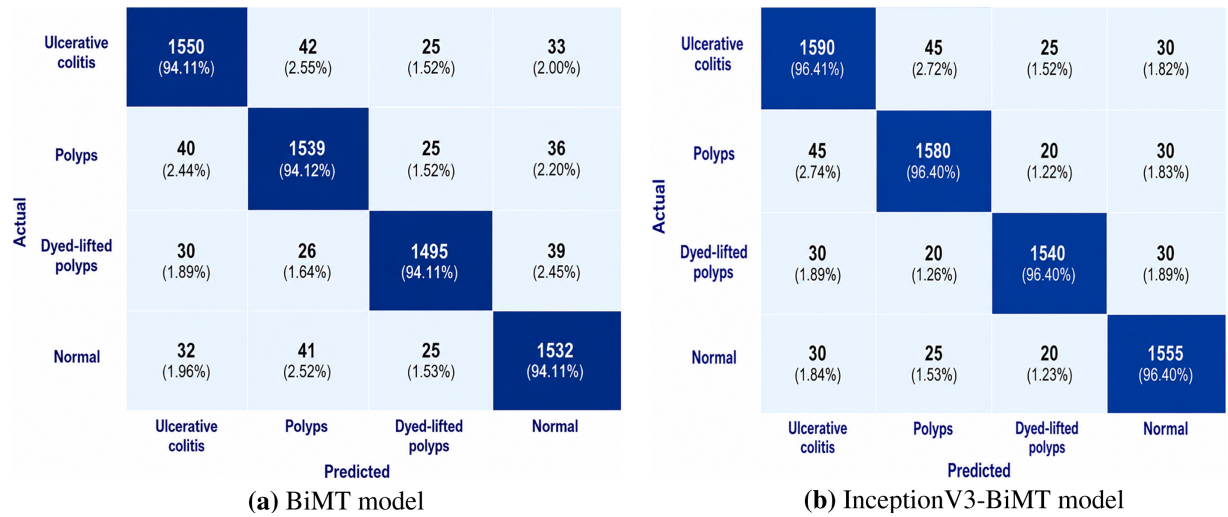


Figure 11: (Continued)

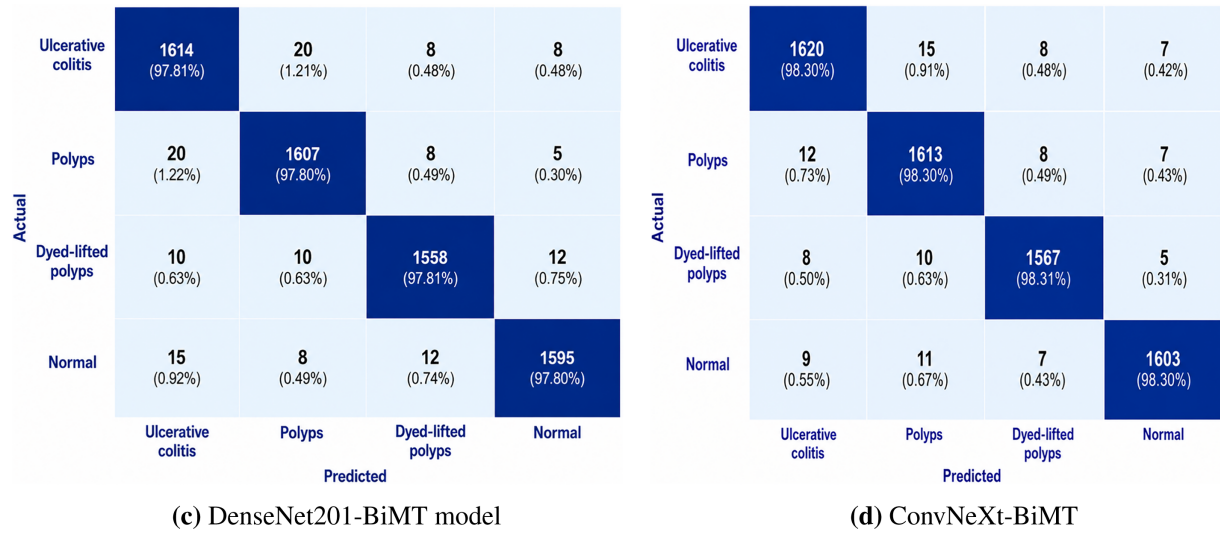


Figure 11: The confusion matrix of the four proposed models.

Table 9 compares the proposed model's performance with existing methods. Fig. 12 represents the result of the proposed model.

4.6 Ablation Study

The ablation study demonstrates that the proposed model achieved progressively improved classification performance across ulcerative colitis, polyps, dyed-lifted polyps, and normal as the batch size increased and the learning rate decreased. A batch size of 32 yielded the best overall performance, achieving 98.30% accuracy, recall 98.30%, precision 98.31%, and F1-score 98.31%, specifically 98.31%, which outperforms batch sizes 8 and 16. Likewise, a learning rate of 0.0001 consistently achieved superior results compared with learning rates of 0.1 and 0.001. Among the evaluated optimization algorithms, Stochastic Gradient Descent (SGD) consistently outperformed Adam and RMSProp, achieving the highest scores across all evaluation metrics. Overall, the combination of batch size 32, a learning rate of 0.0001, and the SGD optimizer proved to be the optimal configuration, delivering the most accurate, robust, and reliable classification performance across four classes.

Table 9: Comparison of the classification results of this study and other studies.

Author, Year, Reference	Methodology	Accuracy (%)
Rahman et al., 2024 [48]	EfficientNetB5	98.02
Bajhaiya et al., 2023 [49]	DenseNet121	94.01
El-Ghany et al., 2024 [50]	EfficientNet-B0	98.03
Nam et al., 2024 [51]	CNN + LSTM	95.80
Khan et al., 2025 [22]	DenseNet201 with Self-Attention	95.10
Chen et al., 2024 [52]	CNN + PyQt5	87.17
Proposed Model	ConvNeXtvl-BiMT	98.41

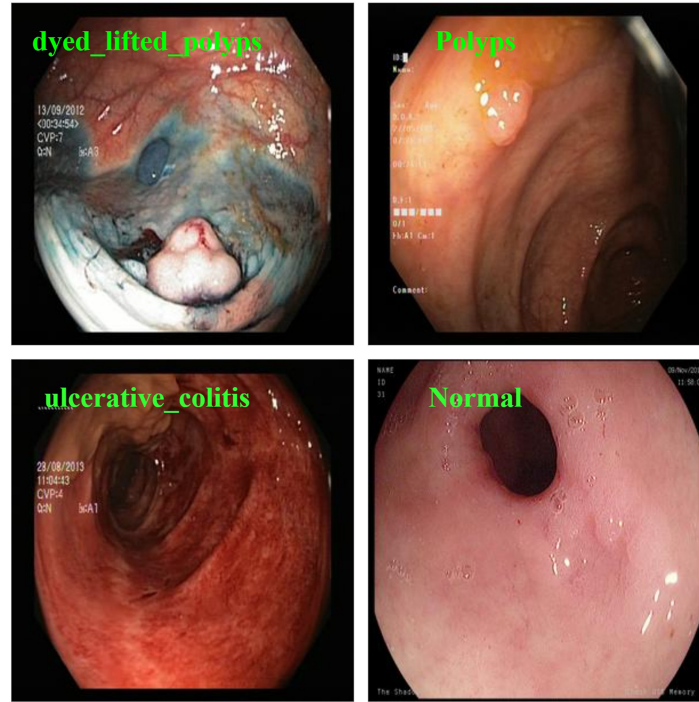


Figure 12: The proposed ConvNeXt v2-BiMT model's test results.

5 Conclusion, Limitation and Future Directions

In this study, we developed a deep learning-based fusion framework for the multi-class classification of ulcerative colitis, polyps, dyed-lifted polyps, and normal in the WCE dataset. The proposed framework combines advanced feature extraction and sequential learning techniques to effectively identify complex patterns in WCE data. BiLSTM model is a combination of an enhanced transformer module and CNN with MHSA to extract the best contextual features and visual information from successive endoscopic images. The ConvNeXt-BiMT achieved the best performance with an accuracy of 98.41%, followed by DenseNet201-BiMT (97.48%), InceptionV3-BiMT (96.33%), and the original BiMT model (94.23%). The suggested method improved the performance by using the temporal and contextual information with intricate visual representations. This approach improves the performance and accuracy of gastrointestinal lesion classification in the WCE dataset by taking into account lesion features and overall image dependencies. These results show that ConvNeXt is a viable method for automated WCE image analysis, as it increases classification accuracy, improves generalization, and enhances model stability when integrated into the BiMT framework. Despite the promising performance of the proposed model, it has a few limitations, such as the study focusing primarily on the three gastrointestinal disorders. Therefore, the framework does not currently address other clinically important tasks such as lesion localization or segmentation, which are essential for comprehensive gastrointestinal examinations. In future work, we aim to further develop the proposed framework for more clinically relevant applications, particularly lesion localization and segmentation, while incorporating temporal modeling to better capture the relationships between consecutive frames in video data. Moreover, improving human-AI interaction through clear and interpretable visualizations, along with a confidence-aware decision support mechanism.

Acknowledgement: The authors present their appreciation to King Saud University for funding this research through the Researchers Supporting Program number (ORF-2026-1776), King Saud University, Riyadh, Saudi Arabia.

Funding Statement: The authors present their appreciation to King Saud University for funding this research through the Researchers Supporting Program number (ORF-2026-1776), King Saud University, Riyadh, Saudi Arabia.

Author Contributions: Conceptualization, Sarfaraz Natha, and Mohammad Siraj; methodology, Kashan Memon, and Mohammad Siraj; software, Sarfaraz Natha, and Mohammad Siraj; validation, Mohammed Muflih Alamer, and Aaqid Syed; formal analysis, Ayesha Shafique, and Kashan Memon; investigation, Sarfaraz Natha and Aaqid Syed; resources, Sarfaraz Natha and Aaqid Syed; data curation, Ayesha Shafique; writing original draft preparation, Sarfaraz Natha, and Mohammed Muflih Alamer; writing, review and editing, Kashan Memon, Aaqid Syed and Mohammad Siraj; visualization, Ayesha Shafique, and Sarfaraz Natha; supervision, Mohammed Muflih Alamer, Sarfaraz Natha, and Aaqid Syed; project administration, Mohammed Muflih Alamer, and Kashan Memon; funding acquisition, Mohammed Muflih Alamer, and Mohammad Siraj. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets are publicly available at: Kvasir (<https://www.kaggle.com/datasets/meetnagadia/kvasir-dataset>), and Kvasir v2 (<https://www.kaggle.com/datasets/plhalvorsen/kvasir-v2-a-gastroin-testinal-tract-dataset>) Accessed on 28 June 2025.

Ethics Approval: Not applicable.

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

References

1. Salahi-Niri A, Nabavi-Rad A, Monaghan TM, Rokkas T, Doulberis M, Sadeghi A, et al. Global prevalence of *Helicobacter pylori* antibiotic resistance among children in the world health organization regions between 2000 and 2023: a systematic review and meta-analysis. *BMC Med.* 2024;22(1):598. doi:10.1186/s12916-024-03816-y.
2. 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.
3. 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.
4. Su CC, Chou CK, Mukundan A, Karmakar R, Sanbatcha BF, Huang CW, et al. Capsule endoscopy: current trends, technological advancements, and future perspectives in gastrointestinal diagnostics. *Bioengineering.* 2025;12(6):613. doi:10.3390/bioengineering12060613.
5. Rochmawati N, Fatichah C, Amaliah B, Raharjo BA, Dumont F, Thibaudeau E, et al. Deep learning-based lesion detection in endoscopy: a systematic literature review. *IEEE Access.* 2025;13(1):43532–56. doi:10.1109/ACCESS.2025.3548167.
6. Sharmila V, Geetha S. A recurrent multimodal sparse transformer framework for gastrointestinal disease classification. *Sci Rep.* 2025;15(1):24206. doi:10.1038/s41598-025-08897-0.
7. Habe TT, Haataja K, Toivanen P. Precision enhancement in wireless capsule endoscopy: a novel transformer-based approach for real-time video object detection. *Front Artif Intell.* 2025;8:1529814. doi:10.3389/frai.2025.1529814.
8. Kuo HY, Lee KH, Chou CK, Mukundan A, Karmakar R, Chen TH, et al. Deep learning-enhanced prediction of small intestinal bleeding points using long short-term memory networks. *World J Gastroenterol.* 2026;32(15):116105. doi:10.3748/wjg.v32.i15.116105.
9. Yuan Y, Li B, Meng MQH. WCE abnormality detection based on Saliency and adaptive Locality-Constrained linear coding. *IEEE Trans Autom Sci Eng.* 2017;14(1):149–59. doi:10.1109/TASE.2016.2610579.
10. Sahafi A, Koulaouzidis A, Naemi A. Artificial intelligence in gastrointestinal wireless capsule endoscopy: a systematic literature review and meta-analysis. *Diagnostics.* 2026;16(9):1269. doi:10.3390/diagnostics16091269.

11. 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.
12. Abbas MJ, Alshaya H, Bouchelligua W, Hassan N, Nasir IM. Hierarchical multi-stage attention and dynamic expert routing for explainable gastrointestinal disease diagnosis. *Diagnostics.* 2025;15(21):2714. doi:10.3390/diagnostics15212714.
13. Chou YT, Hsieh SY, Lin PC, Kuo HY, Chou HH. A GAN-based with expert-validated data augmentation method for wireless capsule endoscopy images of small intestine polyp. *J Supercomput.* 2025;81(5):653. doi:10.1007/s11227-025-07146-5.
14. Ramzan M, Raza M, Khan ZF, Khan MA, Baćanin-Džakula N, Damaševičius R, et al. A review on computer-aided diagnostic system to classify the disorders of the gastrointestinal tract. *Eur J Med Res.* 2025;30(1):674. doi:10.1186/s40001-025-02718-w.
15. Saraei M, Lalinia M, Lee EJ. Deep learning-based medical object detection: a survey. *IEEE Access.* 2025;13(5):53019–38. doi:10.1109/ACCESS.2025.3553087.
16. Ahmad S, Kim JS, Park DK, Whangbo T. Automated detection of gastric lesions in endoscopic images by leveraging attention-based YOLOv7. *IEEE Access.* 2023;11:87166–77. doi:10.1109/ACCESS.2023.3296710.
17. Liu Y, Zhang L, Hao Z, Yang Z, Wang S, Zhou X, et al. An xception model based on residual attention mechanism for the classification of benign and malignant gastric ulcers. *Sci Rep.* 2022;12(1):15365. doi:10.1038/s41598-022-19639-x.
18. Rubab S, Jamshed M, Khan MA, Almujaally NA, Damaševičius R, Hussain A, et al. Gastrointestinal tract disease classification from wireless capsule endoscopy images based on deep learning information fusion and newton Raphson controlled marine predator algorithm. *Sci Rep.* 2025;15(1):32180. doi:10.1038/s41598-025-17204-w.
19. Dinola AA, Madhubashini V, Poornimadevi K, Sridharan M, Venkatesh R, Yogeshwar MJ. Optimized deep CNN model for gastrointestinal imaging and abnormality detection using ResNet50. In: *Proceedings of the 2025 4th International Conference on Automation, Computing and Renewable Systems (ICACRS); 2025 Dec 10–12; Pudukkottai, India.* p. 884–90. doi:10.1109/ICACRS67045.2025.11324311.
20. Garbaz A, Lafraxo S, Charfi S, Ansari EM, Koutti L, Salihoun M. Combined deep convolutional neural networks for abnormality classification in wireless capsule endoscopy images. *Multimed Tools Appl.* 2025;84(33):40809–37. doi:10.1007/s11042-025-20749-7.
21. Asif S, Ying R, Qu J, Wang VY, yao J, GastricNet X D. An efficient and lightweight deep neural network in mobile edge computing for gastrointestinal disease detection. *J Big Data.* 2026;13(1):97. doi:10.1186/s40537-026-01446-0.
22. Khan MA, Shafiq U, Hamza A, Mirza AM, Baili J, AlHammadi DA, et al. A novel network-level fused deep learning architecture with shallow neural network classifier for gastrointestinal cancer classification from wireless capsule endoscopy images. *BMC Med Inform Decis Mak.* 2025;25(1):150. doi:10.1186/s12911-025-02966-0.
23. Bajhaiya D, Unni NS. Deep learning-enabled detection and localization of gastrointestinal diseases using wireless-capsule endoscopic images. *Biomed Signal Process Control.* 2024;93(1):106125. doi:10.1016/j.bspc.2024.106125.
24. Zhang K, Yu Q, Liu Y, Duan Y, Lou Y, Xu W. AFR: an image-aided diagnostic approach for ulcerative colitis. *Biomed Signal Process Control.* 2025;105:107542. doi:10.1016/j.bspc.2025.107542.
25. Xie W, Hu J, Liang P, Mei Q, Wang A, Liu Q, et al. Deep learning-based lesion detection and severity grading of small-bowel Crohn's disease ulcers on double-balloon endoscopy images. *Gastrointest Endosc.* 2024;99(5):767–77.e5. doi:10.1016/j.gie.2023.11.059.
26. Murugesan M, Arieth MR, Balraj S, Nirmala R. Colon cancer stage detection in colonoscopy images using YOLOv3 MSF deep learning architecture. *Biomed Signal Process Control.* 2023;80(7):104283. doi:10.1016/j.bspc.2022.104283.
27. Karthikha R, Jamal ND, Rafiammal S. An approach of polyp segmentation from colonoscopy images using Dilated-U-net-Seg—a deep learning network. *Biomed Signal Process Control.* 2024;93(4):106197. doi:10.1016/j.bspc.2024.106197.

28. Oukdach Y, Garbaz A, Kerkaou Z, Ansari EM, Koutti L, Ouafdi EAF, et al. UViT-Seg: an efficient ViT and U-net-based framework for accurate colorectal polyp segmentation in colonoscopy and WCE images. *J Imaging Inform Med.* 2024;37(5):2354–74. doi:10.1007/s10278-024-01124-8.
29. Zhang X, Zheng X, Dong M, Zhang M. An endoscopic images and diagnostic records based multimodal method for severity grading of gastric cancer. *Biomed Signal Process Control.* 2025;109(3):107891. doi:10.1016/j.bspc.2025.107891.
30. Prakash M, Krishnamurthy GN. An innovative transfer learning-polyp detection from wireless capsule endoscopy videos with optimal key frame selection and depth estimation. *Biomed Signal Process Control.* 2025;108(1):107963. doi:10.1016/j.bspc.2025.107963.
31. Balasubramanian S, Devi SM, Kavitha K, Balasubramaniam S. Attention-enhanced residual U-net with histogram equalization for automated classification of gastrointestinal bleeding disorders. *Biomed Signal Process Control.* 2026;111(7):108371. doi:10.1016/j.bspc.2025.108371.
32. Malik H, Naeem A, Sadeghi-Niaraki A, Naqvi RA, Lee SW. Multi-classification deep learning models for detection of ulcerative colitis, polyps, and dyed-lifted polyps using wireless capsule endoscopy images. *Complex Intell Syst.* 2024;10(2):2477–97. doi:10.1007/s40747-023-01271-5.
33. Li X, Wu Q, Chen Y, Wu K, Meng L. Wireless capsule endoscopy diagnosis using prototype self-attention and dynamic curriculum learning. *IEEE Trans Autom Sci Eng.* 2025;22:14260–71. doi:10.1109/TASE.2025.3558934.
34. Singh P, Singh S, Shukla MK. LightGastroFormer: a lightweight multi-resolution transformer for gastrointestinal disease classification. *Sci Rep.* 2026;16(1):23709. doi:10.1038/s41598-026-56751-8.
35. Li X, Wu Q, Wu K. Wireless capsule endoscopy anomaly classification via dynamic multi-task learning. *Biomed Signal Process Control.* 2025;100(5):107081. doi:10.1016/j.bspc.2024.107081.
36. Yakar İ, Kuçak RA, Bilgi S, Ferhanoglu O, Akinci TC. A hybrid deep learning and optical flow framework for monocular capsule endoscopy localization. *Electronics.* 2025;14(18):3722. doi:10.3390/electronics14183722.
37. Attallah O, Aslan MF, Sabanci K. EndoNet: a multiscale deep learning framework for multiple gastrointestinal disease classification via endoscopic images. *Diagnostics.* 2025;15(16):2009. doi:10.3390/diagnostics15162009.
38. 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.
39. Siraj M, Natha SAS, Alamer MM, Telba A, Syed A, Shafique A. High-precision classification of WCE-based gastrointestinal abnormality using a fusion deep learning approach. *Sci Rep.* 2026;16(1):21498. doi:10.1038/s41598-026-49370-w.
40. Güler O. StackDeVNet: an Explainable stacking ensemble of DenseNets and vision transformers for advanced gastrointestinal disease detection. *Int J Imaging Syst Technol.* 2026;36(1):e70275. doi:10.1002/ima.70275.
41. Coelho P, Pereira A, Leite A, Salgado M, Cunha A. A deep learning approach for red lesions detection in video capsule endoscopies. In: Campilho A, Karray F, Ter Haar Romeny B, editors. *Image analysis and recognition, Lecture notes in computer science.* Vol. 10882. Cham, Switzerland: Springer International Publishing; 2018. p. 553–61.
42. Jeeva S, Anand S. Diagnosis of hemorrhage in wireless capsule endoscopy images using a hybrid convolutional neural network-long short term memory (CNN-LSTM) model. In: *Proceedings of the 2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAIAI); 2025 March 28–29; Chennai, India.* p. 1–6.
43. Vaswani A, Shazeer N, Parmar N. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA.* p. 6000–10.
44. Kerkhof M, Wu L, Perin G, Picek S. Focus is key to success: a focal loss function for deep learning-based side-channel analysis. In: Balasch J, O’Flynn C, editors. *Constructive side-channel analysis and secure design, Lecture notes in computer science.* Vol. 13211. Cham, Switzerland: Springer International Publishing; 2022. p. 29–48. ISBN 978-3-030-99765-6.
45. Smedsrud PH, Thambawita V, Hicks SA, Gjestang H, Nedrejord OO, Næss E, et al. Kvasir-capsule, a video capsule endoscopy dataset. *Sci Data.* 2021;8(1):142. doi:10.1038/s41597-021-00920-z.

46. Tamimi AKA, Zena HK, Aljuboori AM. Deep learning for multi-class gastrointestinal endoscopy: a survey of recent advances and reliability challenges. *J Al-Qadisiyah Comput Sci Math.* 2026;18(2):183–98. doi:10.29304/jqcs.2026.18.22677.
47. Fasihi-Shirehjini O, Babapour-Mofrad F. Effectiveness of ConvNeXt variants in diabetic feet diagnosis using plantar thermal images. *Quant InfraRed Thermogr J.* 2025;22(2):155–72. doi:10.1080/17686733.2024.2310794.
48. Rahman ZAMJD, Mythili R, Chokkanathan K, Mahesh TR, Vanitha K, Yimer TE. Enhancing image-based diagnosis of gastrointestinal tract diseases through deep learning with EfficientNet and advanced data augmentation techniques. *BMC Med Imaging.* 2024;24(1):306. doi:10.1186/s12880-024-01479-y.
49. Bajhaiya D, Unni SN, Koushik AK. Deep learning-powered generation of artificial endoscopic images of GI tract ulcers. *iGIE.* 2023;2(4):452–63.e2. doi:10.1016/j.igie.2023.08.002.
50. El-Ghany SA, Mahmood MA, El-Aziz A. An accurate deep learning-based computer-aided diagnosis system for Gastrointestinal disease detection using wireless capsule endoscopy image analysis. *Appl Sci.* 2024;14(22):10243. doi:10.3390/app142210243.
51. Nam SJ, Moon G, Park JH, Kim Y, Lim YJ, Choi HS. Deep learning-based real-time organ localization and transit time estimation in wireless capsule endoscopy. *Biomedicines.* 2024;12(8):1704. doi:10.3390/biomedicines12081704.
52. Chen J, Xia K, Zhang Z, Ding Y, Wang G, Xu X. Establishing an AI model and application for automated capsule endoscopy recognition based on convolutional neural networks (with video). *BMC Gastroenterol.* 2024;24(1):394. doi:10.1186/s12876-024-03482-7.