EAI Endorsed Transactions
on AI and Robotics

Research Article EAI.EU

An Interpretable Hybrid Deep Learning Framework for Computer-Aided Detection of Gastrointestinal Diseases in Endoscopic Imaging

Shafiqul Islam Talukder[1], Md Jobaer Ahmed[2], Farhad Uddin Mahmud[3], Md Fokrul Islam Khan[4,*], Emon Hasan[5], Abu Kowshir Bitto[6]

[1]Department of Computer Science, Westcliff University, CA, USA
[2]College of Technology & Engineering, Westcliff University, CA, USA
[3]School of Business, International American University, CA, USA
[4]College of Business, Westcliff University, CA, USA
[5]Department of Information Technology, Washington University of Science and Technology, Alexandria, USA
[6]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh

Abstract

Gastrointestinal (GI) illnesses, especially gastric polyps and gastroesophageal reflux disease (GERD), are still ubiquitous diagnostic challenges with their complicated presentation and high inter-observer variation on endoscopy. The current research proposes `Xnception`, a new dual-backbone deep learning network that synergistically combines the `Xception` and `InceptionV3` architectures to facilitate classification robustness and feature expressivity for analysis of endoscopic images. Using transfer learning and fold-wise validation, the model is optimized for small-sized medical datasets with ensured generalizability. The fusion mechanism combines deep semantic representations of both backbones through specialized dense layers to enable precise discrimination between pathological and non-pathological classes. Explainable AI (XAI) techniques—such as Local Interpretable Model-agnostic Explanations, Integrated Gradients, and `Grad-CAM++`—are employed to visualize important regions impacting the model's predictions, hence ensuring transparency and clinical trustworthiness. Quantitative results on publicly available datasets demonstrate that `Xnception` outperforms its component individual models as well as other common baselines on several measures like accuracy, precision, and AUC. The proposed framework demonstrates promise to improve real-time diagnostic pipelines in gastroenterology and provides a scalable platform for AI-augmented endoscopic screening.

Received on 10 April 2026; accepted on 11 May 2026; published on 27 May 2026

Keywords: Medical Imaging, Gastrointestinal Disease, Endoscopic Image Analysis, Deep Learning, Explainable AI

doi:10.4108/airo.12566

1. Introduction

Gastrointestinal (GI) disease is a significant worldwide health problem, and it causes substantial morbidity and mortality in different populations. According to the World Health Organization (WHO), gastrointestinal diseases are responsible for more than 8 million deaths annually, with increasing prevalence due to aging populations, diet, and lifestyle [1]. Early diagnosis and treatment of GI disease are crucial in avoiding complications and improving patient outcomes. However, traditional endoscopic diagnosis is time-consuming, operator-dependent, and requires highly trained specialists, which in most of the world, and particularly in low-income countries, is a challenge [2].

Of all GI diseases, gastric polyps are mucosal surface growths of abnormal tissue that project from the stomach. Most polyps are benign, but others, such as adenomatous or hyperplastic polyps, are predisposed to malignant transformation to gastric cancer, especially if left undiagnosed or untreated [3]. Current data have indicated that in asymptomatic patients presenting to routine endoscopy, gastric polyps are present in the

*Corresponding author. Email: fokrulkhan837@gmail.com

range of 3% to 6% and occur more often in individuals over the age of 50 [4]. Therefore, early and accurate diagnosis of gastric polyps plays a significant role in cancer prevention and treatment.

Gastroesophageal reflux disease (GERD) is yet another common GI disorder, characterized by the retroflux of the contents of the stomach into the esophagus, producing symptoms like heartburn and regurgitation, and complications like esophagitis and Barrett's esophagus. Epidemiological data suggest that about 20% of adults in Western countries have GERD, with increasing rates in Asia and other developing countries [5]. Untreated chronic GERD can have a profound impact on quality of life and lead to esophageal adenocarcinoma. Visual endoscopy examination remains the first line in diagnosis, but inter-observer variation can make diagnostic performance challenging [6].

Over the past few years, the integration of artificial intelligence (AI) and computer vision techniques into the interpretation of medical images has offered a vast amount of potential for supporting clinical decision-making [7–9]. Deep learning models, particularly convolutional neural networks (CNNs), have demonstrated excellent performance in the automated classification of endoscopic images, typically on par with expert clinicians [10]. Transfer learning, employing pretrained models on large-scale datasets such as `ImageNet`, has been a powerful tool for medical tasks with limited labeled data [11]. Transformer- and deep learning-based approaches have been increasingly applied to clinical imaging to improve diagnostic accuracy and robustness. Recent studies have explored efficient architectures and hybrid learning strategies for medical decision support systems, highlighting the growing importance of AI in healthcare imaging applications [31–33]. Furthermore, explainable AI (XAI) methods such as `Grad-CAM` provide visual rationale for model outputs, increasing confidence and interpretability in clinical applications.

In this paper, we present `Xnception`, an `Xception-InceptionV3` dual-backbone transfer learning framework that combines the strengths of these architectures for the robust classification of endoscopic images of gastric polyps and GERD. The proposed model draws and fuses deep feature representations from both backbones with specialized dense layers to perform fine-grained classification. We evaluate fold-wise performance to enhance statistical stability and include XAI techniques to generate visual explanations which identify disease-relevant areas inside input images. The proposed framework is evaluated with publicly available datasets, and experimental results indicate that it surpasses individual baseline models in performance and interpretability.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the proposed methodology. Section 4 presents the results and discusses the findings and interpretability analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2. Related Work

2.1. Artificial Intelligence in GERD and Polyp

Artificial intelligence (AI) has shown promise in the detection and diagnosis of gastrointestinal diseases, particularly in the detection of polyps and gastroesophageal reflux disease (GERD). Deep learning algorithms have demonstrated high accuracy rates in colonoscopy applications, achieving real-time polyp detection. Urban et al. [12] reported 96% accuracy, while Kavitha et al. described a CNN-based approach with 96.4% accuracy and an AUC of 0.991. Wang et al. [14] illustrated real-time computer-assisted detection systems. AI has also been applied to analyze endoscopic images for GERD diagnosis and impedance metric measurement. Deep convolutional neural networks (DCNNs) have successfully improved polyp detection rates during colonoscopy, enhancing clinical practice efficiency and accuracy.

2.2. Transfer Learning in GERD and Polyp

Transfer learning has been one of the major techniques to enhance the accuracy and efficiency of polyp detection in wireless capsule endoscopy (WCE) and colonoscopy. Transfer learning involves making use of pre-trained models for related tasks to utilize the features from images that would be hard to analyze due to variability in the size, shape, and texture of polyps. Souaidi and Ansari [17] proposed a new multi-scale pyramidal fusion single-shot multibox detector network (MP-FSSD) for small polyp detection from WCE and colonoscopy images. Deep transfer learning was employed in this study to utilize prior knowledge, enhancing the ability of the model to capture representative features under varying illumination conditions. Belabbes et al. [18] highlighted the use of a one-shot detector based on deep transfer learning, where they saw an improvement in small polyp area detection and contextual feature learning, which are critical for correct diagnosis. In the detection of polyps, pre-trained CNNs have been employed as backbone models. Shin et al. [19] employed a region-based CNN in combination with a transfer learning approach to automatically identify colonic polyps from endoscopic images. Their findings indicate that fine-tuning a pre-trained model, such as ResNet, is capable of enhancing detection rates, transcending challenges emanating

from inconsistencies in polyIntegrating transfer learning enables these models to generalize more effectively and better across diverse polyp traits, resulting in more accurate screening outcomes. Improved performance in computer-assisted polyp classification has been demonstrated by Jasim et a,l. [20] who suggested a model utilizing Northern Goshawk Optimization combined with transfer learning methods, with an emphasis on improved feature extraction from polyp images.

2.3. Explainable AI in GERD and Polyp

The use of Explainable Artificial Intelligence (XAI) in gastroenterology, and particularly in gastroesophageal reflux disease (GERD) and colorectal polyp detection, has been in the spotlight for its potential to enhance the accuracy of diagnosis as well as the confidence of clinicians. Given the increasing use of AI in medical diagnostics, particularly image analysis during procedures such as colonoscopy and GERD diagnosis, it becomes essential to develop systems by which healthcare providers can understand the decision-making of these systems. Deep neural networks, or convolutional neural networks (CNNs), have significantly advanced polyp detection during colonoscopy. Using XAI techniques such as Layer-wise Relevance Propagation (LRP), clinicians can visualize why some elements of the input data cause specific predictions from the AI system; this is essential to ensure practitioners' trust [21]. Moreno-Sánchez [22] suggests that XAI systems contribute value to the healthcare professionals' decision-making process. By providing insight into how a model makes its decisions, XAI helps clinicians make informed decisions that can affect patient care directly.

3. Methodology

Figure 1 presents the workflow of the proosed gastrointestinal disease classification framework. A customized gastrointestinal endoscopic image dataset containing four classes (GERD, Normal GERD, Polyp, and Normal Polyp) was first collected and preprocessed through cropping, resizing, and transformation. Data augmentation techniques such as rotation, flipping, and zooming were applied to improve data diversity and reduce overfitting.

Three transfer learning models, Xception, VGG19, and InceptionV3, were initially evaluated for comparative analysis. A hybrid deep learning framework was then developed by combining Xception and InceptionV3 to leverage their complementary feature extraction capabilities. Fold-wise cross-validation was employed to ensure robustness and stability during training. To evaluate generalizability, the proposed framework was further validated on the publicly available Kvasir dataset.

Table 1. Class-wise data distribution of the datasets used in this study

Dataset	Class	Samples	Proportion
GastroEndoNet [25]	GERD	974	24.3%
	GERD Normal	1,103	27.5%
	Polyp	779	19.4%
	Polyp Normal	1,150	28.7%
	Total	**4,006**	**100%**
Kvasir [26]	Dyed Lifted Polyp	500	12.5%
	Dyed Resection Margin	500	12.5%
	Esophagitis	500	12.5%
	Normal Cecum	500	12.5%
	Normal Pylorus	500	12.5%
	Normal Z-Line	500	12.5%
	Polyps	500	12.5%
	Ulcerative Colitis	500	12.5%
	Total	**4,000**	**100%**

Performance was assessed using accuracy, precision, recall, F1-score, TPR, TNR, FPR, and FNR. In addition, Explainable AI (XAI) techniques, including Integrated Gradients (IG), Grad-CAM++, and LIME, were applied to provide visual interpretations of the model predictions.

3.1. Dataset Description

To evaluate the proposed model, two endoscopic image datasets were utilized: an in-house clinical dataset serving as the primary training and evaluation benchmark, and the publicly available Kvasir dataset employed to assess the generalizability of the model across an independent data source.

GastroEndoNet: Primary Dataset The primary dataset employed in this study comprises a total of 4,006 endoscopic images collected from clinical procedures conducted at Zainul Haque Sikder Women's Medical College & Hospital (Pvt.) Ltd. The dataset encompasses four distinct diagnostic categories: Gastroesophageal Reflux D isease (GERD), G ERD N ormal, P olyp, and Polyp Normal. Originally curated by the authors of this study, the dataset has since been made publicly available through *Data in Brief* [25], ensuring direct provenance and clinical relevance of the collected samples. As detailed in Table 1, the class-wise distribution reflects the natural variability observed in clinical endoscopic settings, with sample counts ranging from 779 (Polyp) to 1,150 (Polyp Normal). Representative samples from each category are illustrated in the top row of Figure 2.

Kvasir: External Benchmark Dataset To assess the generalizability of the proposed model beyond the in-house data, the Kvasir dataset [26] was employed as an independent external benchmark. Kvasir is a well-established, publicly available gastrointestinal (GI) endoscopic image dataset developed by researchers

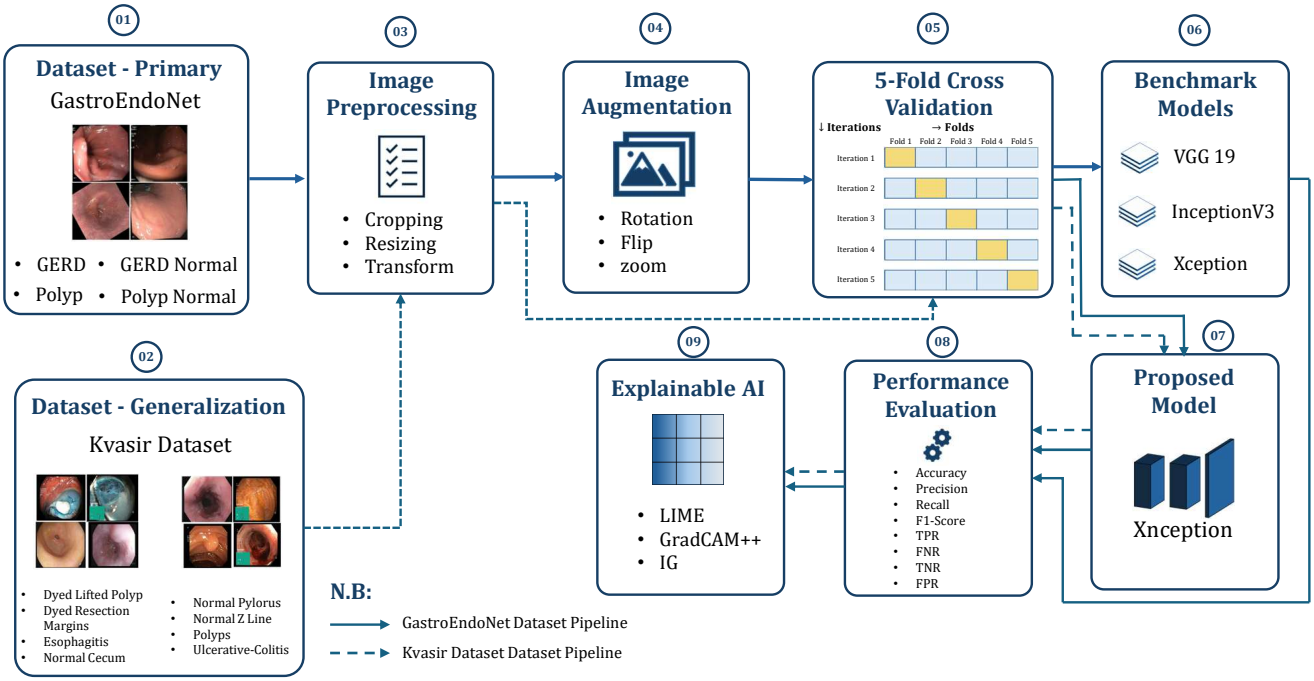

Figure 1. Workflow of the proposed gastrointestinal disease classification framework, including preprocessing, augmentation, transfer learning, hybrid model development, validation, performance evaluation, and XAI-based interpretation

at Simula Research Laboratory, Norway, and is widely adopted in the computerized analysis of GI diseases. The dataset comprises 4,000 images distributed uniformly across eight clinically relevant classes: Dyed Lifted Polyps, Dyed Resection Margins, Esophagitis, Normal Cecum, Normal Pylorus, Normal Z-Line, Polyps, and Ulcerative Colitis, with 500 images per class (12.5% each), as summarized in Table 1. Image resolutions vary from 720×576 to 1920×1072 pixels, and images are organized into category-specific folders. Representative samples from the Kvasir dataset are presented in the bottom rows of Figure 2.

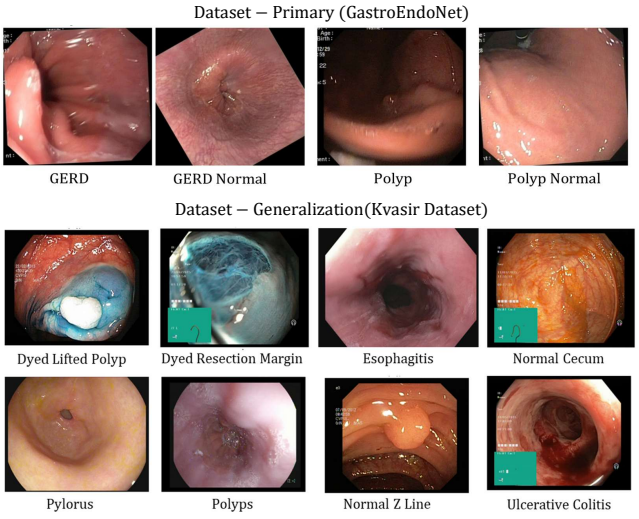

Figure 2. Representative endoscopic image samples from the two datasets used in this study. **Top row:** GastroEndoNet (primary dataset) – (A) GERD, (B) GERD Normal, (C) Polyp, (D) Polyp Normal. **Middle and bottom rows:** Kvasir dataset – (E) Dyed Lifted Polyps, (F) Dyed Resection Margins, (G) Esophagitis, (H) Normal Cecum, (I) Normal Pylorus, (J) Normal Z-Line, (K) Polyps, (L) Ulcerative Colitis

3.2. Data Preprocessing

To ensure consistency across both datasets and to promote robust model generalization, all images from the GastroEndoNet and Kvasir datasets were subjected to a standardized preprocessing and augmentation pipeline prior to model training, as illustrated in Figure 1. The pipeline was designed to address three key challenges common in clinical endoscopic imaging: input heterogeneity, class imbalance, and limited training diversity.

Image Preprocessing The preprocessing stage consisted of three sequential steps applied uniformly to both datasets. First, images were cropped to remove extraneous border regions and irrelevant peripheral artifacts that are commonly present in raw endoscopic captures. Following cropping, all images were resized to a uniform resolution of 224×224 pixels, satisfying the input requirements of the deep learning architectures

employed in this study. Finally, a normalization transform was applied, scaling pixel intensity values to the $[0, 1]$ range using a factor of $\frac{1}{255}$, which standardizes the input distribution and facilitates stable gradient-based optimization during training.

Class Balancing via Oversampling (GastroEndoNet Only) As presented in Table 1, the GastroEndoNet dataset exhibits moderate class imbalance, with sample counts ranging from 779 images (Polyp) to 1,150 images (Polyp Normal). To mitigate the risk of model bias toward over-represented categories, an oversampling strategy was applied exclusively to the GastroEndoNet dataset prior to training. Augmentation-based synthetic image generation was used to upsample all minority classes to match the majority class count of 1,150 samples per category, yielding a fully balanced training set of 4,600 images across the four classes. As the Kvasir dataset maintains a perfectly uniform distribution of 500 images per class, no class balancing was required for this dataset.

Real-Time Data Augmentation To further improve training diversity and enhance the model's robustness to geometric and orientational variability inherent in endoscopic imaging, a real-time data augmentation strategy was applied to both datasets during training, with a batch size of 32. The augmentation pipeline comprised three core transformations:

- **Random rotation:** up to 40°, simulating variations in endoscope orientation during clinical procedures
- **Horizontal flip:** randomly applied, reflecting natural left-right symmetry in gastrointestinal anatomy
- **Zoom:** within a range of 20%, accounting for variable imaging distances and endoscope proximity to tissue

All augmentations were performed on-the-fly during training rather than being pre-stored, effectively increasing the perceptual diversity of training batches at each epoch without expanding the dataset size on disk. A `nearest` fill mode was applied during spatial transformations to preserve boundary integrity at image edges.

Data Partitioning and Cross-Validation To obtain reliable and statistically robust performance estimates, a **5-fold cross-validation** scheme was adopted for model evaluation on the GastroEndoNet dataset. Under this scheme, the dataset was partitioned into five equal and mutually exclusive folds; in each of the five iterations, four folds were used for training and the remaining fold was held out for validation. Performance metrics were subsequently averaged across all five folds to yield a final unbiased estimate of model generalization. The Kvasir dataset was employed as a fully independent external test set to further validate cross-dataset generalizability of the proposed model, as denoted by the dashed pipeline in Figure 1.

3.3. Proposed Model Development

This section describes the benchmark architectures employed for comparative evaluation, followed by a detailed formulation of the proposed dual-backbone model, Xnception, which integrates Xception and InceptionV3 into a unified classification framework.

Benchmark Models Three established deep convolutional neural network architectures were adopted as benchmark models to provide a rigorous comparative baseline: VGG19, InceptionV3, and Xception. Each model was initialized with weights pre-trained on the ImageNet dataset [27], and their convolutional base layers were frozen to preserve the learned low- and mid-level feature representations, with task-specific classification heads appended and fine-tuned to the target endoscopic image classification task.

VGG19 [28] is a deep sequential architecture comprising 19 weight layers organized into blocks of 3×3 convolutional filters followed by fully connected layers. Its uniform and straightforward design makes it a widely adopted baseline in medical image classification tasks. InceptionV3 [29] employs factorized convolutions and inception modules to efficiently capture multi-scale spatial features at reduced computational cost, making it well-suited for the heterogeneous appearance of endoscopic imagery. Xception [30] extends the Inception paradigm by replacing standard inception modules with depthwise separable convolutions, enabling more efficient and expressive feature learning that has demonstrated strong performance across fine-grained visual recognition tasks. All three benchmark models were trained and evaluated under identical experimental conditions, including input resolution, optimization strategy, batch size, and cross-validation scheme, to ensure a fair and reproducible comparison with the proposed model.

Xnception: Proposed Dual-Backbone Model The proposed model, **Xnception**, is a hybrid deep learning architecture that synergistically combines the complementary feature extraction capabilities of Xception and InceptionV3 within a unified dual-backbone framework. Both backbones are initialized with ImageNet pre-trained weights, and their base layers are frozen to retain transferable low-level representations, while only the appended classification heads are fine-tuned to the target task of gastric polyp and GERD classification from endoscopic images.

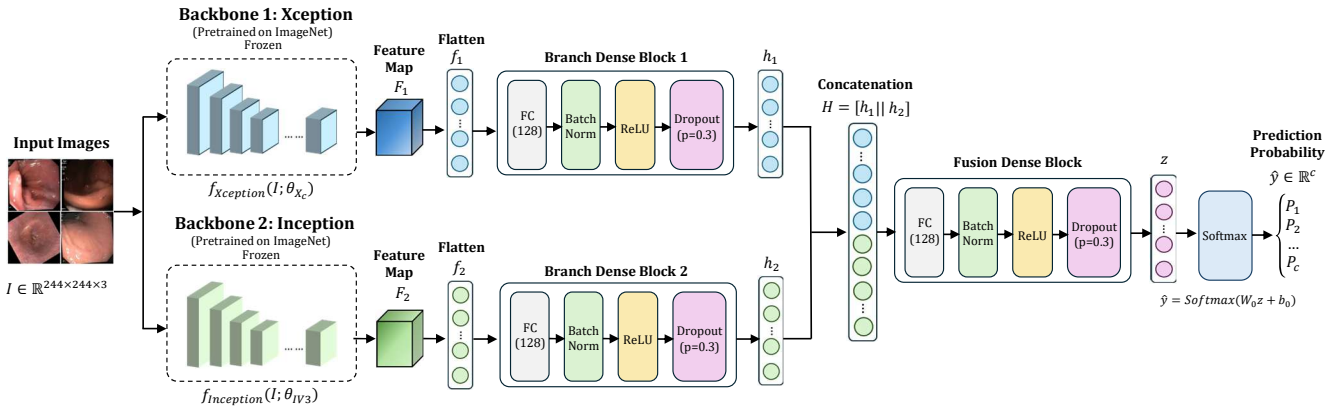

Figure 3. Architecture of the proposed Xnception dual-backbone framework combining Xception and InceptionV3 for feature fusion and gastrointestinal disease classification

Given an input image $\mathbf{I} \in \mathbb{R}^{224\times224\times3}$, the two frozen backbone networks independently extract deep feature maps $\mathbf{F}_1$ and $\mathbf{F}_2$ as follows:

$$\mathbf{F}_1 = f_{\text{Xception}}(\mathbf{I};\, \theta_{\text{Xc}}) \tag{1}$$

$$\mathbf{F}_2 = f_{\text{InceptionV3}}(\mathbf{I};\, \theta_{\text{IV3}}) \tag{2}$$

where θ_{Xc} and θ_{IV3} denote the frozen pre-trained weights of the Xception and InceptionV3 backbones, respectively. The resulting spatial feature maps are then flattened into one-dimensional vectors to prepare them for subsequent dense processing:

$$\mathbf{f}_1 = \text{Flatten}(\mathbf{F}_1) \tag{3}$$

$$\mathbf{f}_2 = \text{Flatten}(\mathbf{F}_2) \tag{4}$$

Each flattened vector is independently passed through a branch-specific dense block comprising a fully connected layer with 128 units, Batch Normalization (BN), ReLU activation, and Dropout ($p = 0.3$) to regularize learning and suppress overfitting:

$$\mathbf{h}_1 = \text{Dropout}_{0.3}(\text{ReLU}(\text{BN}(\mathbf{W}_1\mathbf{f}_1 + \mathbf{b}_1))) \tag{5}$$

$$\mathbf{h}_2 = \text{Dropout}_{0.3}(\text{ReLU}(\text{BN}(\mathbf{W}_2\mathbf{f}_2 + \mathbf{b}_2))) \tag{6}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{b}_1$, $\mathbf{b}_2$ are the learnable weight matrices and bias vectors of each branch dense layer, respectively. The branch-wise representations $\mathbf{h}_1$ and $\mathbf{h}_2$ are then concatenated along the feature dimension to form a unified joint representation that captures the complementary strengths of both backbone architectures:

$$\mathbf{H} = [\mathbf{h}_1 \,\|\, \mathbf{h}_2] \tag{7}$$

The concatenated representation $\mathbf{H}$ is passed through a second dense block with 64 units, Batch Normalization, ReLU activation, and an increased Dropout rate ($p = 0.5$) to further regularize the fused representation:

$$\mathbf{z} = \text{Dropout}_{0.5}(\text{ReLU}(\text{BN}(\mathbf{W}_3\mathbf{H} + \mathbf{b}_3))) \tag{8}$$

Finally, a Softmax output layer maps the processed representation $\mathbf{z}$ to a probability distribution over the C target classes, yielding the predicted class probability vector $\hat{\mathbf{y}} \in \mathbb{R}^C$:

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_o\,\mathbf{z} + \mathbf{b}_o) = \frac{e^{\mathbf{W}_o\,\mathbf{z}+\mathbf{b}_o}}{\sum_{c=1}^{C} e^{(\mathbf{W}_o\,\mathbf{z}+\mathbf{b}_o)_c}} \tag{9}$$

where $C = 4$ for the GastroEndoNet dataset (GERD, GERD Normal, Polyp, Polyp Normal) and $C = 8$ for the Kvasir dataset. The complete forward pass of the proposed Xnception framework is illustrated in Figure 3.

3.4. Performance Evaluation

The predictive performance of all models was assessed on the held-out test data following the completion of training. Evaluation was conducted using a comprehensive set of classification metrics derived from the confusion matrix, providing a thorough and multi-faceted characterization of model behavior across all target classes [31–33]. The metrics employed are formally defined as follows, where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

Accuracy measures the overall proportion of correctly classified samples across all classes:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{10}$$

Precision, also referred to as positive predictive value, quantifies the fraction of positive predictions that are truly positive:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{11}$$

Recall, equivalently the True Positive Rate (TPR) or sensitivity, measures the proportion of actual positives correctly identified by the model:

$$\text{Recall} = \text{TPR} = \frac{TP}{TP + FN} \times 100\% \tag{12}$$

The F1-Score represents the harmonic mean of Precision and Recall, providing a balanced measure particularly suited to imbalanced class distributions:

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{13}$$

The True Negative Rate (TNR), or specificity, measures the proportion of actual negatives correctly identified:

$$\text{TNR} = \frac{TN}{TN + FP} \times 100\% \tag{14}$$

The False Positive Rate (FPR) quantifies the proportion of actual negatives incorrectly classified as positive:

$$\text{FPR} = \frac{FP}{FP + TN} \times 100\% \tag{15}$$

The False Negative Rate (FNR) measures the proportion of actual positives incorrectly classified as negative, and is directly related to Recall as FNR = 1 – TPR:

$$\text{FNR} = \frac{FN}{FN + TP} \times 100\% \tag{16}$$

4. Results and Discussion

This section presents a comprehensive evaluation of the four deep learning models: InceptionV3, VGG19, Xception, and the proposed hybrid Xnception across two gastrointestinal endoscopic image datasets. All models were trained and evaluated using a 5-fold cross-validation scheme, with 10 epochs per fold, to ensure statistically robust and reproducible performance estimates. For each fold, class-wise metrics including Accuracy, Precision, Recall, F1-Score, True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) were recorded. The results are organized into three subsections: benchmark model evaluation on Dataset 1, proposed model evaluation on both datasets, and a comparative analysis supplemented by Explainable AI (XAI) interpretability assessment.

4.1. Benchmark Model Evaluation on Dataset 1

InceptionV3 The fold-wise classification performance of InceptionV3 on Dataset 1 GastroEndoNet is presented in Table 2. The model achieves a mean accuracy of 80.00% across all five folds, with its best performance recorded in Fold 4 at 82.02%. The overall TPR of 0.76 and TNR of 0.92 indicate that the model maintains a reasonable balance between sensitivity and specificity. A moderate FPR of 0.06 and a relatively elevated FNR of 0.20 suggest that while the model rarely raises false alarms, it occasionally fails to identify true positive cases, particularly for GERD, where recall drops as low as 0.67 in Fold 3. Polyp detection, by contrast, consistently exhibits high precision and recall across folds. Overall, InceptionV3 demonstrates stable and well-balanced predictive behavior with minor sensitivity limitations in GERD classification.

The training and validation accuracy curves in Figure 4(A) confirm a consistent upward trend across all folds, with validation accuracies approaching or exceeding 0.80 by the final epoch. Fold 4 and Fold 5 exhibit the strongest generalization, achieving training and validation accuracies in the range of 0.85 to 0.90. Folds 1 and 2 display slightly higher variability in validation accuracy, likely attributable to differences in the underlying data distribution of those folds. The loss curves in Figure 4(B) reflect a steep and consistent decline in both training and validation loss across epochs, confirming effective optimization. Folds 3, 4, and 5 exhibit the smoothest convergence profiles, while Fold 2 shows elevated validation loss variability, possibly indicative of localized overfitting. Taken together, these results confirm that InceptionV3 is a capable and stable baseline with strong generalization in high-performing folds.

VGG19 The fold-wise performance of VGG19 on Dataset 1 is summarized in Table 3. VGG19 records the lowest mean accuracy of 73.72% among all evaluated models, with its best fold performance of 74.42% achieved in Fold 3. The TPR of 0.72 and TNR of 0.90 are the lowest among the benchmark models, and the elevated FNR of 0.26 and FPR of 0.08 indicate a greater tendency to miss true positive cases and generate false alarms relative to InceptionV3 and Xception. GERD recall varies considerably across folds, ranging from 0.53 in Fold 4 to 0.77 in Fold 5, reflecting inconsistent sensitivity to this class. While the model performs reasonably well on Polyp Normal detection, its overall precision-recall trade-off is less stable, suggesting that VGG19's sequential architecture is less well-suited to capturing the fine-grained feature variability present in endoscopic imagery. The accuracy curves in Figure 5(A) demonstrate a general increasing trend across all folds, though validation accuracy remains comparatively moderate

Table 2. Fold-wise classification performance of InceptionV3 on Dataset 1

Fold	Class	Acc.	Prec.	Rec.	F1	TPR	TNR	FPR	FNR
1	GERD	80.25	0.79	0.70	0.74	0.69	0.94	0.05	0.30
	GERD Normal		0.74	0.86	0.79	0.85	0.88	0.11	0.14
	Polyp		0.92	0.72	0.81	0.71	0.98	0.01	0.28
	Polyp Normal		0.81	0.89	0.85	0.89	0.91	0.08	0.10
2	GERD	78.15	0.79	0.73	0.76	0.73	0.93	0.06	0.26
	GERD Normal		0.78	0.82	0.80	0.81	0.90	0.09	0.18
	Polyp		0.69	0.83	0.75	0.82	0.91	0.08	0.17
	Polyp Normal		0.86	0.76	0.81	0.75	0.95	0.04	0.24
3	GERD	79.47	0.80	0.67	0.73	0.66	0.94	0.05	0.33
	GERD Normal		0.73	0.87	0.80	0.87	0.86	0.13	0.12
	Polyp		0.89	0.73	0.81	0.73	0.98	0.01	0.26
	Polyp Normal		0.82	0.86	0.84	0.86	0.92	0.07	0.13
4	GERD	82.02	0.78	0.75	0.76	0.74	0.93	0.06	0.25
	GERD Normal		0.75	0.86	0.80	0.85	0.90	0.09	0.14
	Polyp		0.86	0.84	0.85	0.83	0.96	0.03	0.16
	Polyp Normal		0.90	0.84	0.87	0.84	0.95	0.04	0.15
5	GERD	80.16	0.83	0.67	0.74	0.74	0.66	0.95	0.33
	GERD Normal		0.71	0.87	0.78	0.78	0.86	0.86	0.13
	Polyp		0.90	0.78	0.83	0.83	0.77	0.97	0.22
	Polyp Normal		0.83	0.87	0.85	0.85	0.86	0.92	0.13

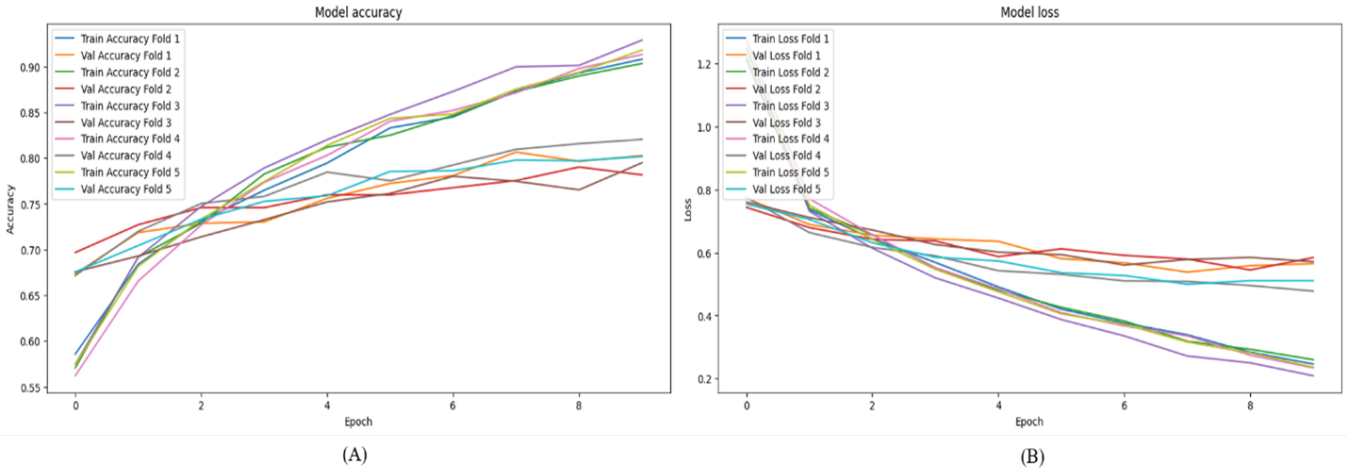

Figure 4. Fold-wise training and validation curves of InceptionV3 on Dataset 1. (A) Training vs. validation accuracy. (B) Training vs. validation loss

Table 3. Fold-wise classification performance of VGG19 on Dataset 1

Fold	Class	Acc.	Prec.	Rec.	F1	TPR	TNR	FPR	FNR
1	GERD	73.87	0.72	0.59	0.65	0.58	0.91	0.08	0.41
	GERD Normal		0.64	0.79	0.71	0.78	0.84	0.15	0.21
	Polyp		0.90	0.66	0.76	0.66	0.98	0.01	0.33
	Polyp Normal		0.78	0.89	0.83	0.88	0.90	0.09	0.11
2	GERD	73.87	0.72	0.55	0.62	0.54	0.93	0.06	0.45
	GERD Normal		0.65	0.84	0.73	0.83	0.83	0.16	0.16
	Polyp		0.93	0.64	0.76	0.64	0.98	0.01	0.35
	Polyp Normal		0.76	0.88	0.81	0.87	0.89	0.10	0.12
3	GERD	74.42	0.66	0.70	0.68	0.69	0.87	0.12	0.30
	GERD Normal		0.72	0.72	0.72	0.72	0.89	0.10	0.27
	Polyp		0.74	0.73	0.74	0.73	0.94	0.05	0.26
	Polyp Normal		0.85	0.81	0.83	0.80	0.94	0.05	0.19
4	GERD	73.54	0.67	0.53	0.59	0.52	0.92	0.07	0.47
	GERD Normal		0.66	0.83	0.74	0.83	0.82	0.17	0.16
	Polyp		0.87	0.68	0.76	0.67	0.97	0.02	0.32
	Polyp Normal		0.79	0.85	0.82	0.85	0.91	0.08	0.14
5	GERD	72.92	0.58	0.77	0.66	0.77	0.83	0.16	0.22
	GERD Normal		0.73	0.57	0.64	0.57	0.91	0.08	0.42
	Polyp		0.78	0.77	0.78	0.76	0.94	0.05	0.23
	Polyp Normal		0.85	0.82	0.83	0.81	0.94	0.06	0.18

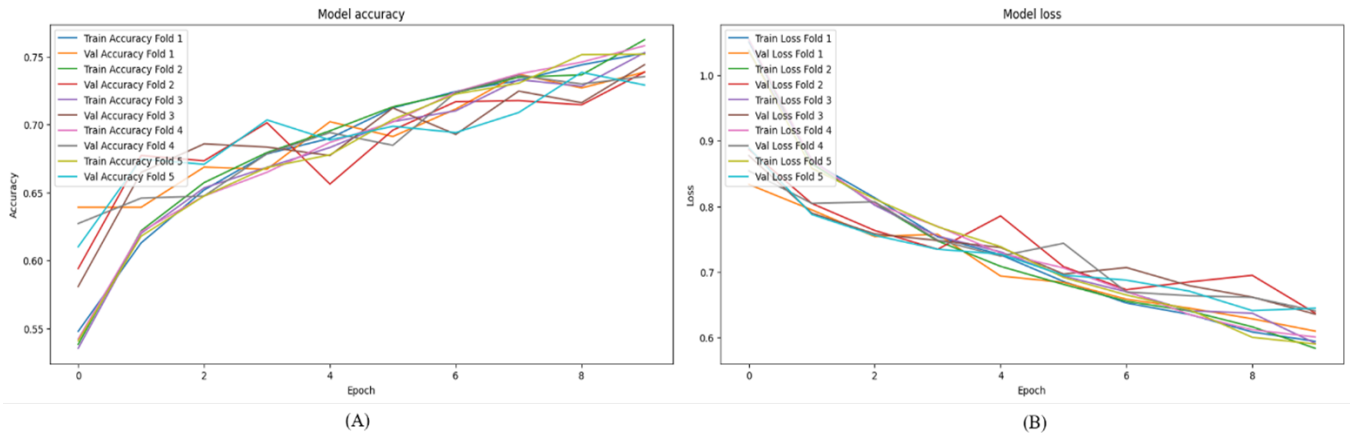

Figure 5. Fold-wise training and validation curves of VGG19 on Dataset 1. (A) Training vs. validation accuracy. (B) Training vs. validation loss

Table 4. Fold-wise classification performance of Xception on Dataset 1

Fold	Class	Acc.	Prec.	Rec.	F1	TPR	TNR	FPR	FNR
1	GERD	82.12	0.80	0.76	0.78	0.75	0.93	0.06	0.24
	GERD Normal		0.75	0.84	0.79	0.84	0.89	0.10	0.15
	Polyp		0.89	0.77	0.83	0.77	0.97	0.02	0.22
	Polyp Normal		0.87	0.89	0.88	0.88	0.94	0.05	0.11
2	GERD	80.64	0.74	0.81	0.77	0.80	0.90	0.09	0.19
	GERD Normal		0.80	0.74	0.77	0.74	0.92	0.07	0.25
	Polyp		0.82	0.83	0.83	0.83	0.95	0.04	0.16
	Polyp Normal		0.87	0.85	0.86	0.84	0.94	0.05	0.15
3	GERD	80.64	0.75	0.78	0.76	0.77	0.91	0.08	0.22
	GERD Normal		0.78	0.77	0.78	0.77	0.91	0.08	0.22
	Polyp		0.79	0.85	0.82	0.85	0.94	0.05	0.14
	Polyp Normal		0.90	0.83	0.86	0.83	0.96	0.03	0.16
4	GERD	81.95	0.77	0.78	0.77	0.77	0.92	0.07	0.22
	GERD Normal		0.77	0.84	0.80	0.84	0.90	0.09	0.15
	Polyp		0.84	0.85	0.85	0.85	0.95	0.04	0.14
	Polyp Normal		0.92	0.81	0.86	0.80	0.97	0.02	0.19
5	GERD	82.33	0.75	0.82	0.78	0.81	0.91	0.08	0.18
	GERD Normal		0.83	0.77	0.80	0.77	0.93	0.06	0.22
	Polyp		0.93	0.76	0.84	0.75	0.98	0.01	0.24
	Polyp Normal		0.82	0.92	0.87	0.92	0.91	0.08	0.07

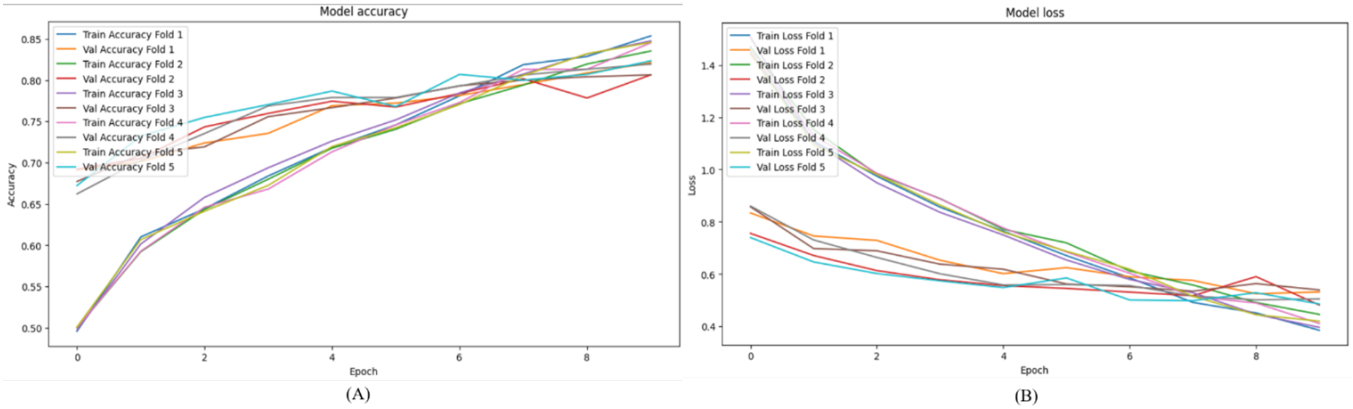

Figure 6. Fold-wise training and validation curves of Xception on Dataset 1. (A) Training vs. validation accuracy. (B) Training vs. validation loss

, predominantly in the 0.70–0.75 range. Training curves are smoother than their validation counterparts, with Fold 2 exhibiting notable oscillations indicative of learning instability. The loss curves in Figure 5(B) show a consistent downward trend in training loss; however, validation loss follows with greater fluctuation, particularly in Fold 2, suggesting susceptibility to overfitting in certain data splits. Despite these limitations, the model does converge across all folds, confirming that VGG19 is trainable on this task but is outperformed by architectures with more expressive feature extraction mechanisms.

Xception Table 4 presents the fold-wise results of the Xception model on Dataset 1. Xception outperforms both InceptionV3 and VGG19, achieving a mean accuracy of 81.53% and a best-fold accuracy of 82.33% in Fold 5. Its TPR of 0.80 and TNR of 0.93 reflect strong and balanced detection capability across both positive and negative classes. Notably, Xception achieves the lowest FPR (0.05) and FNR (0.18) among the three benchmark models, confirming its superior ability to minimize both false alarm rates and missed detections. Class-wise, Xception consistently produces well-balanced precision and recall values, particularly for the Polyp and Polyp Normal categories, demonstrating its effectiveness in capturing fine-grained morphological features through depthwise separable convolutions. The model's stable fold-wise performance further underscores its suitability for medical image classification tasks requiring simultaneous high sensitivity and specificity.

The accuracy curves in Figure 6(A) exhibit a sharp and consistent upward trajectory across all folds, converging to values between 0.80 and 0.85 by the final epoch. The curves are noticeably smoother and more compact than those of VGG19, indicating more stable learning dynamics with lower inter-fold variability. Fold 5 achieves the closest alignment between training and validation accuracy, reflecting minimal overfitting. The loss curves in Figure 6(B) confirms this pattern, with training and validation loss decreasing steadily across all folds and final validation losses stabilizing in the range of 0.40 to 0.55, a marked improvement over VGG19. Minor oscillations in Fold 2 around epoch 7 are present but do not indicate systemic instability, and the overall convergence behavior confirms Xception as the strongest benchmark model on this dataset.

4.2. Proposed Model Evaluation

Xnception on Dataset 1 The fold-wise performance of the proposed Xnception model on Dataset 1 is detailed in Table 5. Xnception achieves the highest mean accuracy of 85.96% and the best single-fold accuracy of 86.15% in Fold 5, surpassing all three benchmark models by a substantial margin. The model records the best TPR (0.84) and TNR (0.94) among all evaluated models, alongside the lowest FPR (0.04) and FNR (0.14), confirming that it not only detects diseased cases with high sensitivity but also minimizes misclassification of healthy samples. F1-scores consistently exceed 0.80 across all folds and all four classes, reflecting a well-calibrated balance between precision and recall. Notably, the Polyp Normal and GERD classes — which present the greatest inter-class visual ambiguity — achieve strong and stable performance across folds, validating the advantage of fusing the complementary feature extraction capabilities of Xception and InceptionV3 within the dual-backbone architecture. The training and validation accuracy curves in Figure 7(A) demonstrate a progressive and consistent improvement across all five folds, with training accuracy surpassing 95% by the tenth epoch, reflecting effective pattern learning on the training set. Validation accuracy improves steadily across folds, with Fold 3 and Fold 5 achieving validation accuracies closely aligned with their training counterparts, indicating strong generalization with minimal overfitting. The loss curves in Figure 7(B) confirm consistent convergence of training loss across all folds. Validation loss follows a broadly decreasing trend, though with moderate fold-wise variability after epoch 3, which is expected under cross-validation given the inherent distributional differences between validation splits. Overall, the learning dynamics confirm that Xnception generalizes effectively to unseen data on Dataset 1.

Xnception on Dataset 2 To assess the cross-dataset generalizability of the proposed Xnception model, it was further evaluated on the Kvasir dataset (Dataset 2) under the same 5-fold cross-validation protocol. The fold-wise results are presented in Table 6. The model achieves fold-wise accuracies ranging from 86.25% (Fold 2) to 90.00% (Fold 4), with Fold 4 representing the best overall training instance. Classes such as normal-cecum and esophagitis consistently achieve TPR values at or near 0.99 across all folds, reflecting near-perfect sensitivity for these categories. The dyed-lifted-polyps class presents the greatest classification challenge, with recall values ranging from 0.60 to 0.72 and moderate F1-scores between 0.70 and 0.79, likely attributable to inter-class visual similarity with dyed-resection-margins. Across all remaining classes, including ulcerative-colitis, polyps, and dyed-resection margins. The model achieves consistently high precision and recall, with FPR values below 0.04 and FNR values below 0.13 for the majority of classes, confirming low misclassification rates and strong diagnostic reliability.

The accuracy curves in Figure 8(A) show a progressive improvement in training accuracy across all five

Table 5. Fold-wise classification performance of Xnception on Dataset 1

Fold	Class	Acc.	Prec.	Rec.	F1	TPR	TNR	FPR	FNR
1	GERD	84.76	0.80	0.80	0.80	0.80	0.92	0.07	0.19
	GERD Normal		0.82	0.82	0.82	0.81	0.93	0.06	0.18
	Polyp		0.88	0.87	0.88	0.87	0.97	0.02	0.12
	Polyp Normal		0.90	0.90	0.90	0.90	0.95	0.04	0.09
2	GERD	84.59	0.87	0.75	0.81	0.75	0.96	0.03	0.24
	GERD Normal		0.80	0.87	0.83	0.86	0.91	0.08	0.13
	Polyp		0.83	0.84	0.83	0.83	0.95	0.04	0.16
	Polyp Normal		0.85	0.88	0.86	0.87	0.93	0.06	0.12
3	GERD	85.00	0.82	0.75	0.78	0.75	0.95	0.04	0.25
	GERD Normal		0.79	0.87	0.83	0.87	0.91	0.08	0.12
	Polyp		0.91	0.84	0.88	0.84	0.97	0.02	0.15
	Polyp Normal		0.89	0.92	0.90	0.91	0.95	0.04	0.08
4	GERD	86.00	0.91	0.77	0.84	0.77	0.97	0.02	0.22
	GERD Normal		0.82	0.91	0.87	0.91	0.92	0.07	0.08
	Polyp		0.89	0.80	0.84	0.79	0.97	0.02	0.20
	Polyp Normal		0.84	0.92	0.88	0.92	0.92	0.07	0.07
5	GERD	86.15	0.82	0.85	0.83	0.85	0.94	0.05	0.14
	GERD Normal		0.87	0.78	0.82	0.78	0.95	0.04	0.21
	Polyp		0.93	0.86	0.89	0.86	0.98	0.01	0.13
	Polyp Normal		0.85	0.94	0.89	0.94	0.93	0.06	0.05

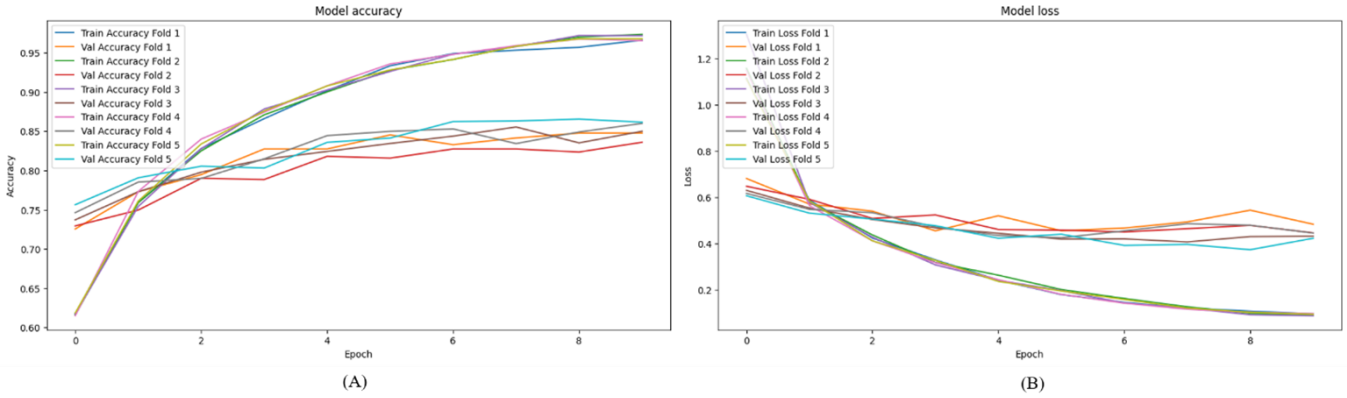

Figure 7. Fold-wise training and validation curves of Xnception on Dataset 1. (A) Training vs. validation accuracy. (B) Training vs. validation loss

folds, reaching approximately 90% by epoch 9. Validation accuracy stabilizes between 85% and 90% for most folds, indicating balanced and consistent generalization across data splits. The loss curves in Figure 8(B) confirm steady training loss reduction across all folds, with validation loss following a smooth and parallel decreasing trend exhibiting markedly lower oscillation compared to Dataset 1 results. The proximity of training and validation loss curves across folds confirms the stability and robustness of the model on this independent external dataset, further demonstrating the cross-dataset generalizability of the Xnception architecture.

4.3. Comparative Analysis and Explainability

Comparative Performance Summary Table 7 presents a consolidated summary of the average and best-fold performance of all four models on Dataset 1. Xnception outperforms all benchmark architectures across every reported metric. With a mean accuracy of 85.96% and a best-fold accuracy of 86.15% (Fold 5), Xnception surpasses Xception (81.53%), InceptionV3 (80.00%), and VGG19 (73.72%) by margins of 4.43, 5.96, and 12.24 percentage points, respectively. Its TPR of 0.84 and TNR of 0.94 represent the highest sensitivity and specificity a mong a ll m odels, w hile t he l owest FPR

Table 6. Fold-wise classification performance of Xnception on Dataset 2 (Kvasir)

Fold	Class	Acc.	Prec.	Rec.	F1	TPR	TNR	FPR	FNR
1	Dyed Lifted Polyps	86.72	0.88	0.70	0.77	0.69	0.98	0.01	0.30
	Normal Z-Line		0.73	0.90	0.80	0.90	0.95	0.04	0.09
	Dyed Resection Margins		0.91	0.72	0.80	0.71	0.98	0.01	0.28
	Normal Pylorus		0.88	0.96	0.92	0.96	0.98	0.01	0.03
	Normal Cecum		0.95	0.99	0.97	0.98	0.99	0.00	0.01
	Polyps		0.76	0.92	0.83	0.92	0.95	0.04	0.07
	Ulcerative Colitis		0.91	0.84	0.87	0.83	0.98	0.01	0.16
	Esophagitis		0.97	0.92	0.95	0.92	0.99	0.00	0.07
2	Dyed Lifted Polyps	86.25	0.76	0.71	0.73	0.70	0.96	0.03	0.29
	Normal Z-Line		0.77	0.79	0.78	0.79	0.96	0.03	0.20
	Dyed Resection Margins		0.89	0.77	0.83	0.77	0.98	0.01	0.22
	Normal Pylorus		0.92	0.96	0.94	0.95	0.98	0.01	0.04
	Normal Cecum		0.94	0.99	0.96	0.98	0.99	0.00	0.01
	Polyps		0.78	0.89	0.83	0.88	0.96	0.03	0.11
	Ulcerative Colitis		0.94	0.89	0.91	0.89	0.99	0.00	0.10
	Esophagitis		0.92	0.94	0.93	0.94	0.98	0.01	0.05
3	Dyed Lifted Polyps	88.75	0.90	0.70	0.78	0.67	0.98	0.01	0.32
	Normal Z-Line		0.74	0.91	0.82	0.90	0.95	0.04	0.09
	Dyed Resection Margins		0.92	0.85	0.88	0.85	0.98	0.01	0.14
	Normal Pylorus		0.83	1.00	0.91	0.99	0.96	0.03	0.00
	Normal Cecum		0.97	0.99	0.98	0.98	0.99	0.00	0.01
	Polyps		0.87	0.93	0.90	0.92	0.97	0.02	0.07
	Ulcerative Colitis		0.92	0.80	0.86	0.80	0.99	0.00	0.02
	Esophagitis		1.00	0.96	0.98	0.96	0.99	0.00	0.03
4	Dyed Lifted Polyps	90.00	0.88	0.72	0.79	0.72	0.98	0.01	0.28
	Normal Z-Line		0.77	0.91	0.83	0.90	0.96	0.03	0.09
	Dyed Resection Margins		0.91	0.85	0.88	0.85	0.98	0.01	0.15
	Normal Pylorus		0.96	0.95	0.96	0.95	0.99	0.00	0.04
	Normal Cecum		0.96	0.98	0.97	0.97	0.99	0.00	0.02
	Polyps		0.82	0.85	0.83	0.85	0.97	0.02	0.14
	Ulcerative Colitis		0.96	0.90	0.93	0.90	0.99	0.00	0.09
	Esophagitis		0.90	0.97	0.93	0.96	0.98	0.01	0.03
5	Dyed Lifted Polyps	87.03	0.88	0.60	0.70	0.58	0.98	0.01	0.41
	Normal Z-Line		0.75	0.93	0.83	0.92	0.95	0.04	0.07
	Dyed Resection Margins		0.86	0.83	0.85	0.82	0.98	0.01	0.17
	Normal Pylorus		0.90	0.95	0.92	0.95	0.98	0.01	0.04
	Normal Cecum		0.96	0.96	0.96	0.96	0.99	0.00	0.03
	Polyps		0.85	0.90	0.87	0.89	0.97	0.02	0.10
	Ulcerative Colitis		0.92	0.87	0.89	0.86	0.98	0.01	0.13
	Esophagitis		0.86	0.93	0.90	0.92	0.98	0.01	0.07

of 0.04 and FNR of 0.14 confirm its superior ability to Simultaneously minimize false alarms and missed detections. These results demonstrate that the fusion of Xception and InceptionV3 feature streams in the dual-backbone architecture yields measurably stronger and more robust classification performance than any single-backbone alternative on the GastroEndoNet dataset.

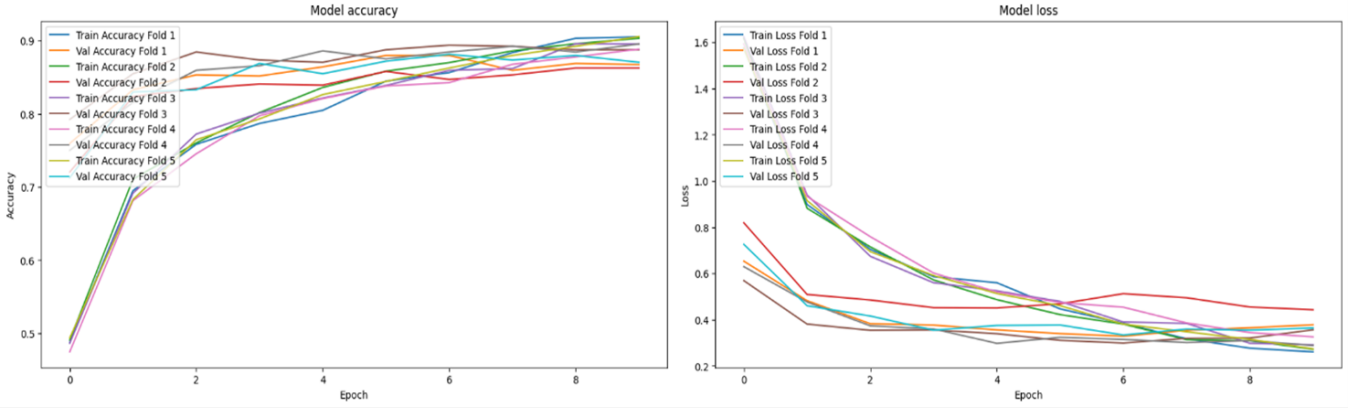

Figure 8. Fold-wise training and validation curves of Xnception on Dataset 2 (Kvasir). (A) Training vs. validation accuracy. (B) Training vs. validation loss

Table 7. Average and best-fold performance comparison of all models on Dataset 1

Model	Avg. Acc. (%)	TPR	TNR	FPR	FNR	Best Fold	Best Acc. (%)
InceptionV3	80.00	0.76	0.92	0.06	0.20	Fold 4	82.02
VGG19	73.72	0.72	0.90	0.08	0.26	Fold 3	74.42
Xception	81.53	0.80	0.93	0.05	0.18	Fold 5	82.33
Xnception (Ours)	**85.96**	**0.84**	**0.94**	**0.04**	**0.14**	**Fold 5**	**86.15**

Table 8. Comparative summary with recent state-of-the-art methods

Ref.	Year	Methodology	Dataset	Accuracy	Key Strength
[23]	2025	MSFF with CMA and CIR modules	Kvasir	88.16%	Robust generalization across multiple endoscopic datasets
[24]	2025	FRAN with dual-branch residual attention learning	Private four-class polyp dataset	85.73%	Effective detection of critical regions and edge features
This Study	2025	Hybrid dual-backbone CNN (Xnception) with LIME, Grad-CAM++, and IG	GastroEndoNet (Dataset 1) and Kvasir (Dataset 2)	86.15% / 90.00%	Dual-dataset validation with integrated Explainable AI

Table 8 situates the proposed model within the context of recent state-of-the-art methods for gastrointestinal endoscopic image classification. The Multi-Stage Feature Fusion Network (MSFF) [23], published in 2025, achieved 88.16% accuracy on the Kvasir dataset by integrating Context Modulation Attention (CMA) with a Cross Interaction Residual (CIR) module to enhance discriminative feature representation. While the method demonstrates strong generalizability across multiple endoscopic datasets, it does not surpass the 90% threshold, suggesting residual limitations in capturing fine-grained features. The Fused Residual Attention Network (FRAN) [24], also published in 2025, employed a dual-branch structure with residual attention learning, achieving 85.73% accuracy on a private four-class polyp dataset. In contrast, the proposed Xnception model achieves 86.15% on Dataset 1 and 90.00% on Dataset 2, positioning it competitively with these recent developments. A further distinguishing contribution of this work is the integration of three complementary Explainable AI (XAI) techniques, including LIME, Grad-CAM++, and Integrated

Gradients (IG), which substantially enhance the interpretability and clinical transparency of model predictions, a dimension not addressed by the compared methods.

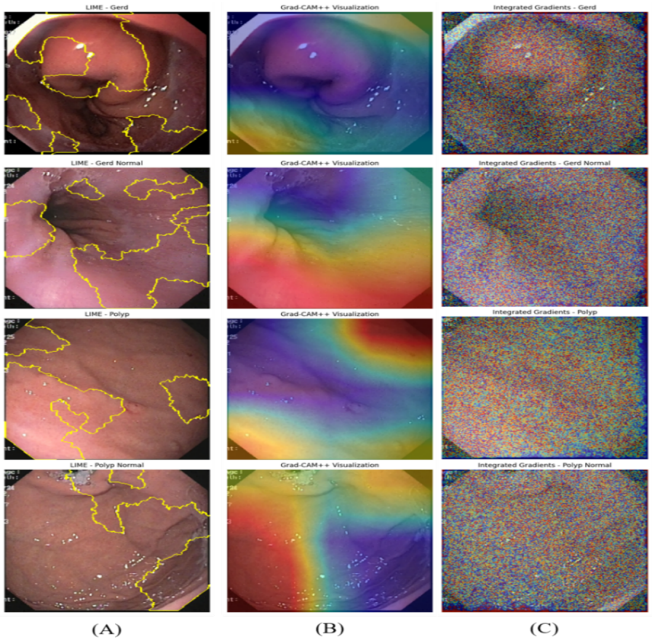

Figure 9. Explainable AI visualizations of the proposed Xception model on Dataset 1 across four classes (GERD, GERD Normal, Polyp, Polyp Normal). (A) LIME superpixel importance maps. (B) Grad-CAM++ class activation heatmaps. (C) Integrated Gradients pixel-level attribution maps

Explainability Analysis To enhance the clinical transparency and interpretability of the proposed Xnception model, three complementary Explainable AI (XAI) techniques were applied: Local Interpretable Model-Agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping++ (Grad-CAM++), and Integrated Gradients (IG). These methods operate at different levels of granularity and collectively form a comprehensive interpretability framework for understanding model decision-making.

Figure 9 presents the XAI visualizations for Dataset 1 across all four diagnostic classes: GERD, GERD Normal, Polyp, and Polyp Normal. In the LIME visualizations, superpixel boundaries delineate the image regions that most strongly influenced the model's predictions, with diseased classes such as GERD and Polyp exhibiting focused attention on inflamed or morphologically abnormal mucosal regions. Grad-CAM++ heatmaps highlight class-discriminative anatomical regions with strong spatial precision, particularly in diseased cases, whereas normal classes display more diffuse activation patterns consistent with the absence of localized pathology. Integrated Gradients visualizations provide dense pixel-level attribution maps that corroborate the attention patterns identified by LIME and Grad-CAM++, confirming that the model consistently focuses on clinically meaningful visual features across all classes. The concordance across all three XAI methods confirms that Xnception has learned disease-specific feature representations rather than spurious correlates, thereby supporting its diagnostic credibility and potential utility in clinical decision-support workflows.

Figure 10 extends this analysis to Dataset 2, presenting XAI visualizations for selected gastrointestinal disease classes from the Kvasir dataset, including dyed-lifted polyps, esophagitis, and normal z-line. LIME superpixel maps effectively identify textural and structural features relevant to classification, though with reduced precision in visually complex classes such as ulcerative colitis. Grad-CAM++ produces anatomically coherent activation maps that align well with regions of pathological significance, particularly in lesion and inflamed tissue cases. Integrated Gradients maps, while denser and more fine-grained, provide high-fidelity pixel-level attribution that proves particularly informative in borderline or diagnostically ambiguous cases. The complementary nature of the three XAI techniques, local superpixel-level, spatially coherent class-activation, and pixel-wise gradient-based, collectively confirms that the Xception model attends to semantically and clinically relevant regions across the diverse pathological categories of Dataset 2, further validating its generalizability and interpretability beyond the primary training domain.

While LIME, Grad-CAM++, and Integrated Gradients provide qualitative interpretability insights, the current evaluation does not incorporate quantitative explainability metrics. Future work will address this limitation by integrating faithfulness-based evaluation measures such as insertion/deletion scores and localization accuracy against expert-annotated regions, providing a more rigorous and objective validation of model interpretability. Furthermore, the proposed framework is designed with clinical deployment in mind: by leveraging widely adopted transfer learning backbones and modular explainability techniques, Xnception can be integrated into existing endoscopic diagnostic pipelines with minimal architectural modification, serving as an interpretable assistive tool for gastrointestinal disease screening where both predictive performance and decision transparency are essential.

5. Conclusion

This study presented Xnception, a dual-stream deep learning framework that integrates the complementary feature extraction capabilities of Xception and InceptionV3 for GERD and gastric polyp classification from endoscopic images. Extensive fold-wise evaluations demonstrated that the proposed architecture

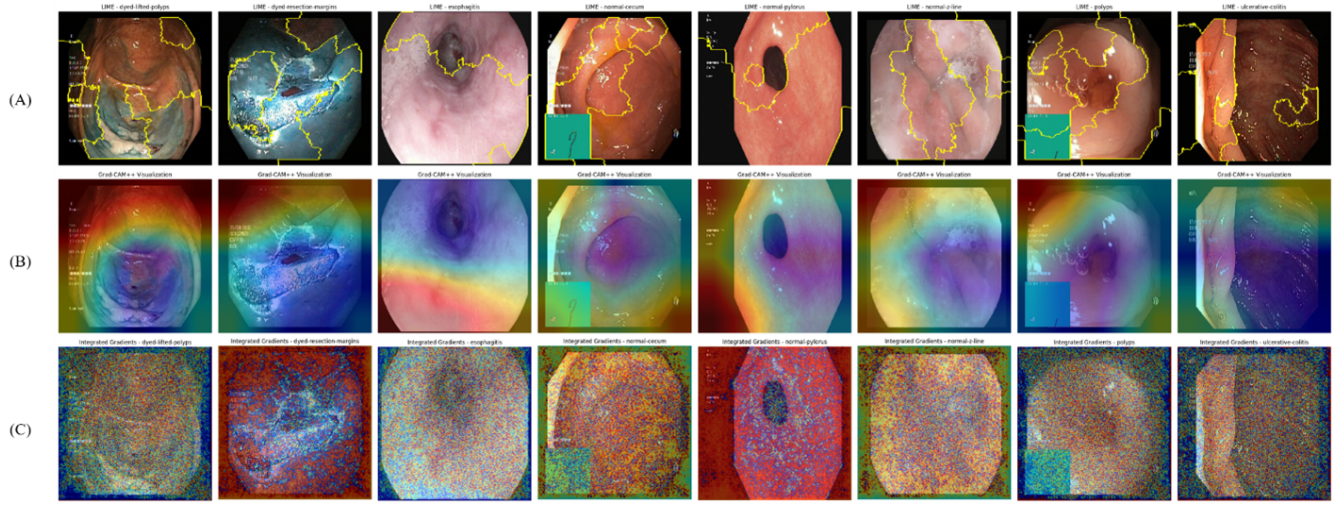

Figure 10. Explainable AI visualizations of the proposed Xnception model on Dataset 2 (Kvasir) across selected gastrointestinal disease classes. (A) LIME superpixel importance maps. (B) Grad-CAM++ class activation heatmaps. (C) Integrated Gradients pixel-level attribution maps

consistently outperformed individual backbone models across multiple performance metrics, highlighting the effectiveness of architectural fusion and transfer learning in constrained medical imaging scenarios. In addition, the incorporation of explainability techniques, including LIME, Integrated Gradients (IG), and Grad-CAM++, improved the interpretability of the framework by providing clinically meaningful visual explanations for model predictions. The experimental findings indicate that Xnception can serve as both a high-performance classification framework and an explainable computer-aided diagnostic tool for gastroenterological analysis. The proposed system further demonstrates the growing potential of intelligent medical imaging systems to reduce diagnostic burden and support decision-making in resource-constrained and high-demand clinical environments. Future work will focus on improving the representational capacity and computational efficiency of the framework through the integration of attention mechanisms and transformer-based modules for enhanced contextual understanding. Comparative evaluations against recent architectures, including Vision Transformers (ViT), Swin Transformers, and hybrid CNN-transformer models, will also be conducted to assess scalability across diverse medical imaging settings. Additionally, domain generalization strategies such as self-supervised learning and few-shot learning will be explored to improve robustness across varying imaging devices and acquisition conditions. Finally, the proposed framework will be validated on larger multi-center and multi-device endoscopic datasets to further strengthen its clinical applicability and real-world deployment potential.

References

[1] Milivojevic V. and Milosavljevic T. (2020) *Burden of gastroduodenal diseases from the global perspective.* Current Treatment Options in Gastroenterology, 18, 148-157.

[2] Sharma N. and Srivastava S. (2025) *Transforming Pancreatic Cancer Diagnosis: Conventional Methods, Challenges, and Future Innovations.* Exon, 2(2), 86-111.

[3] Jain R. and Chetty R. (2009) *Gastric hyperplastic polyps: a review.* Digestive diseases and sciences, 54, 1839-1846.

[4] Wang F.W., Young S.C. et al. (2018) *The prevalence and risk factors of gastric polyp in asymptomatic patients receiving health examination.* Gastroenterology research and practice, 2018(1), 9451905.

[5] Wong B.C. and Kinoshita Y. (2006) *Systematic review on epidemiology of gastroesophageal reflux disease in Asia.* Clinical gastroenterology and hepatology, 4(4), 398-407.

[6] Cortegoso Valdivia P., Deding U. et al. (2022) *Inter/intra-observer agreement in video-capsule endoscopy: are we getting it all wrong? A systematic review and meta-analysis.* Diagnostics, 12(10), 2400.

[7] Badruzzaman Biplob K.B., Sammak M.H. et al. (2024) *COVID-19 and Suicide Tendency: Prediction and Risk Factor Analysis Using Machine Learning and Explainable AI.* EAI Endorsed Transactions on Pervasive Health & Technology, 10(1).

[8] Ringky N.A., Bitto A.K. et al. (2024) *Enhancing Skin Disease Diagnosis Through Fine-Tune Convolutional Neural Network: A Comparative Study with Multi-class Approach.* Journal of ICT Research & Applications, 18(2).

[9] Li X., Zhang L., Yang J. and Teng F. (2024) *Role of artificial intelligence in medical image analysis: A review of current trends and future directions.* Journal of Medical and Biological Engineering, 44(2), 231-243.

[10] Ozawa T., Ishihara S. et al. (2020) *Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks.* Therapeutic advances in gastroenterology, 13, 1756284820910659.

[11] Bitto A.K. and Mahmud I. (2022) *Multi categorical of common eye disease detect using convolutional neural network: a transfer learning approach.* Bulletin of Electrical Engineering and Informatics, 11(4), 2378-2387.

[12] Urban G., Tripathi P. et al. (2018) *Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy.* Gastroenterology, 155(4), 1069-1078.

[13] Kavitha M.S., Gangadaran P. et al. (2022) *Deep neural network models for colon cancer screening.* Cancers, 14(15), 3707.

[14] Wang P., Berzin T.M. et al. (2019) *Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study.* Gut, 68(10), 1813-1819.

[15] Wong M.W., Hung J.S. et al. (2023) *Esophageal secondary peristalsis following acid infusion and chemical clearance correlate with mucosal integrity and acid sensitivity in GERD patients.* Therapeutic Advances in Gastroenterology, 16, 17562848231179329.

[16] Aziz M., Haghbin H. et al. (2020) *Gastrointestinal predictors of severe COVID-19: systematic review and meta-analysis.* Annals of Gastroenterology, 33(6), 615.

[17] Souaidi M. and El Ansari M. (2022) *A new automated polyp detection network MP-FSSD in WCE and colonoscopy images based fusion single shot multibox detector and transfer learning.* IEEE access, 10, 47124-47140.

[18] Belabbes M.A., Oukdach Y. et al. (2024) *Advancements in polyp detection: a developed single shot multibox detector approach.* IEEE Access, 12, 19199-19215.

[19] Shin Y., Qadir H.A. et al. (2018) *Automatic colon polyp detection using region based deep CNN and post learning approaches.* IEEE access, 6, 40950-40962.

[20] Mohammed Jasim M.J., Hussan B.K. et al. (2023) *Automated Colonic Polyp Detection and Classification Enabled Northern Goshawk Optimization with Deep Learning.* Computers, Materials & Continua, 75(2).

[21] Charmet F., Tanuwidjaja H.C. et al. (2022) *Explainable artificial intelligence for cybersecurity: a literature survey.* Annals of Telecommunications, 77(11), 789-812.

[22] Moreno-Sánchez P.A. (2023) *Data-driven early diagnosis of chronic kidney disease: development and evaluation of an explainable AI model.* IEEE Access, 11, 38359-38369.

[23] Zhang R., Luo X. et al. (2025) *Enhancing Medical Image Classification with Context Modulated Attention and Multi-scale Feature Fusion.* IEEE Access.

[24] Li S., Guo X. et al. (2025) *Multi-classification of colorectal polyps with fused residual attention.* Signal, Image and Video Processing, 19(1), 144.

[25] Bitto A.K., Bijoy M.H.I. et al. (2025) *GastroEndoNet: Comprehensive Endoscopy Image Dataset for GERD and Polyp Detection.* Data in Brief, 111572.

[26] Pogorelov K., Randel K.R. et al. (2017, June) *Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection.* In Proceedings of the 8th ACM on Multimedia Systems Conference (pp. 164-169).

[27] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[28] Bansal, Monika, et al. "Transfer learning for image classification using VGG19: Caltech-101 image data set." Journal of ambient intelligence and humanized computing 14.4 (2023): 3609-3620.

[29] Wang, Cheng, et al. "Pulmonary image classification based on inception-v3 transfer learning model." IEEE Access 7 (2019): 146533-146541.

[30] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[31] Jahid Hassan Akash, Mia R, Bitto A.K. et al. (2026) *A Stacking Based Ensemble Learning Approach for Accurate Identification of Tumor Homing Peptides in Precision Cancer Therapeutics.* EAI Endorsed Trans AI Robotics, 5. Available from: https://publications.eai.eu/index.php/airo/article/view/10265

[32] Mia R, Hasan T, Bitto A.K. et al. (2025) *Enhancing the Prediction of IL-4 Inducing Peptides Using Stacking Ensemble Model.* EAI Endorsed Trans AI Robotics, 4. Available from: https://publications.eai.eu/index.php/airo/article/view/9867

[33] Abu Kowshir Bitto, Rezwana Karim, Begum M.H. et al. (2025) *Explainable AI Based Deep Ensemble Convolutional Learning for Multi-Categorical Ocular Disease Prediction.* EAI Endorsed Trans AI Robotics, 4. Available from: https://publications.eai.eu/index.php/airo/article/view/9234