EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

An Explainable AI Based Deep Ensemble Transformer Framework for Gastrointestinal Disease Prediction from Endoscopic Images

Abdul Kadar Muhammad Masum[1*], Abu Kowshir Bitto[2], Shafiqul Islam Talukder[3], Md Fokrul Islam Khan[4], Mohammed Shamsul Alam[5], Khandaker Mohammad Mohi Uddin[1]

[1]Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh
[2]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[3]Department of Computer Science, Westcliff University, Los Angeles, United States
[4]Department of Management Information Systems, International American University, Los Angeles, United States
[5]Department of Computer Science and Engineering, International Islamic University Chittagong, Chittagong, Bangladesh

Abstract

Gastrointestinal diseases such as gastroesophageal reflux disease (GERD) and polyps remain prevalent and challenging to diagnose accurately due to overlapping visual features and inconsistent endoscopic image quality. In this study, we investigate the application of transformer-based deep learning models—Vision Transformer (ViT), Swin Transformer, and a novel Ensemble Transformer model—for classifying four categories: GERD, GERD Normal, Polyp, and Polyp Normal from endoscopic images. The dataset was curated and collected in collaboration with Zainul Haque Sikder Women's Medical College & Hospital, ensuring high-quality clinical annotations. All models were evaluated using precision, recall, F1 score, and overall classification accuracy. Our proposed Ensemble Transformer model, which fuses the outputs of ViT and Swin Transformer, achieved superior performance by delivering well-balanced F1 scores across all classes, reducing misclassification, and improving robustness with an overall accuracy of 87%. Furthermore, we incorporated explainable AI (XAI) techniques such as Grad-CAM and Grad-CAM++ to generate visual explanations of the model's predictions, enhancing interpretability for clinical validation. This work demonstrates the potential of integrating global and local attention mechanisms along with XAI in building reliable, real-time, AI-assisted diagnostic support systems for gastrointestinal disorders, particularly in resource-limited healthcare settings.

Keywords: Gastrointestinal Disease, Medical Image Processing, Transformer Models, Ensemble Model, Explainable AI

Received on 25 July 2025, accepted on 17 August 2025, published on 25 August 2025

doi: 10.4108/airo.9795

1. Introduction

Gastrointestinal (GI) disorders remain a widespread public health concern, causing substantial morbidity and placing immense strain on healthcare systems across both developed and developing nations [1]. Among these, Gastroesophageal Reflux Disease (GERD) and gastrointestinal polyps are two frequently encountered conditions that require timely detection to avoid escalation into life-threatening diseases. GERD is characterized by the chronic reflux of gastric contents into the esophagus, leading to mucosal damage and long-term complications such as esophageal strictures, Barrett's esophagus, and even esophageal adenocarcinoma. It affects an estimated 1 billion people globally, making it one of the most prevalent noncommunicable gastrointestinal disorders [2]. Gastrointestinal polyps especially adenomatous polyps pose significant oncogenic risks and are directly associated with colorectal cancer, which is projected to cause over 1 million deaths annually by 2030 if current trends persist [3].

Globally, colorectal cancer ranks as the third most common cancer and the second leading cause of cancer related deaths, with over 1.9 million new cases diagnosed in 2022 [4]. GERD affects over billion people worldwide, often progressing silently without early intervention [5]. As the demand for accurate and automated endoscopic diagnosis

grows, transformer-based models hold potential to revolutionize GI diagnostics by using their ability to model long range spatial dependencies.

With the development of deep learning in medical image processing, Convolutional Neural Networks (CNNs) have performed very well in computer aided diagnosis systems [6] [7]. Vision Transformer and Swin Transformer are renowned models for feature extraction and have been widely employed in radiological and pathological fields [8] [9]. However, comparative studies regarding their performance in gastrointestinal image classification specifically for GERD and detection of polyps are limited.

In this work, we evaluate the diagnostic performance of Vision Transformer (ViT) and SWIN Transformer models using a specialized dataset comprising 4,006 endoscopic images classified into GERD, GERD Normal, Polyp, and Polyp Normal. Additionally, we construct a novel ensemble of both models to enhance diagnostic precision. To improve clinical interpretability and transparency, we employ explainable AI (XAI) techniques such as Grad-CAM and Grad-CAM++ to visualize the decision-making regions within the endoscopic images, allowing validation of model predictions by medical professionals. This study seeks to bridge the gap between emerging deep transformer models and clinical gastroenterology, offering insights into their adaptability, robustness, and explainability in real-world diagnostic scenarios.

In this manner, the remainder of the study is in order. The section 2 of the paper is a review of the literature. The approach for detection disease is discussed in Section 3. The analysis and results are demonstrated in Section 4. Section 5 of the document, certainly, brings the whole thing to a conclusion.

2. Related Work

The detection of gastroesophageal reflux disease (GERD) and colorectal polyps using artificial intelligence (AI), particularly deep learning models, has emerged as a critical area of research aimed at enhancing diagnostic accuracy and improving patient outcomes [10] - [13]. This pursuit is rooted in the necessity to effectively identify precancerous conditions and other gastrointestinal disorders, which are significant public health concerns.

Deep learning, particularly through convolutional neural networks (CNNs), has advanced polyp detection during colonoscopy [14]. These models can process and analyze vast amounts of imaging data, achieving high accuracy rates in identifying polyps. Quan et al. [15] demonstrated that their AI based system significantly enhanced adenoma detection rates in clinical settings, thereby reducing the risk of colorectal cancer resulting from undetected lesions. Wang et al. [16] reported substantial improvements in polyp detection rates through real time, automatic detection systems integrated into standard colonoscopy practices. Such systems assist in identifying polyps and play a role in characterizing lesions to determine the appropriate treatment strategy, which is vital for effective patient management. Recent studies have also highlighted the importance of optimizing these deep learning models. Al Otaibi et al. [17] introduced an optimized EfficientNet model specifically designed for detecting gastrointestinal disorders through analysis of wireless capsule endoscopy images. This approach represents a shift toward utilizing specialized models tailored to specific clinical needs, enhancing detection capabilities.

The use of regression based CNNs has shown promising results in polyp detection, indicating these models' potential to not only identify polyps but also assess their characteristics and inform clinical management [18]. Wesp et al. [19] demonstrated that deep learning models could differentiate between benign and premalignant colorectal polyps, crucial for making informed clinical decisions and personalized patient care. GPU based implementations of deep learning frameworks have been explored to improve computational efficiency and operational speed during real time procedures, which is essential in busy clinical environments [20].

The effectiveness of deep learning models, however, is often contingent upon the quality and comprehensiveness of the training datasets. Many studies have cited the challenge of limited labelled data, which can impede the training of robust models [21]. To address this, techniques such as transfer learning, transformers and data augmentation have been employed to enhance model performance without requiring extensive new datasets [22] [23].

The integration of deep learning approaches in the detection of GERD and polyps represents a significant advance in gastroenterology, offering tools that improve diagnostic accuracy, optimize treatment strategies, and ultimately enhance patient care [24]. The continuous evolution of these technologies, marked by the deployment of specialized models and innovative datasets, promises to play a pivotal role in the future of gastrointestinal diagnosis and management. Furthermore, the incorporation of explainable AI (XAI) techniques—such as Grad-CAM and Grad-CAM++—enhances the transparency and interpretability of model decisions, allowing clinicians to visually verify AI-predicted regions of concern, thereby fostering trust, and aiding in informed clinical decision-making [25] [26].

3. Methodology

This study proposes an ensemble methodology founded on transformers for precise classification of gastrointestinal endoscopic images. The process starts with the accumulation of labelled gastric polyp and Z line endoscopic images. The images are subjected to intensive preprocessing resizing, cropping, normalization, and extensive augmentation to meet the transformer models' input requirements. These pre-processed images are passed in parallel to two state-of-the-art vision transformer models named Vision Transformer (ViT) and Swin Transformer. Each model captures fine-grained visual dependencies and semantic relationships, which are then fused using a soft voting ensemble mechanism. Final classification maps images to one of four classes based on pathological status. Model performance is evaluated using performance metrics such as confusion matrices and AUC

scores. Finally, XAI techniques such as GradCAM and GradCAM++ improve model interpretability. Figure 1 shows the overall system flow.

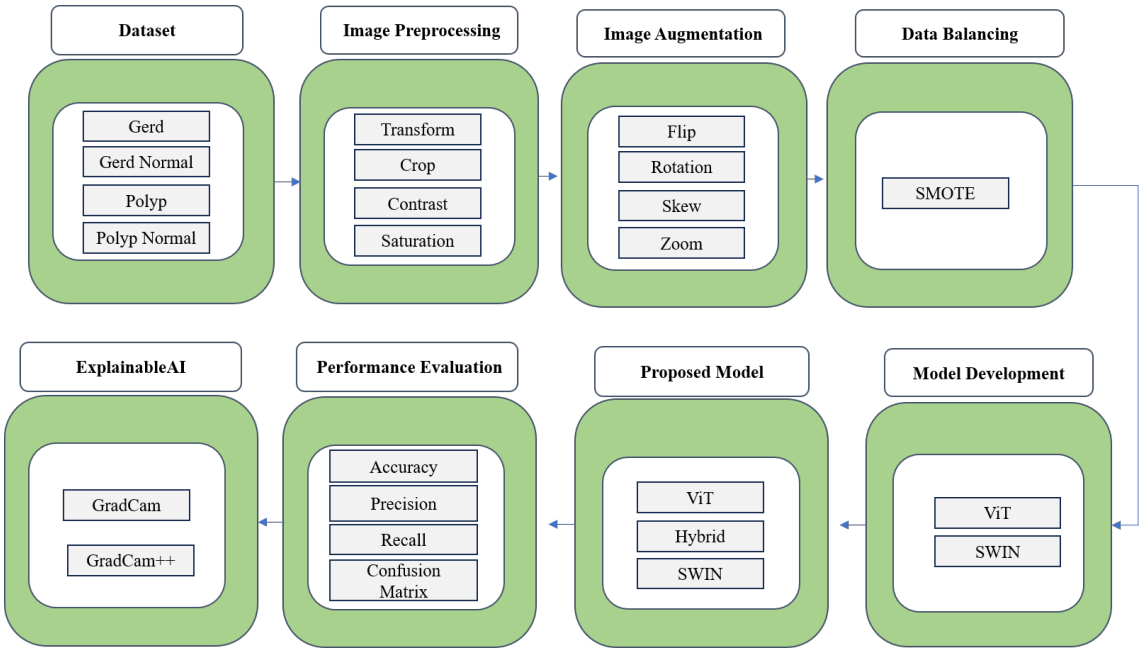

Figure 1. Step by Step Procedure Diagram of Workflow

3.1. Dataset

It needs the availability of large and robust datasets to train precise and trustworthy deep learning models. This is a significant challenge because medical data is sensitive in nature and privacy concerns are involved in its collection and sharing. To address this challenge, the dataset used in this study was obtained from Zainul Haque Sikder Women's Medical College & Hospital (Pvt.) Ltd., prepared under the supervision of Dr. Md. Humayun Kabir, Assistant Registrar, Department of Surgical Gastroenterology. The dataset has 4,006 original high resolution endoscopic images, distributed over four categories: GERD, GERD Normal, Polyp, and Polyp Normal, which is available now on Mendeley and published in data in brief [27]. Figure 2 shows our dataset sample images.

3.2. Data Preprocessing

Preprocessing is a fundamental component of any deep learning pipeline, especially for transformer models like Vision Transformer (ViT) and Swin Transformer. They require some data preprocessing to function optimally, specifically resolution and augmentation policy. The first step in the preprocessing pipeline was resizing images to 460x460 pixels and then cropping them to 384x384 pixels to accommodate the input for ViT and Swin Transformer. Resizing the high-resolution images served to preserve important features that would assist in the classification task. For generalization improvement, we utilized several augmentation techniques such as random rotation, flipping, zooming, and random brightness change. We also augmented each image by applying shearing and scaling transformation to create a more robust dataset. Normalization was done using the pretrained weights' mean and standard deviation of ImageNet for both Swin and ViT models. This aligned the input distribution with the expectations of pretrained transformers. Both models were designed to accept images in RGB color format. The data was thus well suited to match the expected input specs of each model. In the interest of precise performance estimation, the data was divided into 80% training, 10% validation, and 10% test. The validation set was instrumental in model tuning, while the test set was reserved for final testing. Data was organized in minibatches of 32 images to enable efficient model training and leverage hardware acceleration, i.e., utilization of GPU. Class imbalance was dealt with by oversampling minority classes with SMOTE such that all classes had approximately the same number of samples during training. This led to improved model performance, especially in detecting minority classes like GERD and Polyp. Figure 3 illustrates the class distribution before and after applying the SMOTE technique, highlighting how data imbalance was addressed.

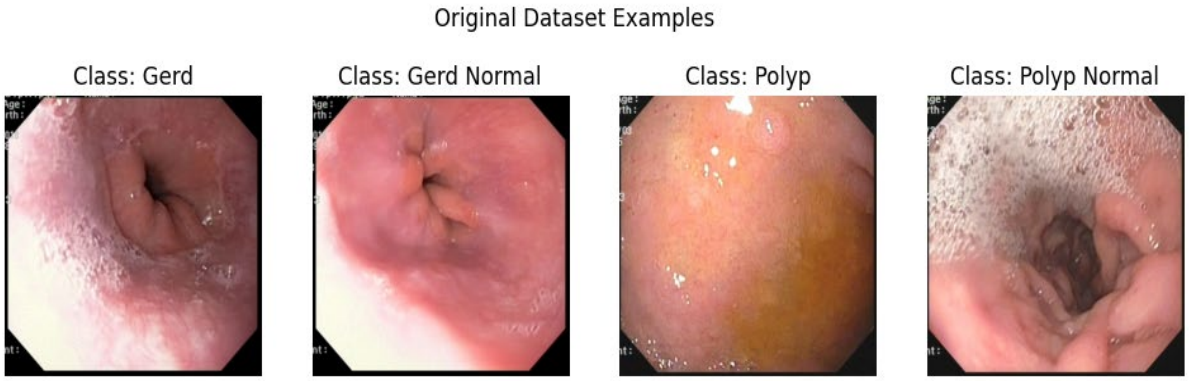

Figure 2. Our Dataset Sample Images

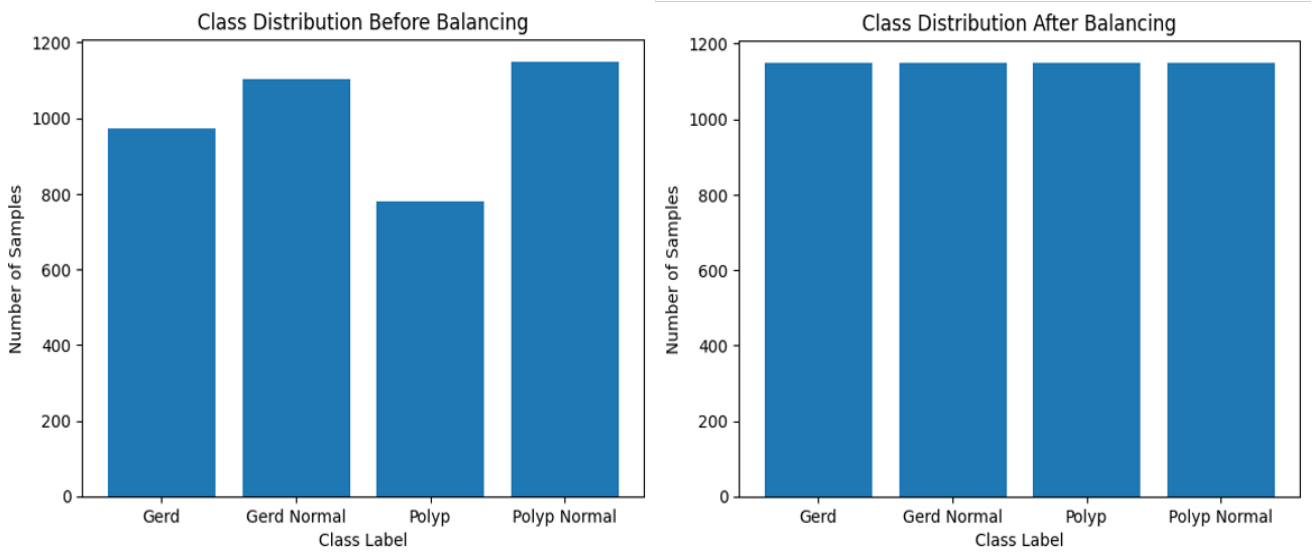

Figure 3. Class Distribution Before and After Balancing

3.3. Proposed Ensemble Model

This paper introduces a two-transformer ensemble model that combines the global attention capability of the Vision Transformer (ViT) and the hierarchical representation capacity of the Swin Transformer for gastrointestinal image classification. The two models were pretrained on largescale image datasets and finetuned on our domain specific endoscopic images. The two transformers' outputs were combined with a weighted soft voting ensemble both local and global contextual information. The ensemble probability P Ensemble shows in:

$$P_{\text{ensemβle}} = \beta \times PViT + (1 - \beta) \times PSwin \quad (1)$$

β is the fusion weight, set to 0.5 in this study to equally prioritize both models. The transformer ensemble further improves the ability of the system to recognize slight visual patterns from medical images for accurate classification of GI disorders. Proposed model architecture diagram shows in figure 4.

3.4. Performance Evaluation

We evaluated performance using various metrics to identify the most effective classifier for detecting eye diseases. Performance indicators, expressed as percentages (%), were calculated using Eqs. (2–5). Additionally, we generated a confusion matrix for each model to assess their performance comprehensively.

$$Accuracy = \frac{True\ Positive + True\ Negative}{All\ Instance} \quad (2)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

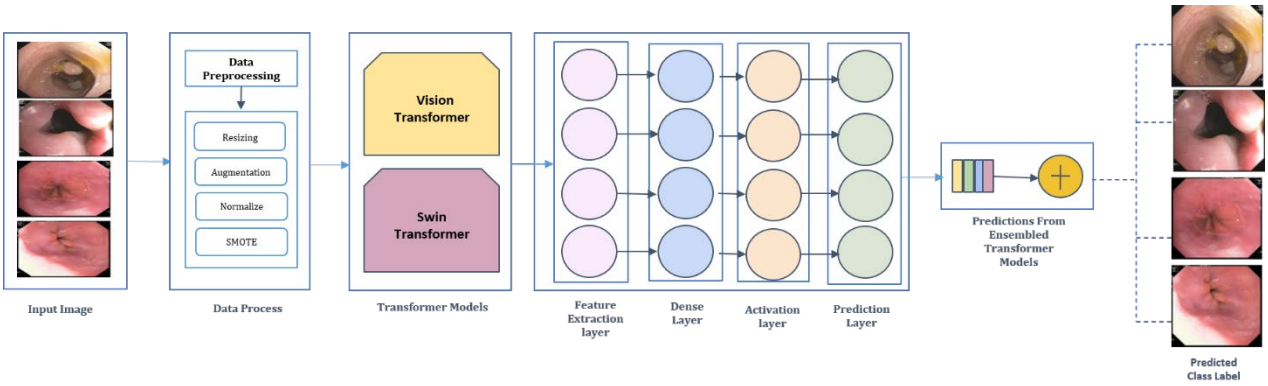

Figure 4. Proposed Ensemble Model Architecture Diagram.

4. Result and Discussion

We introduced a strong transformer based deep learning framework for gastric condition classification from endoscopy images. We contrasted the performance of state-of-the-art transformer-based models such as Vision Transformer (ViT), Swin Transformer, and Ensemble Transformer to classify gastric diseases from endoscopic images. The performance of each model was measured on performance metrics like precision, recall, F1score, and accuracy. All the model run with 20 epoch.

From Table 1, can see that the Swin Transformer demonstrated strong classification ability, particularly for Polyp detection, with a significant F1 score of 0.94. It also performed well in Polyp Normal case detection with an F1 score of 0.88, demonstrating its strong modelling of local and global context information. The GERD classification task was unbalanced in some way, while Precision was excellent (0.95), Recall was quite low (0.69), indicating a bias towards missing some GERD positive instances. The GERD Normal classification was reversed, with higher Recall but lower Precision, which means that the model was predicting Normal more often in ambiguous cases. Accuracy was 85% overall, which indicates Swin's strong performance on classes.

Swin Transformer confusion matrix in figure 5 (A) indicates that it correctly diagnosed 94 instances of GERD but incorrectly diagnosed 40 as GERD Normal, indicating a challenge in distinguishing between pathological and near normal presentations. It correctly classified 124 instances of Polyp with an extremely low level of misclassification, highlighting its ability to detect anomalies. It also correctly labelled 84 instances of Polyp Normal with a low amount of overlap into the Polyp class. These results indicate Swin's recognition of visual features concerned with Polyps but poor discrimination of features between GERD Normal and GERD.

From the training and validation curves in figure 6 (A), the Swin Transformer model demonstrated great learning stability. The training accuracy rose rapidly and plateaued at around 99%, whereas the validation accuracy oscillated between 85–91%, indicating moderate generalization. The training loss gradually decreased and converged to a near zero minimum, indicating successful optimization. In contrast, validation loss oscillated considerably, perhaps due to ambiguous visual features, especially between GERD and GERD Normal.

Table 1. Performance Calculation of Class-wise Precision, Recall, F1-score, and Overall Accuracy for Individual and Ensemble Models

Models	Class	Precision	Recall	F1-score	Accuracy
SWIN	Gerd	0.95	0.69	0.80	85%
	Gerd Normal	0.67	0.94	0.78	
	Polyp	0.93	0.95	0.94	
	Polyp Normal	0.91	0.86	0.88	
Vision	Gerd	0.88	0.74	0.80	80%
	Gerd Normal	0.66	0.88	0.75	
	Polyp	0.97	0.74	0.84	
	Polyp Normal	0.74	0.91	0.82	
Ensemble	Gerd	0.90	0.84	0.87	87%
	Gerd Normal	0.80	0.84	0.82	
	Polyp	0.93	0.91	0.92	
	Polyp Normal	0.82	0.89	0.85	

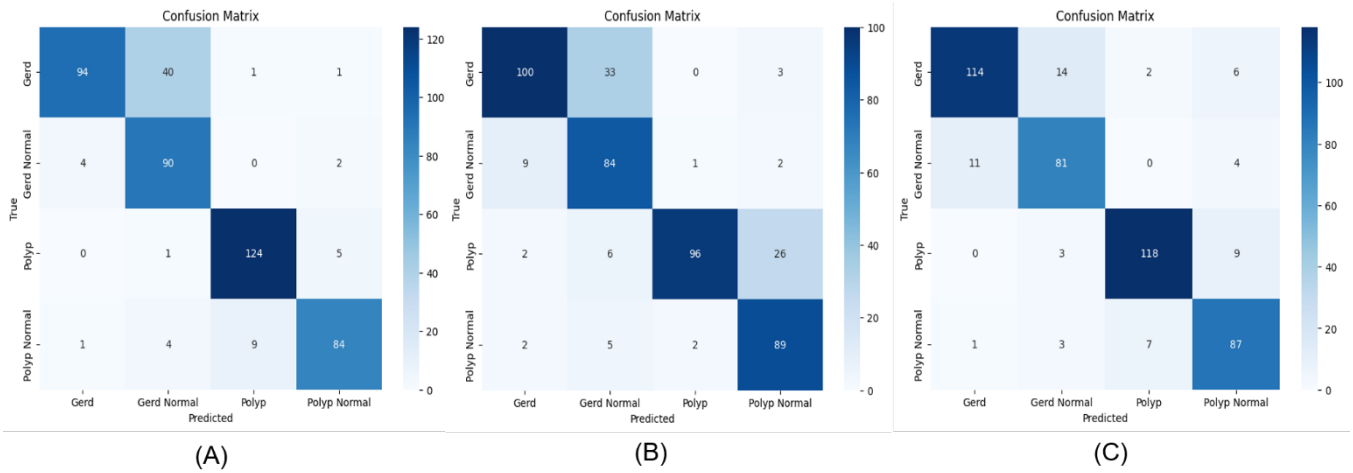

Figure 5. Confusion matrices for individual models and the ensemble model in multi-class gastric disease classification (A) SWIN (B) Vision (C) Ensemble Model

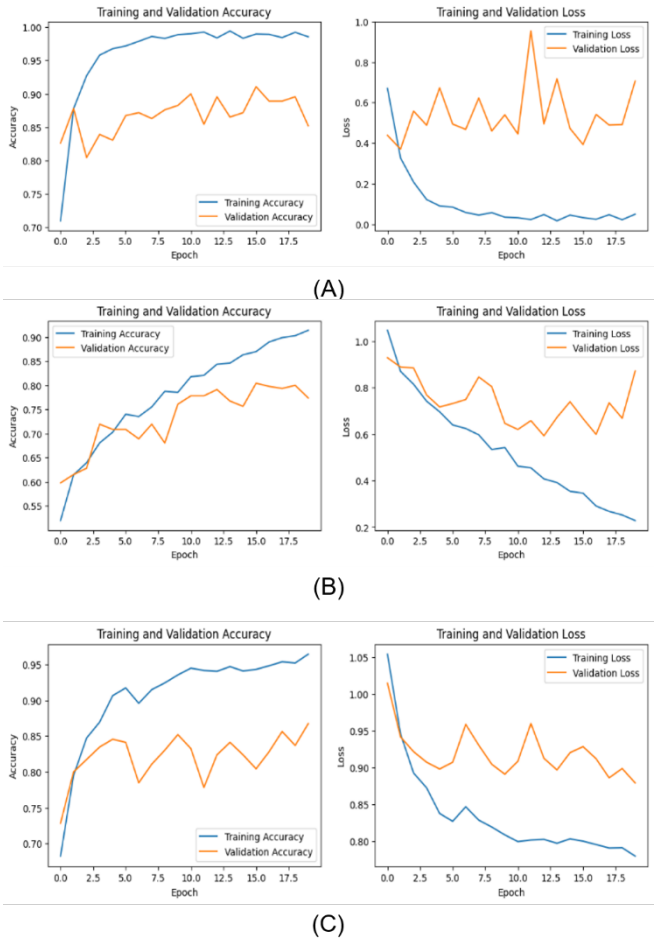

Figure 6. Training and Validation Performance of Transformer Models (A) SWIN (B) Vision (C) Ensemble Model

The ViT model had marginally lower overall performance shows in Table 1 compared to Swin, with a performance of 80% accuracy. It particularly excelled at Precision for Polyp detection (0.97) but lagged at Recall (0.74), exhibiting a lot of false negatives. For the case of GERD classification, the F1 score was equal to Swin (0.80), though ViT showed more balanced precision recall trade-offs. Interestingly, ViT had higher Recall for GERD Normal cases (0.88) compared to Swin, which suggests higher sensitivity towards healthy class detection. Lower Precision scores for Polyp Normal and GERD Normal, though, indicated a certain level of confusion in discriminating these classes. On average, ViT was more sensitive but less precise.

Vision Transformer confusion matrix in figure 5 (B) indicates that it correctly diagnosed 100 instances of GERD but incorrectly diagnosed 9 as GERD Normal, indicating a challenge in distinguishing between pathological and near normal presentations. It correctly classified 96 instances of Polyp with an extremely low level of misclassification, highlighting its ability to detect anomalies. It also correctly labelled 89 instances of Polyp Normal with overlap into the Polyp class. These results indicate ViT recognition of visual features concerned with average for each class as in term of Gerd normal it detects 33 images as Gerd. From the training and validation curves in figure 6 (B), the Vision Transformer model demonstrated great learning stability. The training accuracy rose rapidly and plateaued at around 90%, whereas the validation accuracy oscillated between 80%, indicating moderate generalization.

The best performance across all the measures was posted by the proposed Ensemble Transformer, a blend of Swin's and ViT's output. From Table I we can see that in GERD classification, it significantly improved Recall (0.84) with minimal loss in Precision, earning an outstanding and balanced F1 score of 0.87. Similarly, in Polyp detection, it maintained Swin's high scores with enhancements over ViT's weaknesses, having a strong F1 score of 0.92. The combination further improved class balance in Polyp Normal and GERD Normal differences, indicating the effectiveness of the generalization. The overall highest accuracy of 87% further attests that combining complementary features of the

two models enhances classification credibility and reduces bias.

The confusion matrix of the ensemble from figure 5 (C) reveals high classification accuracy between classes. The model accurately classified 114 GERD cases with minimal misclassification to GERD Normal (14), Polyp (2), and Polyp Normal (6), demonstrating high sensitivity to GERD specific features. The model accurately classified 81 GERD Normal cases with minimal confusion primarily towards GERD, revealing faint visual similarities between pathological and normal presentation in their early stages. The model itself worked very effectively in predicting the Polyp cases, with only 118 exact predictions and minimum misclassifications, validating the model's potential in detecting unique pathological signals. It also classed correctly 87 Polyp Normal cases with minimal overlaps in Polyp, which reflects the high-class separability.

Training and validation graphs of the ensemble depict a smooth and consistent learning process shows in figure 6 (C). Training precision depicts a gradual increase, moving towards 97% and reaching a plateau, whereas validation precision is trailed closely, rising to a maximum of 87%, displaying effective generalization with minimal overfitting. Loss graphs depict a gentle slope downward for training and validation. The trend shows effective optimization and balanced training dynamics.

Compared directly to the two individual models, Ensemble Transformer outperformed both ViT and Swin. The Ensemble Transformer had high F1 and accuracy scores in all four classes with low misclassifications and detecting subtle patterns better than the two individual models. the Ensemble model utilized the global context of ViT and the local accuracy of Swin to develop a synergistic approach that yielded the most consistent and accurate results among disease groups. From figure 7 we can see the performance how accurately it predicts the disease.

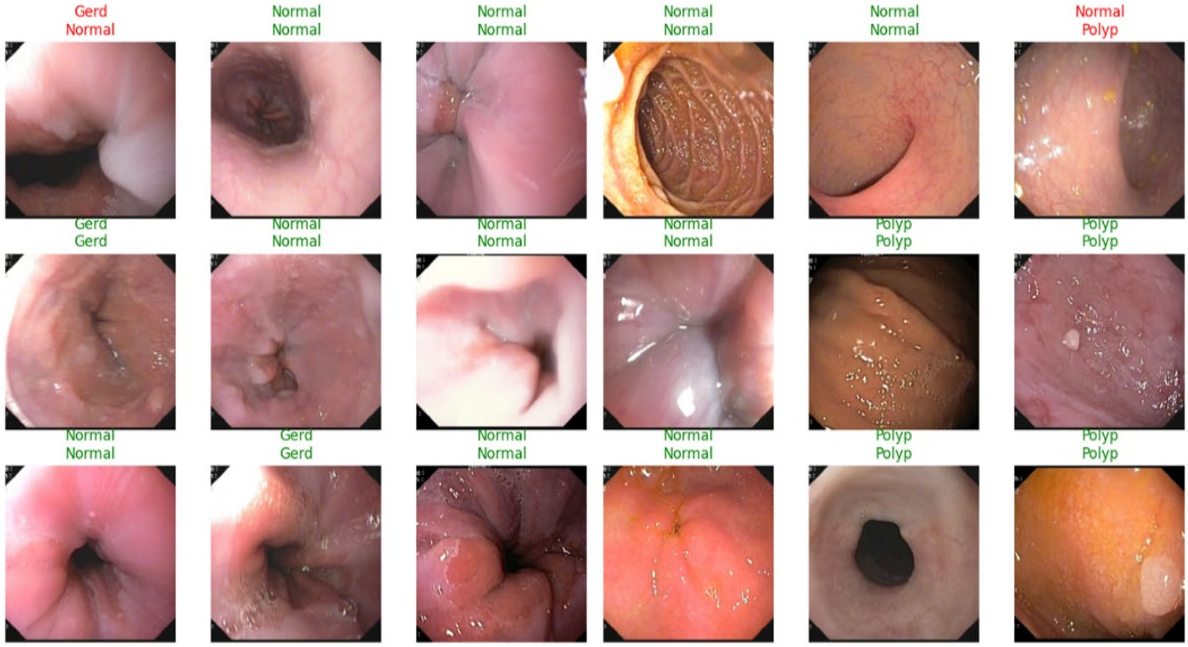

Figure 7. Visual Examples of Correct and Incorrect Classifications by the Ensemble Transformer Model

The Grad-CAM and Grad-CAM++ visualization outcomes offer valuable insights into the interpretability of our proposed transformer-based models in gastrointestinal diagnosis. As shown in Figure X, both techniques were able to highlight the discriminative regions that governed the model's predictions for GERD, GERD Normal, Polyp, and Polyp Normal classes. In cases of GERD, the heatmaps concentrated on inflamed or eroded mucosal areas, resonating with clinically relevant features of reflux. On the other hand, for GERD Normal, the regions of attention were more broadly distributed, consistent with normal anatomy. In the case of Polyp classification, Grad-CAM and Grad-CAM++ directed attention to protruding lesions, indicating appropriate model localization of polyps. For Polyp Normal images, heatmaps showed minimal or diffuse focus, consistent with the absence of pathology. Interestingly, Grad-CAM++ produced more accurate and clearer visualizations compared to vanilla Grad-CAM, with greater clarity in delineating pathological regions. These visual explanations validate that not only is our Ensemble Transformer model achieving high classification performance, but also that it is basing its decisions on clinically interpretable regions. This level of transparency is critical for real-world medical deployment, where trust and interpretability are as important as accuracy.

For its performance, our model has its limitations. Possibly the most significant criticism is that transformer-based models, and ensembles in particular, are computationally

intensive, which may pose barriers to deployment in low-resource clinical environments. Additionally, the models are still reliant on the quality and consistency of endoscopic images; variations in lighting, angle, or resolution may affect generalizability. Nevertheless, the therapeutic utility of this ensemble approach is substantial—by incorporating both local and global attention mechanisms, it allows for robust detection of complex patterns in gastrointestinal images, leading to early diagnosis of GERD and polyps. As a policy recommendation, we propose gradual implementation of these AI-powered diagnostic aid systems in public and private hospitals, starting with pilot programs in teaching hospitals and gastroenterology departments. Apart from this, we recommend investment in digital infrastructure, education, and training programs for clinicians on AI interpretability tools like Grad-CAM++, and the development of regulatory frameworks that ensure safe, ethical, and explainable use of AI in medical imaging.

Grad-CAM and Grad-CAM++ heatmaps were assessed by gastroenterologists with a high degree of endoscopic diagnostic experience. Both reviewers confirmed that the emphasized regions were consistent with clinically relevant features—e.g., mucosal disruption, erythematous plaques, and projecting lesions—in line with traditional diagnostic criteria. For GERD examples, attention maps correctly pointed out the inflammation of the Z-line, but for polyp examples, the heatmaps consistently indicated the polyp head and stalk. The clinicians noted that Grad-CAM++ pinpointed edges better, which might be beneficial for surgical planning or focused biopsy.

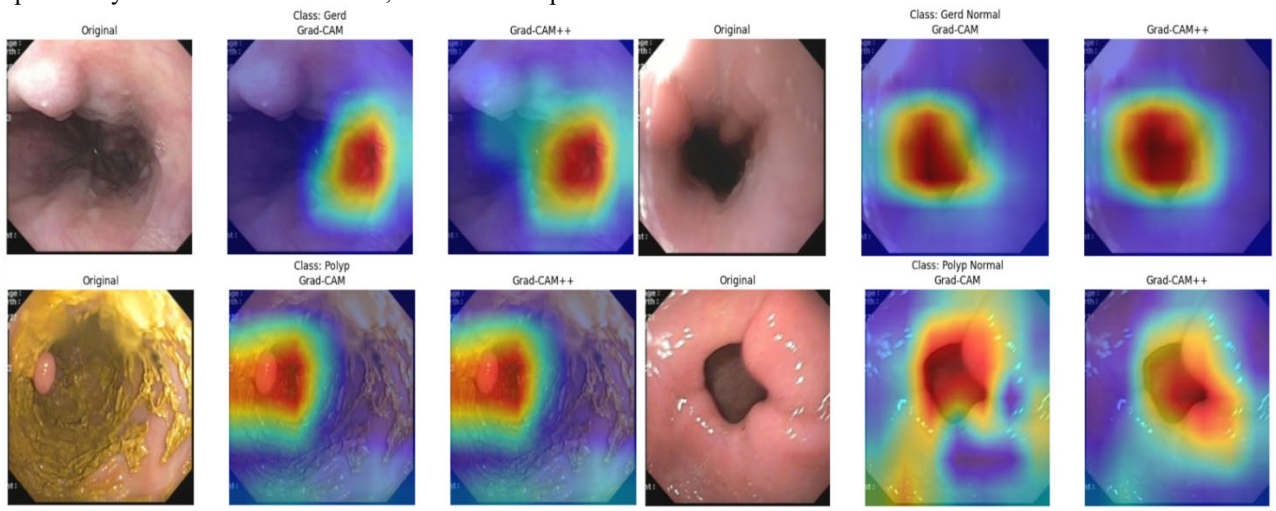

Figure 8. Explainability Analysis Using Grad-CAM and Grad-CAM++ for Gastrointestinal Disease

5. Conclusion

The comparative study of ViT, Swin Transformer, and Ensemble Transformer model for gastric disease prediction presents meaningful observations on the capabilities and limitations of transformer-based models in medical image classification. Swin Transformer, with its spatial locality and hierarchical self-attention, performed incredibly well on Polyp and Polyp Normal class classification and showed robust anomaly detection. But it was moderately challenging to separate GERD from its control counterpart, possibly due to the fine or fuzzy visual cues common in GERD pathology. Vision Transformer, as promising as it is, was comparatively low in spatial accuracy compared with Swin and showed uneven performance in GERD-related classes classification. Its flat attention mechanism can theoretically limit its capacity to find the fine-grained information demanded in subtle diagnostic discrimination. The Ensemble Transformer model combining the output of ViT and Swin offered a generalized and balanced approach. By combining the global contextual features of ViT with the spatial localization capabilities of Swin, the ensemble offered improved classification accuracy for each of the four classes, reduced misclassifications, and maintained better model robustness. To aid model interpretability, we also employed explainable AI (XAI) techniques such as Grad-CAM and Grad-CAM++, which provided us with visual explanations of decision regions. The visualizations guaranteed the models were directing attention to clinically significant anatomical features, thereby ensuring diagnostic confidence, and enabling cross-validation by clinical professionals. Future endeavors on this research include expanding the dataset with more diverse patient populations and endoscopy equipment to increase generalizability. Incorporation of clinical metadata such as patient history, symptoms, age, and biopsy results would further make the model context-aware and predictive. We also plan to explore light transformer architectures for real-time deployment in resource-constrained clinical settings. Further advancements in XAI, including temporal explainability for video endoscopy and user-oriented interpretability dashboards for clinicians, remain highest on the priority list. Lastly, collaborations with gastroenterologists to facilitate potential clinical validation

studies will be crucial to bringing this AI platform into safe, ethical, and effective clinical use.

References

[1] Al-Worafi YM. Epidemiology and Burden of Respiratory Diseases in Developing Countries. InHandbook of Medical and Health Sciences in Developing Countries: Education, Practice, and Research 2023 Dec 19 (pp. 1-24). Cham: Springer International Publishing.

[2] Habib SH, Saha S. Burden of non-communicable disease: global overview. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2010 Jan 1;4(1):41-7.

[3] Peery AF, Crockett SD, Barritt AS, Dellon ES, Eluri S, Gangarosa LM, Jensen ET, Lund JL, Pasricha S, Runge T, Schmidt M. Burden of gastrointestinal, liver, and pancreatic diseases in the United States. Gastroenterology. 2015 Dec 1;149(7):1731-41.

[4] Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. CA: a cancer journal for clinicians. 2023 May;73(3):233-54.

[5] Al-Worafi YM. Gastroesophageal Reflux Disease Management in Developing Countries. InHandbook of Medical and Health Sciences in Developing Countries: Education, Practice, and Research 2024 Feb 6 (pp. 1-43). Cham: Springer International Publishing.

[6] Islam MN, Hasan M, Hossain MK, Alam MG, Uddin MZ, Soylu A. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. Scientific Reports. 2022 Jul 6;12(1):11440.

[7] Alsulami AA, Albarakati A, Al-Ghamdi AA, Ragab M. Identification of anomalies in lung and colon cancer using computer vision-based Swin Transformer with ensemble model on histopathological images. Bioengineering. 2024 Sep 28;11(10):978.

[8] Islam MN, Hasan M, Hossain MK, Alam MG, Uddin MZ, Soylu A. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. Scientific Reports. 2022 Jul 6;12(1):11440.

[9] Atabansi CC, Nie J, Liu H, Song Q, Yan L, Zhou X. A survey of Transformer applications for histopathological image analysis: New developments and future directions. BioMedical Engineering OnLine. 2023 Sep 25;22(1):96.

[10] Horie Y, Yoshio T, Aoyama K, Yoshimizu S, Horiuchi Y, Ishiyama A, Hirasawa T, Tsuchida T, Ozawa T, Ishihara S, Kumagai Y. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointestinal endoscopy. 2019 Jan 1;89(1):25-32.

[11] Yang YJ, Bang CS. Application of artificial intelligence in gastroenterology. World journal of gastroenterology. 2019 Apr 14;25(14):1666.

[12] Suzuki H, Yoshitaka T, Yoshio T, Tada T. Artificial intelligence for cancer detection of the upper gastrointestinal tract. Digestive Endoscopy. 2021 Jan;33(2):254-62.

[13] Visaggi P, De Bortoli N, Barberio B, Savarino V, Oleas R, Rosi EM, Marchi S, Ribolsi M, Savarino E. Artificial intelligence in the diagnosis of upper gastrointestinal diseases. Journal of Clinical Gastroenterology. 2022 Jan 1;56(1):23-35.

[14] Nogueira-Rodríguez A, Domínguez-Carbajales R, López-Fernández H, Iglesias Á, Cubiella J, Fdez-Riverola F, Reboiro-Jato M, Glez-Pena D. Deep neural networks approaches for detecting and classifying colorectal polyps. Neurocomputing. 2021 Jan 29;423:721-34.

[15] Quan SY, Wei MT, Lee J, Mohi-Ud-Din R, Mostaghim R, Sachdev R, Siegel D, Friedlander Y, Friedland S. Clinical evaluation of a real-time artificial intelligence-based polyp detection system: a US multi-center pilot study. Scientific Reports. 2022 Apr 21;12(1):6598.

[16] Wang P, Berzin TM, Brown JR, Bharadwaj S, Becq A, Xiao X, Liu P, Li L, Song Y, Zhang D, Li Y. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019 Oct 1;68(10):1813-9.

[17] Al-Otaibi S, Rehman A, Mujahid M, Alotaibi S, Saba T. Efficient-gastro: optimized EfficientNet model for the detection of gastrointestinal disorders using transfer learning and wireless capsule endoscopy images. PeerJ Computer Science. 2024 Mar 11;10:e1902.

[18] Zhang R, Zheng Y, Poon CC, Shen D, Lau JY. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern recognition. 2018 Nov 1;83:209-19.

[19] Wesp P, Grosu S, Graser A, Maurus S, Schulz C, Knösel T, Fabritius MP, Schachtner B, Yeh BM, Cyran CC, Ricke J. Deep learning in CT colonography: differentiating premalignant from benign colorectal polyps. European Radiology. 2022 Jul;32(7):4749-59.

[20] Owais M, Arsalan M, Mahmood T, Kang JK, Park KR. Automated diagnosis of various gastrointestinal lesions using a deep learning–based classification and retrieval framework with a large endoscopic database: model development and validation. Journal of medical Internet research. 2020 Nov 26;22(11):e18563.

[21] Nogueira-Rodríguez A, Domínguez-Carbajales R, Campos-Tato F, Herrero J, Puga M, Remedios D, Rivas L, Sánchez E, Iglesias A, Cubiella J, Fdez-Riverola F. Real-time polyp detection model using convolutional neural networks. Neural Computing and Applications. 2022 Jul;34(13):10375-96.

[22] Li R, Li J, Wang Y, Liu X, Xu W, Sun R, Xue B, Zhang X, Ai Y, Du Y, Jiang J. The artificial intelligence revolution in gastric cancer management: clinical applications. Cancer Cell International. 2025 Mar 21;25(1):111.

[23] Lei C, Sun W, Wang K, Weng R, Kan X, Li R. Artificial intelligence-assisted diagnosis of early gastric cancer: present practice and future prospects. Annals of medicine. 2025 Dec 31;57(1):2461679.

[24] Chaudhary RG, Dhangar P, Chaudhary AG. Artificial Intelligence in Gastrointestinal Endoscopy: A Comprehensive Systematic Review. medRxiv. 2025:2025-07.

[25] Abu Kowshir Bitto, Rezwana Karim, M. H. Begum, M. F. I. K. Khan, Dr. Md. Maruf Hassan, and Prof. Dr. Abdul kadar Muhammad Masum, "Explainable AI Based Deep Ensemble Convolutional Learning for Multi-Categorical Ocular Disease Prediction", *EAI Endorsed Trans AI Robotics*, vol. 4, Jul. 2025.

[26] Badruzzaman Biplob KB, Sammak MH, Bitto AK, Mahmud I. COVID-19 and Suicide Tendency: Prediction and Risk Factor Analysis Using Machine Learning and Explainable AI. EAI Endorsed Transactions on Pervasive Health & Technology. 2024 Jan 1;10(1).

[27] Bitto AK, Bijoy MH, Shakil KH, Das A, Biplob KB, Mahmud I, Hossain SM. GastroEndoNet: Comprehensive endoscopy image dataset for GERD and polyp detection. Data in Brief. 2025 Jun 1;60:1115