IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 7 September 2025, accepted 27 September 2025, date of publication 6 October 2025, date of current version 16 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3616796

RESEARCH ARTICLE

Exploring Vision Transformers and Explainable AI for Enhanced Artefact Classification in Esophageal Endoscopic Images

PREETI BISSOONAUTH-DAIBOO[1], MUHAMMAD MUZZAMMIL AUZINE[1], MUHAMMAD I. KHAN[2,3], FATIMA ALSHANNAQ[2], TANZILA SABA[2], XIAOHONG GAO[4], AND MALEIKA HEENAYE-MAMODE KHAN[1], (Senior Member, IEEE)

[1]Department of Software and Information Systems, University of Mauritius, Réduit 80837, Mauritius
[2]Artificial Intelligence and Data Analytics Laboratory (AIDA), CCIS, Prince Sultan University, Riyadh 11586, Saudi Arabia
[3]School of Engineering, University of Warwick, CV4 8UW Coventry, U.K.
[4]Department of Computer Science, Middlesex University London, NW4 4BT London, U.K.

Corresponding author: Tanzila Saba (tsaba@psu.edu.sa)

ABSTRACT Esophageal cancer (EC) remains the disease that has the highest incidence and highest mortality rate in global cancer statistics, emphasising the imperative to enhance diagnostic precision and reliability through the use of advancing technologies. While AI-enhanced systems can improve the early detection of EC considerably, the prevalence of artefacts (1 in 4 frames) during endoscopy procedures compromises the developed systems significantly, leading to unreliable medical decision making. Vision transformer (ViT) networks, initially designed for natural language processing tasks, have demonstrated outstanding performance in handling medical images by presenting distinctive features advantageous for image processing. The application of ViT for detecting and classifying artefacts in endoscopic images, particularly in classifying colour misalignment artefacts is still subject to continual refinement and enhancement. This work aims to investigate the implementation of ViT for classification of colour misalignment artefacts in esophagus endoscopy images. Moreover, even though ViT has been a major breakthrough, its acceptance for real world applications is often jeopardised due to the lack of interpretability of how the classification results have been reached. Consequently, Explainable Artificial Intelligence (XAI) techniques have been explored to understand the criteria used to achieve the outcome. Several variants of the ViT and Data Efficient image Transformer (DeiT) networks have been fine-tuned and applied to our dataset in order to improve and evaluate their performance in colour misalignment classification in esophagus endoscopic images. Furthermore, XAI methods have been implemented to provide the criteria used by the network in reaching the classification results. Our fine-tuned ViT model, achieves an accuracy of 93.46%, precision of 93.48%, recall of 93.46% and F1 score of 93.46% surpassing InceptionResNetV2, a state-of-the-art model based on CNN, with an accuracy of 89.10%, precision of 89.10%, recall of 89.10% and F1 score of 88.23%. Additionally, the GradCAM XAI technique has been found to highlight the deterministic features used by the ViT model better than other XAI methods applied in this work. ViT achieves remarkable performance in classification of colour misalignment artefact outperforming CNNs, attributed to ViT's enhanced ability to capture pixel relationships through self-attention weights. In addition, the intrinsic self-attention technique provides novel insights into the model's decision-making mechanism.

INDEX TERMS Endoscopic artefact, vision transformer, explainable AI, healthcare.

I. INTRODUCTION

Esophageal cancer was ranked as the seventh most common cancer and the sixth leading cause of cancer-related deaths

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

globally in 2020 [1]. Among men, it was the seventh most prevalent cancer, while in women, it was the thirteenth. In that same year, there were over 600,000 new esophageal cancer cases in the world [2]. Esophageal cancer can develop when a malignant tumor arises in the lining of the esophagus or at the junction where the esophagus connects to the stomach [3]. The two main histological types of esophageal cancer are esophageal adenocarcinoma (EAC) and squamous cell carcinoma cancer (SCC). Esophagogastroduodenoscopy (EGD) is the endoscopy of the upper part of the gastrointestinal (GI) tract which is made up of the esophagus, stomach and duodenum (top of the small intestine). EGD is widely used to identify potential suspected areas for further diagnosis of esophageal cancers [4]. Endoscopy has been used for several decades since the first endoscope was developed [5]. An endoscope is a long tube with a camera and light at its end which is inserted into the organ during EGD. Endoscopic images are taken and displayed in real time on a video monitor or stored for future analysis [6]. One major hindrance to the visual interpretation of endoscopic images are the presence of artefacts. The latter, by definition, are artificial elements or distortions that are discovered during scientific investigations or experiments that do not exist naturally but develop as a consequence of the preparative or investigative processes. Several factors can introduce artefacts during endoscopy. These include the limited space within the organ, the hand motion of the endoscopists as well as over or under exposure to changes of illumination as the endoscopic camera moves through the esophagus. Artefacts can also be in the form of body fluids and debris. Artefacts can often be misinterpreted for suspicious areas or can occlude endoscopists' view of possible lesions which can probably lead to cancer. Various studies effected in 2020 have concluded that from 2018 to 2020 the rate of upper gastrointestinal cancers, which go undetected, has increased considerably from 9.8% to 25.8%. The foremost cause of this rise is error during endoscopy [7]. It has been reported that more than 60% of endoscopic video frames and approximately 70% of endoscopy video sequences can often be damaged with artefacts [8].

Fig. 1 and Table 1 show the most common artefacts that can be found in esophagus endoscopic images. Amongst the artefacts, a prominent one is the colour misalignment which can significantly impact the accuracy and interpretation of endoscopic images. Colour misalignment in esophagus endoscopic images occurs when there is a discrepancy between the 3 colour channels (Red, Green and Blue) captured by the endoscopic camera and the actual ones represented in the images. This discrepancy can be attributed to several factors. These include the lighting conditions during the endoscopy procedure, the unsteady movement of the endoscopist's hands or the tight space within the organ being diagnosed. Addressing and minimising colour misalignment artefact is crucial for enhancing the reliability of esophagus endoscopic images and ensuring the effectiveness of diagnostic and therapeutic interventions [9].

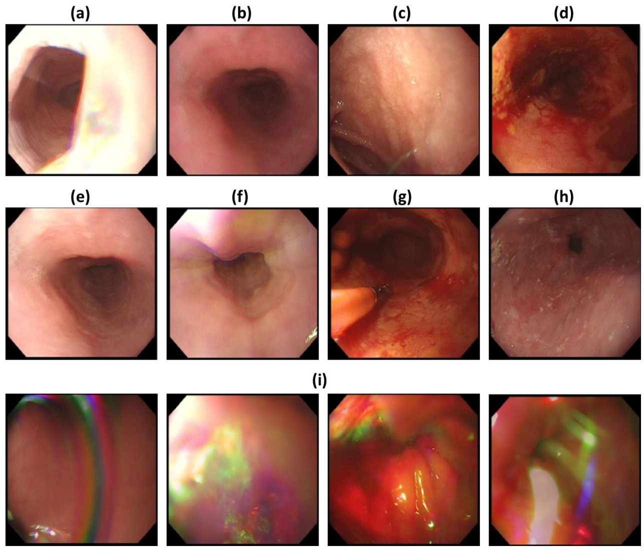


FIGURE 1. Common artefacts: (a) Saturation; (b) Contrast; (c) Specularity; (d) Blood; (e) Bubble; (f) Blur; (g) Artifacts; (h)Debris and Mucosa; (i) Colour misalignment.

TABLE 1. Details of common artefacts.

Artefact	Source of occurence
Saturation	Due to over exposure of image regions because of changes in illumination [6]
Contrast	Due to under exposure to light because of changes in illumination [6].
Specularity	Due to non-optimal reflection of light since surface of organs are smooth [6].
Blood, debris and mucosa	Bodily fluids and medicines used during endoscopy [6].
Bubble	Thin film of liquid with air that distorts appearance of tissue [8].
Blur	Caused by the unsteady hand motion of endoscopists [10].
Artifacts	Apparatus or other materials used during the endoscopy [9]
Colour misalignment	Images are taken in sequence in red, blue and green by the endoscopic camera. The three colour channels are then combined to create a single image. However, because the esophagus is narrow and the camera moves quickly, the three colour channels might not be gathered from the exact same location, resulting in a distorted appearance [12].

Recently, ViT has attained cutting-edge performance in Natural Language Processing. The achievement of the work by [13] has spurred research in ViT application for image processing. There is indeed a recent expansion in the use of ViTs for image identification tasks [14], [15], [16]. ViT has attained groundbreaking advancement in standard computer vision work with respect to natural images such as classification of ImageNet, object detection and semantic segmentation. The attention mechanism integral to transformers presents several notable advantages: (1) it captures long-range dependencies effectively, (2) it adaptively models relationships through computed self-attention weights, thereby elucidating the interactions between tokens, and (3) it incorporates an inherent form of saliency, revealing the focal points of the model's attention [17]. Through the publications of the performance of ViT, there is definitely much potential in the exploitation of this technique. Therefore, ViT can decisively contribute to the effective detection and classification of artefacts in endoscopic images.

The main contributions of this study can, therefore, be summarised as follows:

1) **Application of ViT to Artefact Detection and Classification in Endoscopic Images:** ViT, a novel architecture has been utilized for the task of artefact detection and classification within endoscopic images. By leveraging on ViT's attention mechanism, we address challenges affecting artefact detection, thereby, supporting more accurate early cancer detection.
2) **Analysis of ViT Performance for Colour Misalignment Artefact Classification:** The ability of ViT to accurately identify and classify colour misalignment artefacts, commonly encountered in endoscopic imaging, has been assessed by providing insights into the model's strengths and potential limitations.
3) **Comparison of the results of the identified best ViT model against state-of-the-art ViTs applied to medical image classification:** A thorough comparison is performed by benchmarking the performance of our selected high performing ViT model against other state-of-the-art ViTs utilised in medical image classification to assess its competitiveness and efficiency in medical imaging in general.
4) **Comparative Performance Evaluation of ViT and CNN:** This study conducts an in-depth comparative evaluation by contrasting the performance of ViT with that of CNNs used for colour misalignment artefact classification in endoscopic images to provide the relative strengths and weaknesses of ViT and CNN in artefact recognition in this domain.
5) **Review and application of XAI to provide explainability of the outcome of our enhanced model:** XAI has been applied to the artefact detection models developed to provide explanations for their predictions. Understanding the rationale behind the decisions made by deep learning models can provide insights into how they arrive at their conclusions thus, increasing trust and confidence for the integration of the ViT models into clinical practice.

To the best of our knowledge, our work consists of one of the first application of ViT to artefact detection and classification in esophagus endoscopic images. In the following section, we outline the materials and methods utilised in our proposed approach. Section IV elaborates on the experiments performed and results obtained. Finally, we provide a detailed discussion of the results before concluding the paper in the last section.

II. RELATED WORK

Several works involving ViT have been conducted over the last decade in this field. Reference [17] has explored the potential of ViT models to replace CNNs in the processing of medical images. Multiple experiments have been conducted with three datasets: APTOS 2019 including images of diabetic retinopathy, ISIC 2019 comprising of dermoscopic images depicting skin lesions as well as CBIS-DDSM containing mammography images. ResNet50 has been employed as the model based on CNN and as ViT, DeiT-S was utilised. DINO served as the learning technique based on self-supervision. The ViT models have outperformed the CNN based methods. For instance, a Cohen Kappa of 89.6 has been obtained with DeiT for classifying APTOS2019 dataset into 5 disease severity categories, compared to 89.3 with ResNet50. For classification of ISIC2019 into skin diseases comprising of 9 classes, a Recall of 84.4 was obtained on DeiT compared to 81 for ResNet50. On the other hand, for detection of masses in the DDSM mammography images, an ROC-AUC of 94.7 was obtained for DeiT model whereas with ResNet50, the value obtained is 95.3. Similarly, with supervised learning, for classification on APTOS2019 dataset, a Cohen Kappa of 89.6 was obtained with DeiT compared to 89.4 with ResNet50; on ISIC2019, a Recall of 85.3 was obtained on DeiT compared to 83.3 for ResNet50 while with DDSM (object detection), an ROC-AUC of 95.6 was obtained for DeiT model whereas with ResNet50, the value obtained is 95.5.

Reference [18] has assessed ViT performance against CNN for classifying breast cancer ultrasound images into normal, benign and malignant categories. Among the CNN based models compared are Inceptionv3, ResNet50, VGG16, and NASNET. ResNet50 model achieved the best performance with an accuracy of 85.3% and Area Under Curve (AUC) of 0.95. ViT model using transfer learning attained an accuracy of 86.7% and AUC of 0.95. The results show the higher performance of the ViT model compared to that of CNN in terms of accuracy but for AUC both achieved similar results. Point-of-Care (POC) Transformer (POCFormer) has been employed by [19] for classification of Covid-19 disease in POC lungs ultrasound images. POCFormer, being a lightweight ViT model, has decreased the complexity of the self-attention mechanism from quadratic to linear with respect to time and space. A linear transformer model (Linformer) is used in place of the usual transformer model to guarantee that the suggested transformer architecture is memory and time efficient. The average accuracy obtained is above 91% [20].

In the context of Alzheimer's disease, [21] introduced a lightweight deep learning model that utilises MRI data for accurate detection. The model is designed to operate efficiently in real world scenarios with limited computational power, while still achieving high classification accuracy. This demonstrates the potential of lightweight architectures in resource-constrained medical settings. Furthermore, [22] proposed a hybrid learning model that combines deep and hybrid learning techniques to identify various progression stages of Alzheimer's disease. Their study highlights the effectiveness of integrating different learning paradigms to improve the sensitivity and specificity of early diagnosis in neurodegenerative disorders using MRI. Similarly, [23] presented a CNN-based deep learning approach to detect Alzheimer's disease through MRI. Their method, trained on brain scan datasets, effectively classified Alzheimer's and non-Alzheimer's cases, reaffirming the role of deep learning models such as CNNs in brain disorder diagnostics.

[24] used Vision Outlooker (VOLO), a novel attention technique, which effectively encodes low level features into ViT, enhancing classification performance. Specifically, an image of size 224 × 224, for instance, is first tokenised using smaller patches such as 8 × 8 and then token representations are encoded at a fine level (e.g., 28 × 28) before global dependencies at a coarse level is built using self-attention. VOLO has achieved an accuracy of 87.1% for classification on ImageNet dataset. [25] made use of Swin Transformer and Transformer-in-transformer for classification of chest X-ray images with COVID-19 from normal and pneumonia images. The accuracy score obtained is 0.9475, sensitivity score of 0.9475 and a specificity score of 0.9509. Hybrid Transformer-CNN models have also been used for the classification of medical images. In [26] the benefits of CNN and ViT have been combined for the extraction of local and global information in gastric histopathological images for the classification of gastric cancer. Inceptionv3 is used as the CNN model. The results for the GasHisTransformer model is accuracy of 97.97%, precision of 98.55%, recall of 97.38% and an F1 score of 97.97%. With the lightweight version, the results are an accuracy of 96.43%, precision of 95.99%, recall of 96.90% and an F1 score of 96.43%.

[27] makes use of the benefits of transformer and convolution for polyp detection by setting up a network of convolution in transformer, the COTR. The COTR consists of a CNN for feature extraction, layers of transformer encoder combined with convolutional layers for the encoding and reconstruction of features, layers of transformer decoder for predicting objects as well as a feed forward network for detection prediction. On the ETIS-LARIB dataset, COTR obtained 91.49% precision, 82.69% sensitivity, and 86.87% F1-score; on the CVC-ColonDB, it obtained 91.67% precision, 93.54% sensitivity, and 92.60% F1-score. For the classification of Wireless Capsule Endoscopic images, [28] made use of a pure ViT model, which does not have any convolutional operations utilised by CNNs. Spatial downsampling is used in order to fit the spatial structure of ViT. The model has been trained from scratch. With an accuracy of 79.15% on the Kvasir-Capsule dataset and 98.63% on the RLE dataset, the technique employed performed significantly better than both CaiT and Swin Transformer on both WCE datasets.

[29] introduces RanMerFormer for classification of brain tumor in MRI images. RanMerFormer consists of a pre-trained ViT as its backbone together with a token-merging mechanism to eliminate redundancy and improve processing efficiency. Additionally, a randomised vector functional-link is utilised as the head for quicker training. An accuracy of around 99% has been achieved on 2 publicly available MRI datasets. However, there is still improvement required for certain classes in the dataset as well as for enhancing the performance of the ViT model. Pre-trained ViT models including ViT base, large and huge variants have been employed by [30], for the classification of breast cancer. The large ViT method has obtained the best accuracy of 94.32%. Data augmentation has been applied for gaining a large dataset for training. [31] has utilised a patch based Vit (PBVit) model for the detection and classification of brain tumor. Compared to other state-of-the-art techniques applied in this work for comparison, the PBVit model has achieved quite high accuracy, precision, recall and F1-score of 95.8%, 95.3%, 93.2%, and 92% respectively. While the model could be enhanced to make it better applicable to more real-life scenarios, interpretability should be considered to ensure its acceptance in the decision making of medical professionals.

Recently, Anticancer peptides (ACPs) have been proposed as alternatives to traditional cancer detection approaches. ACPs are a sub-category of anti-microbial peptides and have potentially anti-cancerous effects. pACPs-DNN has been put forward as an innovative attention-based deep learning model for the prediction of ACPs [32]. In [33], pNPs-CapsNet, an attention based model, has been utilized to predict neuropeptides (NPs), which are essential signaling molecules and which possess therapeutic characteristics critical in cancer treatment. [32] also introduced pACP-HybDeep for accurately predicting anti-cancer peptides. The model is based on attention and CNN/RNN is used for training. All these models obtained remarkable accuracies in the prediction of peptides.

After scrutinising related works regarding utilisation of ViT, it has been observed that the latter has been applied to medical image processing tasks such as object detection, classification and segmentation, attaining quite high performance. However, its application has been relatively limited in the context of artefact detection and classification within endoscopic images. Furthermore, detecting esophageal cancers is challenging and the existence of colour misalignment artefacts further complicates this task for both human specialists and automated systems due to the lack of realism it brings to endoscopic images. There is an urgent need to detect colour misalignment artefacts.

On the other hand, medical practitioners are very reluctant in using AI applications since they are considered to be 'closed-box' systems. To address this challenge, XAI models have emerged. However, some are still at their embryonic stage. Some visualisation techniques that can be used for

XAI are GradCAM, SHAP and LIME. GradCAM is found to perform better than SHAP and LIME for the classification of WCE images [34]. GradCAM is able to depict the areas with potential COVID19 well in the CXR images [35]. Similarly, for identification of pneumonia in CXR images, GradCAM provides better visualisations [36]. It can be seen that different other XAI methods have been used as well. TransLRP has been used for classification of chest X-ray images. In some cases, it performs better than LIME and Attention rollout based on the high value of faithfulness and small sensitivity while LIME is found to be best for visualisation. In [37], ViT Shapley provides better explanation than the baseline approaches such as GradCAM, Attention rollout and LRP for classifying images from musculoskeletal radiographs. It is noted that in order to make Shapley values more computer efficient, they are integrated with ViT and it is seen that the approach employed provides more accurate explanations than existing methods for ViTs. [38] have applied different CNN and ViT models for distinguishing between Crohn' s disease and ulcerative colitis and it is found that ViT-S/16 provides the best results both in terms of classification of the disease as well as its interpretability instead of Grad-CAM, LIME, SHAP values, RISE, and Occlusion sensitivity. The ViT-B/16 model applied by [39] for classification together with Gradient Attention Rollout algorithm for explainability is found to highlight suspected areas correctly. The ViT (vit-base-patch16-224) model applied by [40] with Score-CAM algorithm as explainability method is able to clearly show the focused regions used during prediction.

Since ViT has recently confirmed its potential in medical image processing, various endeavours have been undertaken to make ViT models better explainable in this field. However, there remains a notable gap in the application of ViT for providing explainability in the esophageal endoscopic image classification and more so to artefact detection and classification. Therefore, XAI has been implemented to provide interpretability of the classification results obtained from the ViT models.

III. MATERIALS AND METHODS

This section details the methodology used in implementing ViT for classification of colour misalignment artefacts.

A. PROPOSED WORKFLOW AND COMPONENTS

Fig. 2 shows the overall structure of the proposed model. The dataset is first pre-processed and data augmentation is applied. Next, several pre-trained ViT models are used on the original as well as the augmented dataset. The results are analysed and the parameters of each model are fine-tuned until the best performance is obtained. The enhanced networks are then applied to the test data from the dataset which was not used for the training phase. The relative performance of the different models is evaluated based on several performance metrics and the best model is identified.

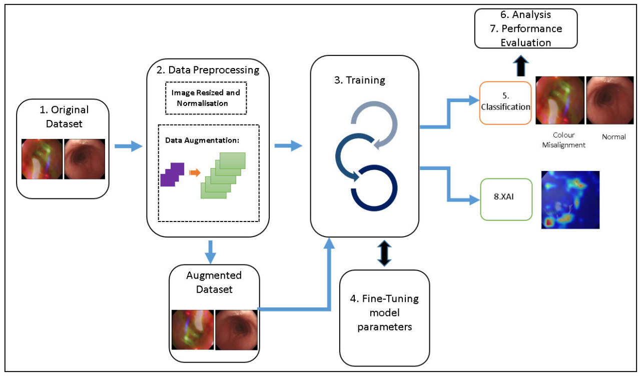


FIGURE 2. Workflow of the proposed model.

The workflow in Fig. 2 is as follows: (1) Preparation of the custom dataset; (2) data pre-processing and augmentation; (3) training of the ViT models using the original and augmented dataset (4) enhancing the ViT models; (5) classification of the dataset; (6) analysis of the results obtained and (7) comparison of the results of our ViT model against state-of-the-art ViT models (8) Application of explainability to the model.

B. DATASET

The dataset utilised in this work originates from the Endoscopy Artefact Detection (EAD) 2019 challenge and is publicly available for research purposes ([10], [11]). It includes seven types of artefacts which are saturations, blur, specular reflections, bubbles, artifacts in imaging, instrument and contrast. It is to be noted that colour misalignment artefacts can form part of the endoscopic images with motion blur. The EAD dataset has been selected instead of the Kvasir dataset [41] or its derivatives such as HyperKvasir [42] or Kvasir-Instrument [43] since the latter do not contain all classes of artefacts. Additionally, conventional endoscopy involves high resolution images which are mainly for the esophagus whereas the Kvasir dataset contains images of intestines only and with much lower resolution.

The EAD dataset was collected from six centers namely John Radcliffe Hospital of UK; ICL Cancer Institute and Ambroise Pare Hospital of France; Instituto Oncologico Veneto of Italy; University Hospital Vaudois of Switzerland and the Botkin Clinical City Hospital of Moscow. The dataset has been procured from different medium including gastroscopy, cystoscopy, gastrooesophageal and colonoscopy. It has also been obtained from multiple modes such as white light, fluorescence and narrowband imaging and involves patients from different countries such as UK, France, Switzerland and Russia. The videos were gathered from patients at Oxford, with randomized sampling at French institutes and cancer patients were exclusively selected from Moscow. The videos were obtained using standard imaging protocols of endoscopes produced by various companies such as Olympus, Biospec as well as Karl Storz. The images were minutely anonymized and there is no patient information visible.

Furthermore, several videos having colour misalignment artefacts have been gathered from a prior study on SCC detection by [44]. These videos are of high definition inclusive of White Light Endoscopy (WLE) and Narrow Band Imaging (NBI). They were obtained from patients who visited Translational Gastroenterology Unit found at the Oxford University Hospital and the Horton General Hospital in Banbury in UK and also from the Beijing General Hospital of China. The videos were captured by Olympus endoscopes – either GIF-H260, GIF-H290 or EVIS Lucera CV260 from the Olympus Medicals Systems of Japan. The patients, from whom the endoscopic video recordings are obtained, have given written consent to provide these recordings together with their biopsies and analysis of their medical data (REC Ref: 16/YH/ 0247) prior to the study [44]. The videos are in .mp4 format at 30 frames per second. The image frames have been extracted from the videos by making use of the VideoProc Converter application [45]. Table 2 exemplifies the proportion of frames with colour misalignment and other artefacts.

TABLE 2. Proportion of artefacts for 5 normal subjects.

Video	Original number of frames extracted	Colour misaligment	Other artefacts
M1	4620	8%	11%
M2	1620	20%	18%
M3	1680	11%	8%
M4	1860	27%	2%
M5	1000	24%	5%

In all in our custom-built dataset, there are thus 1784 images with colour misalignment artefacts and 1866 images with remaining artefacts. Out of these, 2554 images have been used as training data, 629 as validation data and 367 as test data. Overall, there are 3650 images in the custom-built dataset. The images from the EAD dataset as well as the ones obtained from the videos have been verified individually to ensure that they are anonymised, hence, they do not contain any patient's information. Part of our custom-built dataset with endoscopic images containing colour misalignment and the remaining artefacts are shown in Fig. 3 and Fig. 4 respectively.

The dataset has been divided into two classes, one with images containing colour misalignment artefacts, labelled as 'colour' and the other having images with the remaining artefacts, labelled as 'normal'. The images with colour misalignment artefacts represent 51% of the dataset and images with other artefacts represent 49% of the dataset as shown in Fig. 5. Hence, there is clearly imbalance between the 2 classes of data. As discussed in [18] and [26], imbalance in dataset results in poor performance and high misclassification. One possible solution to this problem is data augmentation. Thus, our dataset for training and validation has been augmented both on the fly, that is during processing, as well as using traditional data augmentation techniques before processing.

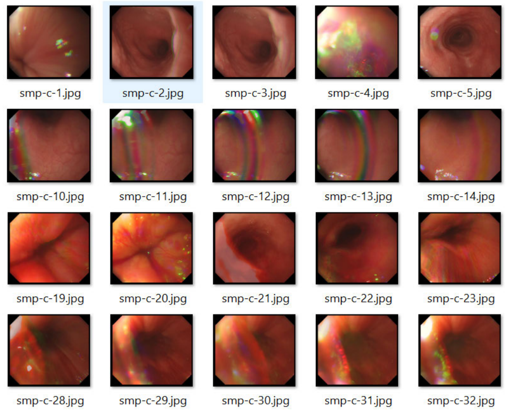


FIGURE 3. Images with colour misalignment.

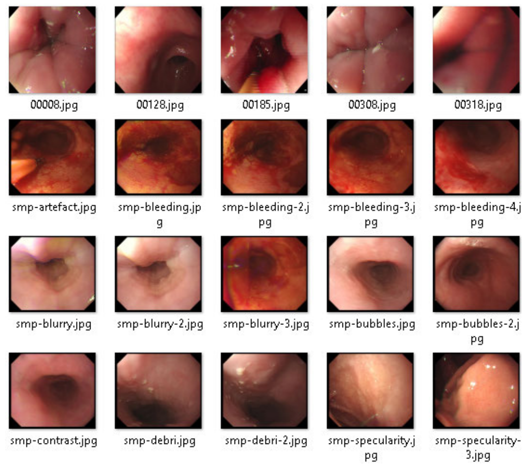


FIGURE 4. Images with other artefacts.

On the fly, the techniques used include rescaling, shear transformation, zooming, horizontal and vertical flip as well as rotation. Another data augmentation technique used is random pixel erasure. To increase the number of images, shearing, zooming and horizontal flip have also been used.

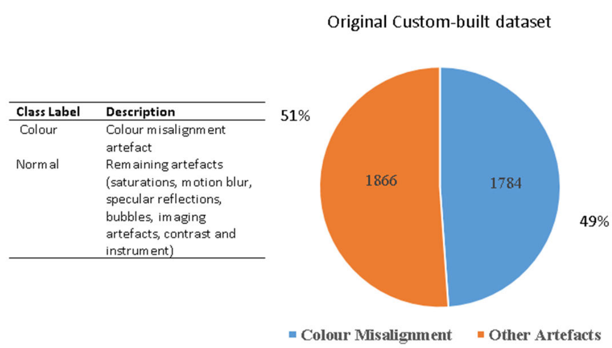


FIGURE 5. Class definition and dataset size.

Table 3 provides the number of images in the 2 classes for both the original custom-built dataset and its augmented version. The images are segregated into training, validation and test dataset in the ratio of 7:2:1.

TABLE 3. Dataset distribution for original and augmented datasets.

	Orignal dataset			Augmented dataset		
Classes	Train	Valid	Test	Train	Valid	Test
Colour misalignment	1248	356	180	4800	1200	180
Other artefacts	1306	373	187	4800	1200	187

C. DATA PRE-PROCESSING

The images in the dataset are of different sizes ranging from 295×299 to 192×1080 pixels. Thus, they need to be reshaped to a fixed size before being passed to the model. Resizing the images allows the training to be faster and less memory intensive. Additionally, the algorithm used in the experiment require the images to be of same size. The images have, therefore, been resized to 224×224 pixels. In order to help in effective processing, the input images should be normalised as well. The pixel values of the input images are in the range of 1 to 255. Hence, the pixel values have been normalised to be in the range of 0 to 1 [46].

D. MODEL ARCHITECTURE

1) VIT ARCHITECTURE

ViT, being a novel architecture which has been applied in our work for colour misalignment classification, is explained in detail in this section. Fig. 6 shows the architecture of a standard ViT model.

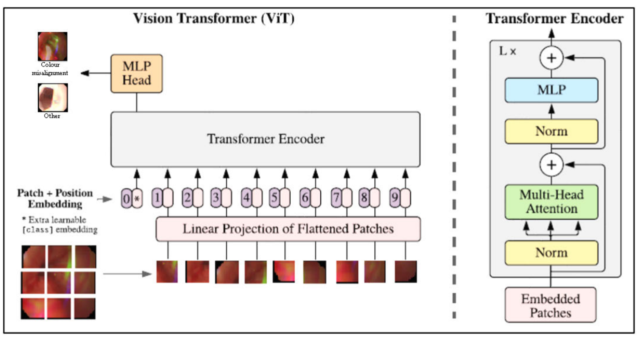


FIGURE 6. Architecture of ViT.

Overall, in ViT, an image is split into patches of fixed size which are flattened and linearly embedded. Positional embeddings are added so as to keep information about the relative positions of the patches in the original image. The resulting sequence of vectors is fed into a standard Transformer encoder module. An extra learnable "classification token" is added to the sequence which is used during pre-training and the parameter fine-tuning phases. The Transformer encoder is made up of alternating layers of Multi-Head Self-Attention (MSA) and Multi-Layer Perception (MLP) blocks. MHSA is derived from the standard Self-Attention mechanism, whereby for each element in the input sequence, a weighted sum with respect to other values in the sequence is computed. Several self-attention operations called 'heads' are carried out in parallel and their concatenated outputs are projected.

The algorithm applied by ViT is further detailed below:

1) Our 2D esophagus image, x is represented in the form of
$$x \in \mathbb{R}^{H \times W \times C} \tag{1}$$
where H is the height, W is the width of the image and C is the number of channels.
2) The 2D images are first reshaped into a sequence of flattened 2D patches
$$x \in \mathbb{R}^{N \times (P^2 C)} \tag{2}$$
where $P^2 = P \times P$ is the resolution of each image patch and $N = \frac{HW}{P^2}$ is the number of patches.
3) The sequence of N flattened patches ($x_1, x_2, x_3, x_4, \ldots, x_N$) are projected using trainable linear projection layer to D dimension which represents each patch
$$X \in \mathbb{R}^{N \times D} \tag{3}$$
4) **MSA Block**
The aim of self-attention is to capture the interaction between the N patches by encoding each patch in terms of the global contextual information. To achieve this, three learnable weight matrices are defined to transform X into
   a) Queries, Q using
   $$W^Q \in \mathbb{R}^{D \times D_q} \tag{4}$$
   b) Keys, K using
   $$W^K \in \mathbb{R}^{D \times D_k} \tag{5}$$
   c) Values, V using
   $$W^V \in \mathbb{R}^{D \times D_v} \tag{6}$$
   where $\quad D_q = D_k$
5) The input sequence X is first projected onto these weight matrices to get
$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{7}$$
Thus, for any entity in the sequence, the self-attention is basically computed by the dot-product of the query with all keys. Each entity is, thus, the weighted sum of all entities in the input sequence, where the weights are provided by the attention scores
6) The output $Z \in \mathbb{R}^{N \times D}$ is given by
$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{D_q}}\right)V \tag{8}$$
7) The output is therefore normalised using softmax operator to obtain the attention scores [48].
8) MSA is made up of multiple self-attention blocks or heads in order to model the complex relationships between the various elements in the input sequence. Each head or block has its own set of learnable weight

matrices $[W^{(Q_i)}, \quad W^{(K_i)}, \quad W^{(V_i)}]$, where i = 0, 1,2, …(h-1) and h is the number of heads in the MSA. For a given input X, the result of the h self-attention heads in the MSA block is concatenated into one matrix

$$[Z_0, Z_1, \ldots, Z_{h-1}] \in \mathbb{R}^{h \cdot D_v \times N}. \tag{9}$$

This is projected onto a weight matrix,

$$W \in \mathbb{R}^{h \cdot D_v \times D} \tag{10}$$

9) The MLP is made up of two layers with a Gaussian Error Linear Unit (GELU) non-linearity as activation function. As in [13], a class token is prepended to the sequence of embedded patches $z_0^0 =$ xclass whose state at the output of the Transformer encoder (z_L^0) serves as the image representation y (Equation 14). The transformer encoder consists of L stacked base blocks, each one consisting of MSA and MLP blocks (Equations 12 and 13). Before each MSA and MLP block, layer normalisation is used while after each block, residual connections are applied.

$$z_0 = [\text{xclass}; x_p^1 E; x_p^2 E; x_p^3 E; \ldots; x_p^N E] + E_{\text{pos}} \tag{11}$$

where $E \in \mathbb{R}^{D \times (P^2 C)}$, $E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$

$$z_\ell^\ell = \text{MSA}(\text{LN}(z_{\ell-1})) + z_{\ell-1}, \quad \ell = 1, \ldots, L \tag{12}$$

$$z_\ell = \text{MLP}(\text{LN}(z_\ell^\ell)) + z_\ell^\ell, \quad \ell = 1, \ldots, L \tag{13}$$

$$y = \text{LN}(z_L^0) \tag{14}$$

DeiT Architecture

ViT requires extensive datasets and substantial computational resources. To address these issues, [47] introduced the Data Efficient Image Transformer, DeiT based on the ViT model. DeiT not only streamlines VIT's training process but also improves image classification performance by incorporating knowledge distillation techniques. Therefore, DeiT models have also been applied in our work. The architecture of a DeiT is depicted in Fig. 7.

After resizing the input images to a specific fixed size, the images are segmented into N patches. As in our case, N is 14 × 14 and each patch is of 16 × 16 pixels. The patches are reorganised linearly that maintains the image dimension in general to 3 × 16 × 16, since the images are in RGB format. Positional information is added which indicates the positions of the patches relative to each other. This information is added before the initial transformer block, that is afterwards passed through subsequent transformer blocks. A Feed Forward Network (FFN), comprising two linear layers separated by a GELU activation function, is integrated on top of the MSA to form a complete transformer block. MSA as well as FFN function as residual layers because of the skip connections together with layer normalisation. A trainable class vector is added to the patch tokens before the foremost layer. This class vector traverses the transformer layers and, together with a linear layer, is projected for predicting the class. The class token originates from the NLP and is different from pooling layers in computer vision and is used to predict the class.

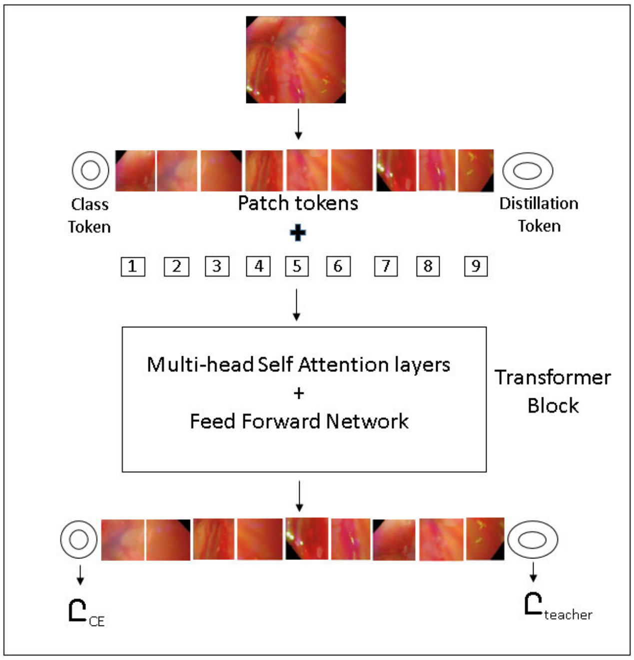


FIGURE 7. DeiT architecture.

The DeiT model makes use of native distillation. The distillation process allows a neural network (the student) to learn from the results of another network, referred to as the teacher. CNN is used as the teacher since it has better priors on images due to inductive bias and hence, can be trained with smaller number of images. A distillation token has been introduced alongside the patches as well as the class tokens. This token interacts with other embeddings through self-attention and is outputted by the network after the final layer. Its aim is to enable the model for learning from the teacher's results meanwhile providing supplementary information to the class vectors. The distillation strategy used by DeiT has proved to enhance the model as compared to vanilla distillation methods [47].

2) IMPLEMENTATION DETAILS

In this research work, two variants of the ViT model, pre-trained on the ImageNet dataset, have been used: ViT-B/16 (ViT_base_patch16_224) and ViT-S/16 (ViT_small_patch16_224). ViT-B/16 refers to the base variant, while ViT-S/16 denotes the smaller variant, both using an input patch size of 16 × 16. ViT-Base consists of 12 transformer blocks, each with a 12-headed self-attention module. In total, this model has 86M trainable parameters [15]. Similarly, ViT-S/16 comprises 12 transformer blocks, each with a 6-headed self-attention module, totaling 21M trainable parameters [14].

The DeiT models chosen are DeiT-Tiny-Patch16-224, DeiT-Small-Patch16-224, and DeiT-Base-Patch16-224, which have also been pre-trained on ImageNet. The motivation behind adopting the ViT and DeiT models is based on the strong performance achieved in prior work [15], [47]. Moreover, ViT and DeiT models have the ability to capture important global features in images using the self-attention mechanism. Given that colour misalignment artefacts are typically spread across several non-contiguous pixels in images, ViT and DeiT are implemented which are designed to grasp long-range interactions between pixels and hence, are capable of finding out the dependencies between pixels in images globally. DeiT-Tiny-Patch16-224 includes 192 embedding dimensions with 3 attention heads, DeiT-Small-Patch16-224 includes 384 embedding dimensions with 6 attention heads, and DeiT-Base-Patch16-224 includes 768 embedding dimensions with 12 attention heads. The number of layers for all three models is the same. Consequently, the number of trainable parameters is 5M, 22M, and 86M for DeiT-Tiny-Patch16-224, DeiT-Small-Patch16-224, and DeiT-Base-Patch16-224, respectively [47].

Table 4 & Table 5 provide a summary of the details of the variables used in the architecture of the ViT and DeiT

TABLE 4. Details on the DeiT models used.

Model	Number of parameters	Embedding Dimension	Number of heads	Number of layers
DeiT-Tiny-patch 16-224	5M	192	3	12
DeiT-Small-patch 16-224	22M	384	6	12
DeiT-Base-patch 16-224	86M	768	12	12

TABLE 5. Details on the ViT models used.

Model	Number of parameters	Input Patch Size	Transformer Blocks	Number of heads
ViT S/16	21M	16 x 16	12	6
ViT B/16	86M	16 x 16	12	12

Moreover, transfer learning is applied to our customized dataset by utilizing the pre-trained weights. Transfer learning involves leveraging features learned from one task and applying them to a different task without starting the training process from scratch. This approach is frequently employed in deep learning model development, as training from scratch requires substantial computational resources, large datasets, and significant time. In our case, the dataset is limited in size, making transfer learning an appropriate choice, as demonstrated in [40], [49], and [50].

In order to preserve the learned representations so that they are not overwritten, the entire network is frozen. We modify and train only the model's head from scratch to adapt it to our task and fine-tune it to enhance its performance. A linear layer followed by an activation function, a dropout layer, and finally an output layer of 2, to match the number of classes in our dataset, are added for the model's head. The hyper-parameters fine-tuned are number of epochs and learning rate. For the model's head, the number of output features of the first linear layer and the probability of the dropout layer are adjusted until best performance is obtained. For each hyperparameter, the algorithm was run 3 times and the average value calculated due to the stochastic nature of our dataset. The optimal parameters of the proposed training model used can be found in Table 6.

TABLE 6. Optimal parameters of proposed training model.

Number of epochs	Learning Rate	Number of output parameters	Probability of dropout layer
100	0.00001	512	0.3

After fine-tuning the hyper parameters, the best results are chosen to build the final model. Rectified Linear Unit (ReLu) has been chosen as activation function since it allows faster training and it has a decreased likelihood for the gradient to vanish [51]. ReLu is basically defined as:

$$\mathrm{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{15}$$

Hence, if the input x is less than or equal to 0, the result is equal to 0, otherwise if the input x is greater than 0, the output is equal to the input [52]. The model is then trained using Adam Optimiser with an initial learning rate of 0.001. Adam Optimiser allows fast convergence and has low memory requirements, hence, allowing fast training even for large networks. It also makes use of adaptive learning rates [53]. LabelSmoothingCrossEntropy is used as the loss function, which is better than the normal CrossEntropyLoss, defined as follows:

$$y_{ls} = (1-\alpha) \cdot y_{\text{hot}} + \frac{\alpha}{K} \tag{16}$$

where one-hot encoded label vector, y_{hot} is replaced by a combination of y_{hot} and the uniform distribution. K represents the number of classes, α defines the amount of smoothing needed [54].

E. CNN IMPLEMENTATION DETAILS

In order to carry out a performance comparison between ViT and CNN based model for colour misalignment artefact detection and classification, four high-performing state-of-the-art CNN architectures have been duly applied on the same dataset. They are ResNet50, VGG16, InceptionResNetv2 and Inceptionv3. Each of the four CNN models has been pre-trained on Imagenet database. The pre-trained weights have been added to our model and the parameters of the latter fine-tuned towards our colour misalignment classification task. VGG16, introduced by [55], is a deep convolutional neural network (CNN) comprising 16 layers, including 13 convolutional layers together with 3 fully connected layers and a softmax classifier. Among the 13 convolutional layers, 5 are for max-pooling. ResNet50, presented by [56], includes CNN with 50 layers, consisting of 48 convolutional layers, a max-pool layer, and an average pool layer. This network is constructed by stacking residual blocks [57]. Inceptionv3, created by Google in 2015, is a cutting-edge CNN model built upon the earlier Inceptionv1. It has 42 layers and produces a lower error rate as compared to its predecessors. This is achieved by factorising larger convolutions into smaller ones and using asymmetric convolutions. The model also employs auxiliary classifiers as regularisers and efficiently reduces grid size [58]. InceptionResNetV2, proposed by [58], is based upon the architecture of Inception. Its stem module includes a double 3×3 convolutional layer and then a max-pooling layer. As for the Inception blocks, to address the degradation problem caused by structures which are deep, residual connections are employed. Reduction blocks allow a decrease in the spatial dimensions.

In this implementation, for loss function, we use categorical entropy and for optimisation, the Adam technique with a learning rate of 0.001 is used. Each CNN model has been trained on the dataset, and their performance has been assessed on the test dataset.

F. XAI IMPLEMENTATION DETAILS

The output of the layers of ViT models typically have the shape $BATCH \times 197 \times 192$. The first element of the 197 element is the class token, thus, the remaining 196 represents the 14×14 patches of the input image. While the class token is used for the final classification, the 196 elements can be considered as a spatial image of size 14×14 together with 192 channels. In order to integrate the different XAI techniques into the ViT model, the image must be reshaped into 2D spatial images. Therefore, a reshape function is passed to the CAM constructor. Since the class token is used to perform the final classification in the last attention layer, the results of the model is not impacted by the 14×14 channels of the last layer and hence, the output gradient is 0. Consequently, any layer before the final attention block has been chosen when producing the class activation maps to understand the behaviour of the model. In addition, the fine-tuned ViT model, with the highest scores, has been loaded into the system, together with the corresponding weights, before processing the images from our test dataset to explain the predictions.

Basically there are two types of methods to develop explainable systems – ante-hoc and post-hoc methods. Ante-hoc techniques incorporate explainability into models from the start while post-hoc ones are applied after the training of the model [34]. From the analysis of the works done relative to explainability for medical images, it is found that GradCAM, a post-hoc explainable method is an efficient and widely used technique [34], [35], [36]. Therefore, our XAI module applies several post-hoc Gradient-based XAI techniques including GradCAM, ScoreCAM, GradCAM++, AblationCAM and XGradCAM to provide the areas used in the input images for prediction. GradCAM makes use of the gradients of a target in the final layer of deep learning models in order to generate a localisation map which highlights the most important areas in an image used to predict the target [59]. ScoreCAM and GradCAM++ are improvements of the GradCAM algorithm which identify the importance of the neurons of the model (Varam et al. 2023). While ScoreCAM makes use of the contribution of each activation map instead of gradients [60], GradCAM++ provides improved localisation of salient areas in images by combining gradients of both positive and negative samples [61]. AblationCAM uses ablation analysis instead of gradients to eliminate gradient saturation issues [62]. On the other hand, XGradCAM utilises a weighted average of gradients to generate class activation maps [63].

G. PERFORMANCE METRICS

For the evaluation of the performance of the ViT model, the metrics used are accuracy, precision, recall, F1 score, Average Precision Score, MCC and AUC. Confusion Matrix also is built to identify whether the model is able to identify the correct images. Accuracy measures how well a model predicts by comparing the predictions of the model with true values [64]. Precision denotes the fraction of predicted instances that have been correctly found.

$$p = \frac{\#\mathrm{TP}}{\#\mathrm{Pred}} \tag{17}$$

where TP represents the number of true positives and $Pred$ is the number of predicted instances.

Recall is the metric that designates the fraction of ground-truth instances that were correctly predicted.

$$r = \frac{\#\mathrm{TP}}{\#\mathrm{GT}} \tag{18}$$

where GT is the number of ground-truth bounding boxes or the number of pixels in image segmentation and TP is the number of true positives.

Average Precision (AP) score is defined as per Equation 19:

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{19}$$

where $P_n$and R_n are the precision and recall at the nth threshold. AP provides a summary of the precision-recall curve and is defined as the weighted mean of precision obtained at each threshold. The increase in the recall value from a previous threshold is used as the weight. To note that this value is different from the area under the precision-recall curve which makes use of linear interpolation and can be excessively optimistic. The Matthews Correlation Coefficient (MCC) measures the quality of binary classifications and is a balance measure which can be applied even when the classes of data are of different sizes. MCC is defined as per Equation 20.

$$\mathrm{MCC} = \frac{tp \times tn - fp \times fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}} \tag{20}$$

where tp, tn, fp *and* fn are the number of true positives, true negatives, false positives and false negatives respectively. F_β scores is a weighted β harmonic means of precision and recall [8].

$$F_\beta = (1+\beta^2) \cdot \frac{p \cdot r}{(\beta^2 \cdot p) + r} \tag{21}$$

IV. RESULTS AND DISCUSSIONS

This section provides the results obtained for the experiments carried out and a discussion on the findings. Figure 8 shows the results obtained on the augmented dataset after applying the ViT and DeiT models. It is to be recalled that before data augmentation, there are 1784 endoscopic images having colour misalignment artefacts and 1866 images having other artefacts in the dataset while after augmentation, there are 6180 images having colour misalignment and 6187 images containing the remaining artefacts. From the results, it is found that after training is performed on augmented data, the performance in terms of the different metrics increases for most models. This indicates that with data augmentation, the models are able to learn the features in the image more efficiently. The DeiT-Base-patch-16-224 model has the best accuracy of 93.46% as well as the best Recall of 93.46%, F1 score of 93.46%, MCC of 86.94% and AUC of 93.48% compared to the other models. This could be attributed to the fact that it does not require extensive amount of data for training as well as its improved training procedure due to the enhanced distillation process. Moreover, there is not much difference in the value of accuracy and other performance metrics of the DeiT-Base-patch-16-226 and ViT-B/16 since both have same architecture which is also confirmed in a research conducted by [47]. The performance of ViT-S/16 also is quite remarkable. The results demonstrate that the attention mechanism of ViT models allows them to better learn the intricate differences in the data patterns in endoscopic images such that the predictions of the images are more accurate. Moreover, the sequential data processing in vision transformers enables enhanced and more efficient parameter utilisation, facilitates the consideration of longer-term dependencies, and fosters broader learning capabilities across diverse inputs.

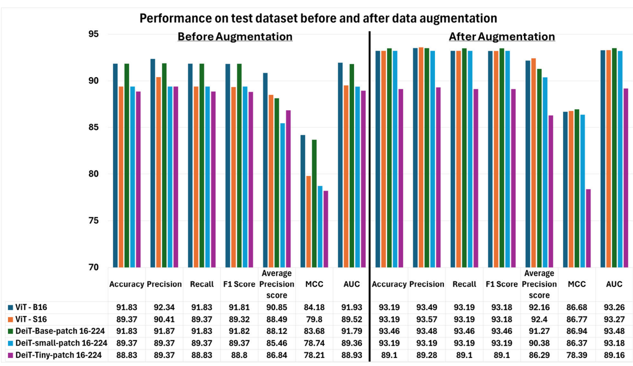
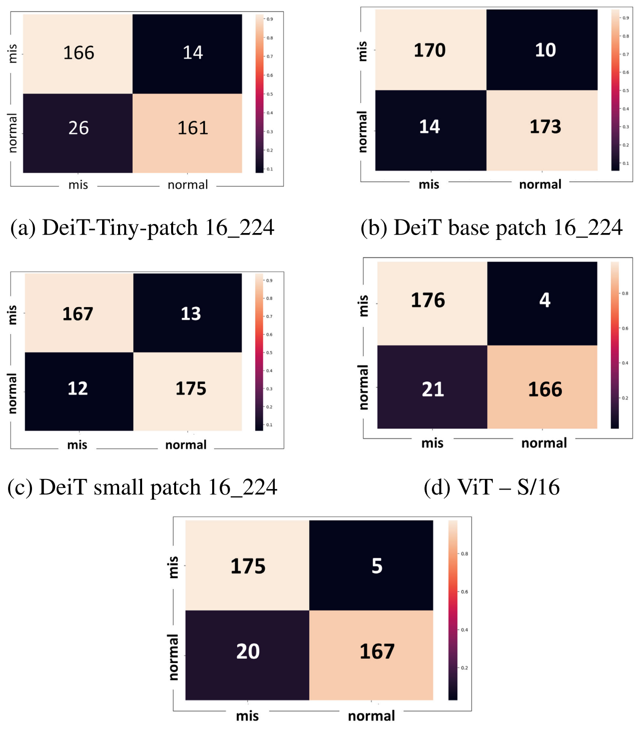


FIGURE 8. Performance results over test dataset before and after data augmentation.

A. CONFUSION MATRIX AFTER DATA AUGMENTATION

The confusion matrix is also generated for the five models when applied to the test dataset after training the models on augmented data. These are illustrated on Figure 9.

A summary of the results from the confusion matrix is provided in Table 7. It is found that ViT-S/16 has the best class accuracy for colour misalignment artefacts, followed by ViT-B/16 and DeiT-Base-patch-16-224. DeiT-Small-patch-16-224 has the best class accuracy for classification of the images with other artefacts, followed by DeiT-Base-patch-16-224 and ViT-B/16. Hence, out of the five models, DeiT-Base-patch-16-224 exhibits a very good accuracy of 94.44% outperforming the other models. This is also confirmed by the AUC of 93.48% obtained. Overall, the DeiT-Base-patch-16-224 model is able to classify both the colour misalignment artefacts as well as the other artefacts quite well. The other models do not perform that well in terms of artefact classification in such a way that only one type can be classified correctly. The foremost reason for the confusion is explained by the quite similar appearance between the colour misalignment artefacts and the other artefacts classes which makes it difficult for the models to differentiate between them. This can also be due to the imbalance in data distribution between the two classes as well as the simple data augmentation techniques applied.



(a) DeiT-Tiny-patch 16_224
(b) DeiT base patch 16_224
(c) DeiT small patch 16_224
(d) ViT – S/16
(e) ViT – B/16

FIGURE 9. Confusion Matrices for different models.

TABLE 7. Confusion Matrix Results - Accuracy obtained for each class.

	Class Accuracy (%)		
Models	Colour Misalignment	Other artefacts	AUC
ViT – S/16	97.78	88.77	93.27
ViT – B/16	97.22	89.3	93.26
DeiT-Base-patch 16-224	94.44	92.51	93.48
DeiT-Small-patch 16-224	92.78	93.58	93.18
DeiT-Tiny-patch 16-224	92.22	86.09	89.15

TABLE 8. Summary of various ViT-based techniques and their performance on different medical datasets.

Authors	Techniques	Dataset	Results
[65]	ViT based method	WCE images from Kvasir dataset	Images Normal – Accuracy: 95.63%, precision: 1.0, recall: 1.0, F1 score: 1.0, Ulcerative Colitis - precision: 1.0, recall: 0.32, F1 score: 0.48, Polyps - precision: 0.61, recall: 0.99, F1 score: 0.76. Esophagitis - precision: 0.97, recall: 1.0, F1 score: 0.99.
[15]	ViT-B/16, ViT-L/16, and ViT-L/32 pre-trained on ImageNet-21k, DeiT-Ti and DeiT-B 384	Three publicly available medical datasets with different modalities - Chest X-ray dataset, Kvasir, Kvasir-Capsule dataset	Chest X-ray: ViT-L/16- F1 score: 0.9532, recall: 0.9533, MCC: 0.9259, accuracy: 0.9533, ROC-AUC score: 0.97. Kvasir dataset ViT-L/16 - F1 score: 0.9436, recall: 0.9437, MCC: 0.9360, accuracy: 0.9437, ROC-AUC score: 0.97. Kvasir-Capsule ViT-B/16 - F1-score: 0.7156, recall: 0.7182, MCC: 0.3705, accuracy: 0.7156, ROC-AUC score: 0.57.
[49]	ViT as the base model and a softmax classifier	Oral cancer histopathological image datasets from Kaggle	ViT- accuracy: 97.78%. Specificity: 96.88%, sensitivity: 98.74%, ROC-AUC score: 97.74%.
[28]	Original ViT with modification of the spatial dimensions in the transformer encoder module.	Kvasir-Capsule dataset, Red Lesion Endoscopy (RLE) dataset	Multi-class Kvasir-Capsule dataset - accuracy: 79.15%, Binary classification RLE - accuracy: 98.63%
[66]	Transformer model based on hybrid shifted windows used	Kvasir v2 and HyperKvasir	Kvasir v2 dataset – accuracy: 95.42% HyperKvasir dataset – accuracy: 86.81%

Continued on next page

TABLE 8. *(Continued.)* Summary of various ViT-based techniques and their performance on different medical datasets.

Authors	Techniques	Dataset	Results
[17]	ViT model - DeiT-S Self-supervised learning method - DINO	APTOS 2019 - diabetic retinopathy images ISIC 2019 - dermoscopic images of skin lesions CBIS-DDSM - mammography images	Classification on APTOS2019 dataset with DeiT- a Cohen Kappa: 89.6; ISIC2019 – Recall: 84.4 DDSM - ROC-AUC: 94.7 Supervised learning, for classification on APTOS2019 - Cohen Kappa: 89.6 ISIC2019 – Recall: 85. DDSM - ROC-AUC: 95.6
[18]	ViT model	Two datasets - breast Ultrasound images	ViT model using transfer learning attained – accuracy: 86.7%, AUC: 0.95.
[27]	COTR - network of convolution in transformer	Publicly available colonoscopic videos CVC-ClinicDB, ETIS-LARIB and CVC-ColonDB.	ETIS-LARIB dataset – Precision: 91.49%, Sensitivity: 82.69%, F1-score: 86.87%; CVC-ColonDB – Precision: 91.67%, Sensitivity: 93.54%, F1-score: 92.60%
[19]	POCFormer – a lightweight ViT	Publicly available ultrasound images of lungs for COVID-19	POCFormer - average accuracy: 91%
[26]	Hybrid Transformer-CNN - GasHisTransformer Inceptionv3 is used as the CNN model	Hematoxylin and Eosin (H&E) stained gastric histopathological image dataset (HE-GHI-DS)	GasHisTransformer – Accuracy: 97.97%, Precision: 98.55%, Recall: 97.38%, F1 score: 97.97%. With the lightweight version – accuracy: 96.43%, Precision: 95.99%, Recall: 96.90% ,F1 score of 96.43%.

B. COMPARISON OF VIT BASED RESULTS WITH CURRENT WORK IN THE FIELD

Our best model obtained is compared with other popular works in classification of artefacts. However, as mentioned in earlier section, there is limited work conducted so far on artefacts in endoscopic images using ViT or ViT based architectures. Therefore, the results of our proposed network are compared with other ViT models applied to classification of GI tract endoscopic images. Table 8 details the techniques deployed by other researchers along with the details of their datasets.

From the performance reported by [15], it is observed that our DeiT-Base-Patch-16-224 model provides better accuracy of 93.46% than the accuracy of 71.56% obtained by the ViT-B/16 model used by the author on the Kvasir-Capsule dataset containing video capsule endoscopic images. This again reiterates the advantage of DeiT models due to the knowledge distillation technique which enhances the training capability of the model. Comparing the value of the accuracy with the results of our ViT-B/16 model, it is found that our fine-tuned ViT-B/16 model has a better accuracy of 93.19% indicating the high performance of our models. Similarly, our DeiT-Base-Patch-16-224 and ViT-B/16 obtain better Recall, MCC and AUC values. This can be attributed to the fact that there is more imbalance between the different classes of the Kvasir-Capsule dataset as opposed to our dataset. Moreover, in the dataset used by [15], there are fewer images in most of the data classes before data augmentation. Likewise, our ViT model has obtained better accuracy than that obtained by [28], that is, 79.15% for disease classification in WCE images, by [66], that is, 86.81% for classification of disease in gastrointestinal endoscopy images and by [18], that is, 86.7% obtained for classification of ultrasound images. This shows that our fine-tuned models have been trained better. This is also explained by the better quality of endoscopic images. In the work conducted by [19] for classification of Covid-19 disease in ultrasound images, an accuracy of 91% is obtained and again our model outperforms it due to the better resolution of the images in our dataset, hence contributing to better learning and training.

Furthermore, contrasting our results with the work by [65], it is found that our Deit-Base-Patch-16-224 model obtains a better recall of 93.46% as compared to 32% and F1 score of 93.46% compared to 48% for classification of Ulcerative Colitis. Similarly, for classifying Polyps in WCE images, our model obtains a better precision of 93.48% as compared to 61% and F1 score of 93.46% against 76% obtained by the author. In addition, our model obtains better precision, recall and F1 score than those of [27] for polyp detection which are 91.49%, 92.69% and 86.87% respectively. In addition, our DeiT-Base-patch-16-224 and Deit-Small-patch-16-224 outperform the DeiT-S of [17] in terms of Recall on APTOS2019 and ISIC2019 datasets. On the other hand, for classification of chest x-ray images, the ViT-L/16 of [15] performs better than our model. Similarly, the model of [26] and [49] outperforms our implementation. However, it is to be noted that the datasets used are not similar. Additionally, endoscopic images are of higher resolution and there are subtle differences between colour misalignment and some other artefacts such as saturation, contrast and specularity. One strategy to improve the performance of our models that could be explored is to perform pre-training on medical images which are more similar to the endoscopic images rather than pretraining on ImageNet database where the features learnt differ from those in endoscopic images.

C. COMPARISON OF VIT PERFORMANCE WITH CNN MODELS

Table 9 provides the values obtained when the four CNN models have been applied to the test dataset as well as a comparison with the VIT results. It is to be noted that on the new data found in the test dataset, it is InceptionResNetV2 which performs the best in terms of accuracy, AUC and Recall. On the other hand, Inceptionv3 performs best in terms of precision and F1 score. This implies that InceptionResNetV2 has a better generalisation with remarkable results over new data. Comparing the results obtained by the best enhanced ViT model, DeiT-base-patch-16-224 with the best CNN based method, InceptionResNetV2, it is seen that the former provides better results with an accuracy of 93.46% compared to 89.37%, a precision of 93.48% against 86.61%, a recall of 93.46% against 89.92% and a F1 score of 93.46% as opposed to 88.23%. As for AUC, InceptionResNetV2 has a better value of 94.4% against 93.48% by our ViT model. This shows that in general the ViT model is able to make better correct predictions (accuracy) which are actual positives (precision). Furthermore, more actual positives are correctly predicted (recall). Similarly, the harmonic mean between precision and recall (F1 score) is better for the ViT model than the CNN model. On the other hand, the CNN model is better able to differentiate between the two classes of data (AUC). InceptionResNetV2 performs better than the other CNN due to its unique architecture which combines the advantages of both inception and ResNet in order to train much deeper networks, hence, exhibiting better extraction of features and helps to avoid overfitting [9]. The main difference between Inception-v3 and Inception-ResNet-v2 lies in the fact that Inception-ResNet-v2 substitutes the filter concatenation phase in the Inception architecture with a residual connection. This enables residual features to be learnt by the network, thereby helping in training by mitigating the vanishing gradients problem [67]. InceptionResNetV2 is able to learn features with differing complexity, due to its distinct design and the fact that filters are used effectively [68]

A comparison is, hereby, provided of the proposed DeiT-base-patch-16-224 model, which has achieved highest performance, with other CNN based state-of-art-models in the classification of artefacts. However, since there is not enough literature on classification of colour misalignment

TABLE 9. Performance comparison of ViT against CNN.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
InceptionResNetv2	89.37	86.61	89.92	88.23	94.4
Inceptionv3	89.1	89.1	89.1	89.1	90.08
VGG16	86.1	82.76	91.55	86.93	94.08
ResNet50	70	84	27	40.86	77
DeiT-Base-patch-16-224	93.46	93.48	93.46	93.46	93.48
ViT - B16	93.19	93.49	93.19	93.18	93.26
ViT - S16	93.19	93.57	93.19	93.18	93.27
DeiT-Small-patch-16-224	93.19	93.19	93.19	93.19	93.18
DeiT-Tiny-patch-16-224	89.1	89.28	89.1	89.1	89.16

artefact in endoscopic images, the results of our proposed network are compared with other models applied to artefact detection and classification as well as classification of GI tract endoscopic images. Table 10 provides details on the techniques applied by other authors in addition to the dataset used and results obtained.

Compared to similar work done by [9] for the classification of colour misalignment artefacts in endoscopic images with an accuracy of 96%, our proposed enhanced model, DeiT-base-patch-16-224 has been able to perform the classification tasks with an accuracy of 93.46%. To note that the dataset used by the author do not have similar images as in our case. Our dataset contains images mostly from normal subjects whereas in the work of [9], colour misalignment artefacts has been detected from endoscopic images of patients potentially with SCC disease. In addition, our ViT model has obtained quite good results in terms of recall, precision and F1 scores while the work of [9] provides only the accuracy results as the performance metrics.

In contrast to the work of [69] to classify artefacts from gastric images, our fine-tuned and customised DeiT-base-patch-16-224 model has achieved higher performance in terms of accuracy(93.46% against 77.65%), precision (93.48% against 75.94%) and recall (93.46% against 77.65%). Moreover, the accuracy of the enhanced DeiT-base-patch-16-224 model is better than the value obtained by the fused method in [70].

Contrary to [71] which has classified motion artefacts and [43] which segmented instrument artefact, the accuracy of our DeiT-base-patch-16- model is less. It is to be noted here also that the dataset used by these authors are different from ours. [43] made use of Kvasir dataset which has images with lower resolution. The InceptionResNetv2 model in [72], the fused method of Inceptionv3 and DenseNet-201 in [73], the VGG16 in [74] and [75] as well as the ResNet50 model of [76] have achieved better accuracy than the corresponding neural network of our enhanced model. But it should be noted that the tasks as well as the dataset utilised are different.

Therefore, CNNs, making use of deep backbones, is able to identify local patterns in images due to the built-in bias. In contrast, ViT models are capable of extracting global features with the Multi-Head Self-Attention modules which improve its ability to focus on important regions within images. Moreover, ViT is able to depict long-range relationships between image pixels which is restrained in CNNs due to the inductive bias. A potential area to gear our study in future would be to integrate both the CNN and ViT models so as to take advantages of the benefits of both models. This is because while ViT models are able to depict dependencies between pixels in images globally, CNN models have a more localised focus on the features.

D. RESULTS OF APPLICATION OF XAI

Given the nature of the feature extraction from ViT, the models are not sufficiently explainable as mentioned previously. Hence, XAI techniques are applied to understand the decision process of ViT. To provide interpretability of the classification results obtained from our fine-tuned ViT models, an additional module, the XAI module, as per Figure 1, has been added to apply XAI techniques in order to display the regions on which the corresponding algorithm has based itself to reach its conclusions. The XAI methods used are GradCAM, ScoreCAM, GradCAM++, AblationCAM and XGradCAM. Figure 10 to Figure13 provide the results obtained when the different XAI methods are applied on some random images from both classes of data classified by the DeiT-Base-patch-16-224 model. Figure 10 and Figure 11 provide the deterministic features for correctly classified images while Figures 12 and Figures 13 illustrates those of misclassified images. The visualisations of the salient regions provide valuable insights into the areas which greatly influenced the decision-making process of the ViT model, thus making the predictions of the latter more transparent and interpretable.

1) CORRECTLY CLASSIFIED SAMPLES

Figure 10 provides the results of correctly classified images having colour misalignment artefacts while Figure 11 provides those for the correctly classified images with other artefacts. In both images, the features which help to identify the particular class has been manually highlighted using dark outlines in the original image. Moreover, the most important regions emphasised by each XAI methods have been highlighted using the dark outlines also.

From all five images in Figure 10, it is found that GradCAM, followed by AblationCAM highlights most of the parts containing the colour misalignment artefacts, suggesting that the model's decisions are influenced by areas relevant to the classification task. As for GradCAM++ and ScoreCAM, in some images the most important regions are emphasised, for example, in the third and fourth images for GradCAM++ and in the second and third images by ScoreCAM while in the remaining images they are not. This possibly indicates a higher sensitivity to local features that are not strongly correlated with colour misalignment artefacts. ScoreCAM and AblationCAM consider more areas in an image as important such as for the first three and last images by ScoreCAM and the second and third images by AblationCAM. However, this makes the visualisation results rather complex since most parts of the images are highlighted. Furthermore, not all parts highlighted by the two algorithms have an impact on the classification results. This could reflect an overestimation of feature importance, leading to

TABLE 10. Details on the techniques and results of other state-of-the-art CNN models.

Authors	Techniques	Dataset used	Results
Gao et al. (2019) [9]	Conventional CNN method, AlexNet, is applied in order to train a binary classifier to determine frames with artefacts or not	350 images from 15 endoscopic videos collected from Oxford NHS University Hospital, UK	The classification accuracy is 96% for determining frames with colour misalignment or not.
Auzine et al. (2022) [69]	Pre-trained CNNs - VGGNet, InceptionV3, and Inception-ResNet applied using transfer learning has been applied to classify artefacts which impede gastric cancer detection.	EAD 2019 and the Kvasir-V2 dataset	InceptionV3 achieves an accuracy of 77.65%, a precision of 75.94% and a recall of 77.65% over InceptionResnetV2 (accuracy of 75.43%, a precision of 75.68% and a recall of 75.31%) and VGG16 (accuracy of 68.02%, a precision of 70.7% and a recall of 63.46%)
Sun et al. (2023) [?]	Cascade R-CNN consisting of a feature extraction network based on ResNeXt101, an FPN, a region proposal network layer, and cascade detectors to detect artefacts in digestive tract are used	Endoscopic image data from 12 medical institutions in Jiangsu Province. 2303 images containing eight types of artefacts were present: blur, bubbles, specularity, saturation, contrast, instrument, blood, and artifacts	Proposed network obtained an AP of 0.37. For each of the artefact, the AP is as follows respectively: Artifact, Blood, Blur, Bubbles, Contrast, Instrument, Saturation, Specularity - 0.318, 0.105, 0.608, 0.242, 0.410, 0.598, 0.333, 0.342
Aubreville et al. (2019) [71]	For detection of motion artifacts automatically in pCLE images, transfer learning from intermediary end points from a pre-trained InceptionV3 network with customised preprocessing	Data set of 15 thousand pCLE images of epithelial tissue acquired in the oral cavity and the vocal folds	ROC curve values of 0.92, accuracy 94.8%
Watanabe et al. (2022) [43]			

Continued on next page

TABLE 10. ***(Continued.)*** **Details on the techniques and results of other state-of-the-art CNN models.**

Authors	Techniques	Dataset used	Results
Aubreville et al. (2019) [71]	The U-Net is an encoder-decoder model consisting of an encoder that extracts features from the input image and a decoder that performs segmentation on the extracted features. Encoder of U-Net changed to one of seven prepared architectures including Inception-ResNet-v2 for segmentation of surgical instruments in endoscopic images	Kvasir-Instrument dataset - 590 endoscopic images contain surgical instruments	The network using Inception-ResNet-v2 as the encoder of U-Net showed the highest values for the four metrics: JSC (0.9167), DSC (0.9501), precision (0.9543), and accuracy (0.9599)
Liu et al. (2020) [70]	A deep neural network structure based on InceptionResNetv2 as the CNN architecture was built. It consisted of 2 streams, the P-stream inputs preprocessed images and extract textual features while the O-stream takes the original images and extracts global features. A concatenation fusion method is used to obtain the final output	A total of 1,272 white-light images were adopted from 748 subjects, including normal cases, premalignant lesions, and cancerous lesions	The proposed structure obtained a recall of 94.23%, a precision of 94.67% and an accuracy of 85.83%. An accuracy of 94.23% was obtained for the classification of the normal type, an accuracy of 82.50% for precancerous lesions and an accuracy of 77.14% was obtained for the cancer type.
Cogan et al. (2019 [72]	To differentiate between eight classes of gastrointestinal landmarks and diseases, 3 pre-trained CNN architecture are used which are Inception-ResNet-v2, Inception-v4, and NASNet	Kvasir dataset	The InceptionResNetv2 obtained an accuracy of 0.9848, Inceptionv4 - 0.9845 and NASNet - 0.9735

Continued on next page

TABLE 10. *(Continued.)* Details on the techniques and results of other state-of-the-art CNN models.

Authors	Techniques	Dataset used	Results
Mohammad and Al-Razgan (2022) [73]	For the classification of stomach diseases in WCE images of the stomach, Inceptionv3 network together with DenseNet-201 has been employed for feature extraction. Binary dragonfly metaheuristic optimisation method is used for feature optimisation.	3000 images of classes bleeding and healthy, stomach gastrointestinal tract images from Kvasirv1 with 500 images and KvasirV2 having 1000 images for classes Esophagitis, polyps, and ulcerative colitis	The proposed fused method of Inceptionv3 and DenseNet-201 obtained an accuracy of 99.8%
Dheir and Naser (2022) [74]	Five CNN models (VGG, ResNet, MobileNet, Inception-v3, and Xception) were retrained to classify eight categories including anatomical landmarks, disease or medical procedures in the Gastrointestinal Tract so as to develop a multi-disease computer aided detection in gastrointestinal tract endoscopic images	Kvasir dataset	Accuracy obtained for VGG, ResNet, MobileNet, Inception-v3, and Xception are 98.3%, 92.3%, 97.6%, 90% and 98.2%, respectively
Yogapriya et al. (2021) [75]	For Gastrointestinal tract disease detection from WCE images, three pretrained CNN based models VGG16, ResNet-18, and GoogLeNet were appplied using transfer learning.	Kvasir version 2 dataset	The highest results were obtained by VGG16 model with 96.33% accuracy, 96.37% recall, 96.5% precision, and 96.5% F1-measure and highest Matthews Correlation Coefficient value of 0.95 and Cohen's kappa score of 0.96

Continued on next page

TABLE 10. *(Continued.)* Details on the techniques and results of other state-of-the-art CNN models.

Authors	Techniques	Dataset used	Results
Kumar and Mubarak (2022) [76]	For the classification of precancerous lesions including Barrett's esophagus and esophagitis, pre-trained deep CNN models (AlexNet, DenseNet201, Resnet50, ResNet101 and Inceptionv3) has been applied in 2 ways, one using traditional classifiers such as SGD, SVM and Random Forest and the other one with transfer learning and softmax classifier.	Kvasir dataset	ResNet50 with SVM classifier achieves an accuracy of 93.5%, recall of 93.5%, precision of 93.4%, f score of 93.5% and AUC of 89.8%. With transfer learning and fine-tuning, ResNet50 model attains improved accuracy of 94.46%, precision 94.46%, f score 94.46%, AUC 96.20%

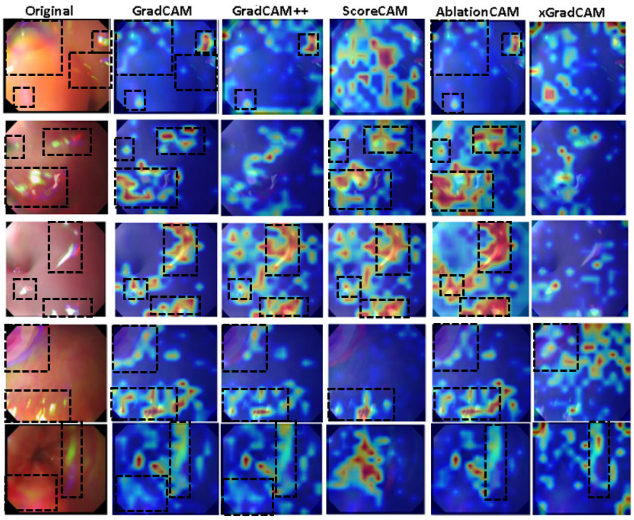


FIGURE 10. Sample of correctly classified colour misalignment artefacts.

explanations that are less precise. In some cases, the dark parts of the images are emphasised as important for example, in the first image by GradCAM and GradCAM++, in the first and second images by ScoreCAM, in the second and third images by AblationCAM and in all five images by XGradCAM. This suggests a possible limitation of the method in accurately localizing discriminative features for this specific task. These findings emphasise the need for careful selection and validation of XAI methods, particularly in tasks involving subtle artefacts. While visualizations can provide insights into model behavior, not all highlighted regions are necessarily indicative of feature importance. Moreover, a method's sensitivity to irrelevant regions can reduce its explanatory value.

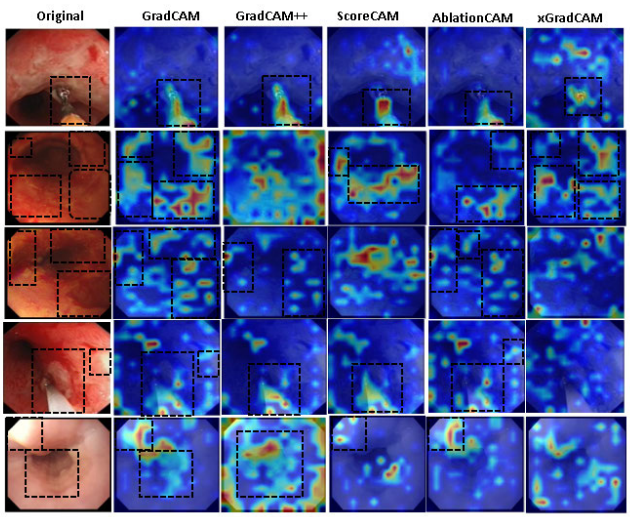


FIGURE 11. Sample of correctly classified other artefacts.

For the images containing other types of artefacts, we find that Deit-Base-patch-16-224 is able to clearly depict the instrument in the first image, the blood in second and third images, the instrument and saturation in the fourth image and saturation in the last image. Here again, GradCAM highlights most of the salient regions which aligns well with the model's focus on relevant features, reflecting its sensitivity to key structures involved in classification. GradCAM++ is highlighting most of the images as important areas for example in the second and fifth image. However, not all highlighted regions have an effect on the classification results. As for AblationCAM and ScoreCAM some of the important regions have been considered but to a lesser extent. Thus, the XAI results emphasises the differences in different visualization methods. While GradCAM excels at capturing the most salient features, it may struggle with fine-tuning its focus, often highlighting non-critical regions. In contrast, AblationCAM and ScoreCAM offer more targeted visualizations, but may miss subtle areas that influence the model's performance.

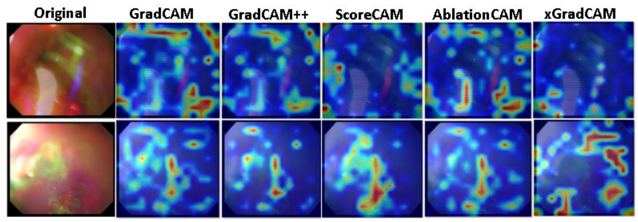


FIGURE 12. Sample of misclassified colour misalignment artefacts.

2) MISCLASSIFIED SAMPLES

For the images in Figure 12, it is observed that not all parts containing the colour misalignment are highlighted by the XAI methods. It is seen that the outputs of GradCAM, GradCAM++ and AblationCAM are quite similar for the first image while for the second image, the outputs are quite similar for all five XAI methods. This suggests that although there are salient regions being depicted by the XAI methods, these do not necessarily correspond to the colour misalignment parts which is the focus of the task.

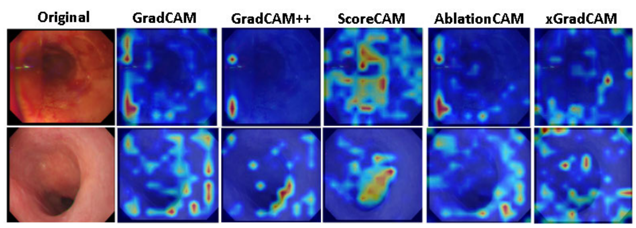


FIGURE 13. Sample of misclassified other artefacts.

From Figure 13, although some parts are highlighted by the XAI methods, they are not enough to contribute to differentiate between colour misalignment and other artefacts. One plausible explanation for the misclassifications in Figure 13 and Figure 13,could be due to the fact that when an image is completely corrupted with colour misalignment, the features highlighted by the algorithm for both classes are quite similar. In such cases, the highlighted areas by the XAI methods may capture global features shared across artefacts, rather than discriminative details specific to colour misalignment. A possible reason for these challenges could be the similarity in pixel patterns between colour misalignment artefacts and other artefacts, which makes it difficult for the XAI methods to clearly delineate between the two.

From Figure 10 and Figure 11, it is observed that with Deit-Base-patch-16-224, the most essential regions in the images are evaluated for classification since the algorithm uses more parameters. References [34], [35], and [36], GradCAM outperforms other XAI techniques to depict areas for enhanced classification. Since GradCAM makes use of back-propagation weights received from the last prediction layer, it extracts visualisations for explanations more efficiently. This is because the weights from the last layer often carries information about essential areas in an image [34]. Therefore, GradCAM is considered as the XAI method that provides the best results in terms of highlighting the important features for most of the ViT models as compared to the other XAI methods used in our study as it better aligns with the model's decision-making process compared to the other techniques explored in the task. GradCAM++, xGradCAM, AblationCAM and ScoreCAM struggle to depict how each of the input features contribute to the prediction done by the model in our case. These methods may be less efficient in isolating the features that are most critical for the classification of colour misalignment artefacts, leading to less precise and sometimes misleading visualizations. One constraint of the GradCAM method is that it faces challenges to localise objects having multiple occurrences in the same class [34]. Alternative XAI methods such as LIME could be used which is a model-agnostic interpretation method, Attention Rollout which is attention-based or TransLRP which is another gradient-based method could be applied. LIME generates locally interpretable models to expound the individual predictions. The attention rollout method describes the attention flow from the previous layer to the following layer of the ViT model while with the TransLRP method, attention scores are merged across the attention graph by amalgamating relevance and gradient details recursively aimed at eliminating negative contributions [77]. SHAP which retrieves the contribution of image pixels using Shapley values [78] could also be implemented. In the same way, ViT Shapley could be used as well whereby Shapley values are made more practical for ViT by utilising an attention masking mechanism [37].

V. CONCLUSION

From the analysis of the literature, it has been found that there is a research gap in the detection and classification of colour misalignment artefact in endoscopic images of the esophagus in comparison with other artefacts. Moreover, our work can be considered as an innovative one as the application of ViT to detect and classify artefacts in endoscopic images has not been explored yet. In this work, different pre-trained ViT as well as CNN models have been fine-tuned and applied using transfer learning for classifying colour misalignment artefact in esophagus endoscopic images. XAI techniques also have been applied to provide insights into the decision process of the models, hence, increasing interpretability and trust in deep learning models. Amongst the CNN based models, InceptionResNetV2 has the best accuracy of 89.37% while DeiT-Base-patch 16-224 has the best performance with an accuracy of 93.46% amongst the ViT models. The results show that CNNs are capable of identifying local patterns in images with noteworthy performance due to the built-in bias while ViT models are better for extracting global features. Moreover, the Multi-Head Self-Attention modules in ViT improve its ability to emphasise important regions within images. The ViT model has outperformed CNN due to the fact that DeiT reduces the requirement for extensive data for training. Moreover, it is able to depict long-range relationships between image pixels which is restrained in CNNs due to the inductive bias. This has been exemplified by applying XAI methods on the ViT model which make it a more robust and transparent one.

For future work, more images with artefacts can be included to have better training and hence, improved performance. In this way a complete dataset can be available for research purposes. There are several artefacts that are introduced during endoscopy of the esophagus. As future work, both CNN and ViT models should be extended to other types of artefacts and compared as whether they are able to classify the other artefacts as well. Currently, only basic data augmentation techniques in the form of geometric transformations have been applied to the dataset for augmentation. GAN, a deep learning data augmentation technique could as well be used to add diversity to he images in our dataset. Moreover, the salient regions highlighted by the XAI methods could be verified by a medical practitioner in the field to ensure that the regions highlighted correctly identifies artefacts instead of GI tract diseases. In the same breath, other ViT models could also be examined such as SWIN transformer, ViT-Large-patch-16-224 and DeiT-Base-patch-16-384 to investigate their appropriateness for artefact detection and classification. Also, other XAI methods could be applied and the results compared such as LIME, Attention Rollout, TransLRP as well as SHAP.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of AIDA Lab, CCIS, Prince Sultan University, Riyadh, Saudi Arabia, for APC support.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] WCRF. *Oesophageal Cancer Statistics | World Cancer Research Fund Internationa*. Accessed: Apr. 5, 2021. [Online]. Available: https://www.wcrf.org/cancer-trends/oesophageal-cancer-statistics/

[3] Healthline. *Esophageal Cancer: Overview, Causes, and Symptoms*. Accessed: Aug. 30, 2025. [Online]. Available: https://www.healthline.com/health/esophageal-cancer

[4] *ASCO, Upper Endoscopy*. Accessed: May 22, 2024. [Online]. Available: https://www.cancer.net/navigating-cancer-care/diagnosing-cancer/tests-and-procedures/upper-endoscopy

[5] E. M. El-Gammal, W. El-Shafai, T. E. Taha, A. S. El-Fishawy, and F. E. A. El-Samie, "A survey of artificial intelligence models for wireless capsule endoscopy videos for superior automatic diagnosis: Problems and solutions," *Multimedia Tools Appl.*, vol. 84, no. 33, pp. 40555–40589, Mar. 2025, doi: 10.1007/s11042-024-18300-1.

[6] S. Ali et al., "Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 102002.

[7] K. Namikawa, T. Hirasawa, T. Yoshio, J. Fujisaki, T. Ozawa, S. Ishihara, T. Aoki, A. Yamada, K. Koike, H. Suzuki, and T. Tada, "Utilizing artificial intelligence in endoscopy: A clinician's guide," *Expert Rev. Gastroenterology Hepatology*, vol. 14, no. 8, pp. 689–706, Aug. 2020, doi: 10.1080/17474124.2020.1779058.

[8] S. Ali, F. Zhou, A. Bailey, B. Braden, J. E. East, X. Lu, and J. Rittscher, "A deep learning framework for quality assessment and restoration in video endoscopy," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101900. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841520302644

[9] W. Malik, R. Javed, F. Tahir, and M. A. Rasheed, "COVID-19 detection by chest X-ray images through efficient neural network techniques," *Int. J. Theor. Appl. Comput. Intell.*, vol. 2025, no. 35, p. 56, Jan. 2025.

[10] S. Ali, F. Zhou, C. Daul, B. Braden, A. Bailey, S. Realdon, J. East, G. Wagniéres, V. Loschenov, E. Grisan, W. Blondel, and J. Rittscher, "Endoscopy artifact detection (EAD 2019) challenge dataset," 2019, *arXiv:1905.03209*.

[11] *EAD2019 EAD2019–Grand Challenge*. Accessed: May 22, 2024. [Online]. Available: https://ead2019.grand-challenge.org/

[12] X. Gao, B. Braden, S. Taylor, and W. Pang, "Towards real-time detection of squamous pre-cancers from oesophageal endoscopic videos," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 1606–1612. [Online]. Available: https://repository.mdx.ac.uk/item/8888x

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[14] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640, doi: 10.1109/ICCV48922.2021.00951. [Online]. Available: https://ieeexplore.ieee.org/document/9709990/

[15] H. Ali, "A meta-review of computational intelligence techniques for early autism disorder diagnosis," *Int. J. Theor. Appl. Comput. Intell.*, vol. 2025, pp. 1–21, Jul. 2025.

[16] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—A contemplative retrospection," *Eng. Appl. Artif. Intell.*, vol. 122, Jun. 2023, Art. no. 106126, doi: 10.1016/j.engappai.2023.106126. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095219762300310X

[17] C. Matsoukas, J. Fredin Haslum, M. Söderberg, and K. Smith, "Is it time to replace CNNs with transformers for medical images?" 2021, *arXiv:2108.09038*.

[18] B. Gheflati and H. Rivaz, "Vision transformer for classification of breast ultrasound images," 2021, *arXiv:2110.14731*.

[19] S. Perera, S. Adhikari, and A. Yilmaz, "POCFormer: A lightweight transformer architecture for detection of COVID-19 using point of care ultrasound," 2021, *arXiv:2105.09913*.

[20] P. Bissoonauth-Daiboo, M. H.-M. Khan, M. M. Auzine, X. Gao, S. Baichoo, and Z. Heetun, "Endoscopic image classification using vision transformers," in *Proc. 7th Int. Conf. Adv. Artif. Intell.*, Oct. 2023, pp. 128–132, doi: 10.1145/3633598.3633623.

[21] A. A. A. El-Latif, S. A. Chelloug, M. Alabdulhafith, and M. Hammad, "Accurate detection of Alzheimer's disease using lightweight deep learning model on MRI data," *Diagnostics*, vol. 13, p. 1216, Jan. 2023.

[22] I. Abunadi, "Deep and hybrid learning of MRI diagnosis for early detection of the progression stages in Alzheimer's disease," *Connection Sci.*, vol. 34, no. 1, pp. 2395–2430, Dec. 2022.

[23] M. F. Ahmad, S. Akbar, S. A. E. Hassan, A. Rehman, and N. Ayesha, "Deep learning approach to diagnose Alzheimer's disease through magnetic resonance images," in *Proc. Int. Conf. Innov. Comput. (ICIC)*, Nov. 2021, pp. 1–6.

[24] T. Yuan, X. Li, H. Xiong, H. Cao, and D. Dou. *Explaining Information Flow Inside Vision Transformers Using Markov Chain*. Accessed: Apr. 4, 2024. [Online]. Available: https://openreview.net/forum?id=TT-cf6QSDaQ

[25] J. Jiang and S. Lin, "COVID-19 detection in chest X-ray images using Swin-transformer and transformer in transformer," 2021, *arXiv:2110.08427*.

[26] J. Chen, X. Li, L. Yu, D. Dou, and H. Xiong. *Beyond Intuition: Rethinking Token Attributions Inside Transformers*. Accessed: Apr. 4, 2024. [Online]. Available: https://openreview.net/forum?id=rm0zIzlhcX

[27] Y. Shen, L. Liu, Z. Tang, Z. Chen, G. Ma, J. Dong, X. Zhang, L. Yang, and Q. Zheng, "Explainable survival analysis with convolution-involved vision transformer," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, pp. 2207–2215, doi: 10.1609/aaai.v36i2.20118. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20118

[28] L. Bai, L. Wang, T. Chen, Y. Zhao, and H. Ren, "Transformer-based disease identification for small-scale imbalanced capsule endoscopy dataset," *Electronics*, vol. 11, no. 14, p. 2747, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/17/2747

[29] J. Wang, S.-Y. Lu, S.-H. Wang, and Y.-D. Zhang, "RanMerFormer: Randomized vision transformer with token merging for brain tumor classification," *Neurocomputing*, vol. 573, Mar. 2024, Art. no. 127216. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223013395

[30] M. Geoda and T. Mousa, "Employing vision transformers for the classification of breast cancer in histopathological images," *Int. Res. J. Adv. Eng. Sci.*, vol. 9, no. 4, pp. 277–280, 2024.

[31] P. Chauhan, M. Lunagaria, D. K. Verma, K. Vaghela, G. Tejani, S. Sharma, and A. R. Khan, "PBVit: A patch-based vision transformer for enhanced brain tumor detection," *IEEE Access*, vol. 13, pp. 13015–13029, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10811909/?arnumber=10811909

[32] Shahid, M. Hayat, A. Raza, S. Akbar, W. Alghamdi, N. Iqbal, and Q. Zou, "PACPs-DNN: Predicting anticancer peptides using novel peptide transformation into evolutionary and structure matrix-based images with self-attention deep learning model," *Comput. Biol. Chem.*, vol. 117, Aug. 2025, Art. no. 108441. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S147692712500101X

[33] S. Akbar, A. Raza, H. H. Awan, Q. Zou, W. Alghamdi, and A. Saeed, "PNPs-CapsNet: Predicting neuropeptides using protein language models and FastText encoding-based weighted multi-view feature integration with deep capsule neural network," *ACS Omega*, vol. 10, no. 12, pp. 12403–12416, Apr. 2025. [Online]. Available: https://pubs.acs.org/doi/10.1021/acsomega.4c11449

[34] D. Varam, R. Mitra, M. Mkadmi, R. A. Riyas, D. A. Abuhani, S. Dhou, and A. Alzaatreh, "Wireless capsule endoscopy image classification: An explainable AI approach," *IEEE Access*, vol. 11, pp. 105262–105280, 2023, doi: 10.1109/ACCESS.2023.3319068. [Online]. Available: https://ieeexplore.ieee.org/document/10262261/

[35] S. I. Nafisah, G. Muhammad, M. S. Hossain, and S. A. AlQahtani, "A comparative evaluation between convolutional neural networks and vision transformers for COVID-19 detection," *Mathematics*, vol. 11, no. 6, p. 1489, Mar. 2023, doi: 10.3390/math11061489.

[36] C. C. Ukwuoma, Z. Qin, M. Belal Bin Heyat, F. Akhtar, O. Bamisile, A. Y. Muaad, D. Addo, and M. A. Al-antari, "A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images," *J. Adv. Res.*, vol. 48, pp. 191–211, Jun. 2023, doi: 10.1016/j.jare.2022.08.021.

[37] I. Covert, C. Kim, and S.-I. Lee, "Learning to estimate Shapley values with vision transformers," 2022, *arXiv:2206.05282*.

[38] J. Maurício and I. Domingues, "Distinguishing between Crohn's disease and ulcerative colitis using deep learning models with interpretability," *Pattern Anal. Appl.*, vol. 27, no. 1, pp. 1–18, Mar. 2024, doi: 10.1007/s10044-023-01206-3.

[39] A. K. Mondal, A. Bhattacharjee, P. Singla, and A. P. Prathosh, "XViTCOS: Explainable vision transformer based COVID-19 screening using radiography," *IEEE J. Transl. Eng. Health Med.*, vol. 10, pp. 1–10, 2022, doi: 10.1109/JTEHM.2021.3134096. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9642985

[40] O. Katar and O. Yildirim, "An explainable vision transformer model based white blood cells classification and localization," *Diagnostics*, vol. 13, no. 14, p. 2459, Jul. 2023, doi: 10.3390/diagnostics13142459. [Online]. Available: https://www.mdpi.com/2075-4418/13/14/2459

[41] Z. M. Lonseko, P. E. Adjei, W. Du, C. Luo, D. Hu, L. Zhu, T. Gan, and N. Rao, "Gastrointestinal disease classification in endoscopic images using attention-guided convolutional neural networks," *Appl. Sci.*, vol. 11, no. 23, p. 11136, Nov. 2021, doi: 10.3390/app112311136. [Online]. Available: https://www.mdpi.com/2076-3417/11/23/11136

[42] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen, and T. de Lange, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Sci. Data*, vol. 7, no. 1, p. 283, Aug. 2020, doi: 10.1038/s41597-020-00622-y.

[43] T. Watanabe, K. Tanioka, S. Hiwa, and T. Hiroyasu, "Performance comparison of deep learning architectures for artifact removal in gastrointestinal endoscopic imaging," 2022, *arXiv:2201.00084*.

[44] X. W. Gao, S. Taylor, W. Pang, R. Hui, X. Lu, and B. Braden, "Fusion of colour contrasted images for early detection of oesophageal squamous cell dysplasia from endoscopic videos in real time," *Inf. Fusion*, vol. 92, pp. 64–79, Apr. 2023, doi: 10.1016/j.inffus.2022.11.023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1566253522002330

[45] *VideoProc Converter–One-Stop Video Processing Software for Windows MAC*. Accessed: Aug. 2, 2023. [Online]. Available: https://www.videoproc.com/

[46] *GitHub*. Accessed: Sep. 4, 2023. [Online]. Available: https://github.com/tejanirla/image_classification/blob/master/transfer_learning.ipynb

[47] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2020, *arXiv:2012.12877*.

[48] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022.

[49] B. S. Deo, M. Pal, P. K. Panigrahi, and A. Pradhan, "Supremacy of attention based convolution neural network in classification of oral cancer using histopathological images," *MedRxiv*, Nov. 2022, Art. no. 22282265, doi: 10.1101/2022.11.13.22282265. [Online]. Available: https://www.medrxiv.org/content/10.1101/2022.11.13.22282265v1

[50] U. Khatri and G.-R. Kwon, "Explainable vision transformer with selfsupervised learning to predict Alzheimer's disease progression using 18f- FDG PET," *Bioengineering*, vol. 10, no. 10, p. 1225, 2023. [Online]. Available: https://www.mdpi.com/2306-5354/10/10/1225

[51] DaemonMaker. *Answer To 'what Are the Advantages of ReLU Over Sigmoid Function in Deep Neural Networks?*. Accessed: Jan. 28, 2024. [Online]. Available: https://tats.stackexchange.com/a/126362

[52] Y. Bai, "RELU-function and derived function review," in *Proc. SHS Web Conf.*, 2022, Paper 02006. [Online]. Available: https://www.shs-conferences.org/10.1051/shsconf/202214402006

[53] N. Vishwakarma. (2023). *What is Adam Optimizer?*. [Online]. Available: https://www.analyticsvidhya.com/blog/2023/09/what-is-adam-optimizer/

[54] W. Wong. *What is Label Smoothing?*. Accessed: Apr. 1, 2024. [Online]. Available: https://towardsdatascience.com/what-is-label-smoothing-108debd7ef06

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[57] Datagen. *ResNet-50: The Basics and a Quick Tutorial*. Accessed: Jul. 31, 2023. [Online]. Available: https://datagen.tech/guides/computer-vision/resnet-50/

[58] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[59] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2019, doi: 10.1007/s11263-019-01228-7.

[60] X. Wang and C. Wang. *Detect Artefacts of Various Sizes on the Right Scale for Each Class in Video Endoscopy*. Accessed: Mar. 15, 2024. [Online]. Available: http://CEUR-WS.org/Vol2366/EAD2019_paoer_4.pdf

[61] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.

[62] K. Ma, P. Xi, K. Habashy, A. Ebadi, S. Tremblay, and A. Wong, "Towards trustworthy healthcare AI: Attention-based feature learning for COVID-19 screening with chest radiography," 2022, *arXiv:2207.09312*.

[63] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based grad-CAM: Towards accurate visualization and explanation of CNNs," 2020, *arXiv:2008.02312*.

[64] M. Riva. *Interpretation of Loss and Accuracy for a Machine Learning Model | Baeldung on Computer Science*. Accessed: Sep. 10, 2023. [Online]. Available: https://www.baeldung.com/cs/ml-loss-accuracy

[65] A. K. M. S. Hosain, M. Islam, M. H. K. Mehedi, I. E. Kabir, and Z. T. Khan, "Gastrointestinal disorder detection with a transformer based approach," in *Proc. 13th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Oct. 2022, pp. 280–285.

[66] W. Wang, X. Yang, and J. Tang, "Vision transformer with hybrid shifted windows for gastrointestinal endoscopy image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4452–4461, Sep. 2023, doi: 10.1109/TCSVT.2023.3277462. [Online]. Available: https://ieeexplore.ieee.org/document/10136531

[67] D. Schumacher. (2023). *Inception-ResNet-v2*. [Online]. Available: https://serp.ai/inception-resnet-v2/

[68] M. Neshat, M. Ahmed, H. Askari, M. Thilakaratne, and S. Mirjalili, "Hybrid inception architecture with residual connection: Fine-tuned inception-ResNet deep learning model for lung inflammation diagnosis from chest radiographs," 2023, *arXiv:2310.02591*.

[69] M. M. Auzine, P. Bissoonauth-Daiboo, M. H. Khan, S. Baichoo, X. Gao, and N. G. Sahib, "Classification of artefacts in endoscopic images using deep neural network," in *Proc. 3rd Int. Conf. Next Gener. Comput. Appl. (NextComp)*, Oct. 2022, pp. 1–5.

[70] G. Liu, J. Hua, Z. Wu, T. Meng, M. Sun, P. Huang, X. He, W. Sun, X. Li, and Y. Chen, "Automatic classification of esophageal lesions in endoscopic images using a convolutional neural network," *Ann. Transl. Med.*, vol. 8, p. 486, Apr. 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7210177/

[71] M. Aubreville, M. Stoeve, N. Oetter, M. Goncalves, C. Knipfer, H. Neumann, C. Bohr, L. Stelzle, and A. Maier, "Deep learning-based detection of motion artifacts in probe-based confocal laser endomicroscopy images," *Int. J. CARS*, vol. 14, pp. 31–42, Jan. 2019, doi: 10.1007/s11548-018-1836-1.

[72] T. Cogan, M. Cogan, and L. Tamil, "MAPGI: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning," *Comput. Biol. Med.*, vol. 111, Aug. 2019, Art. no. 103351, doi: 10.1016/j.compbiomed.2019.103351.

[73] F. Mohammad and M. Al-Razgan, "Deep feature fusion and optimization-based approach for stomach disease classification," *Sensors*, vol. 22, no. 7, p. 2801, Apr. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/7/2801

[74] I. M. Dheir and S. S. Abu-Naser, "Classification of anomalies in gastrointestinal tract using deep learning," *Int. J. Academic Eng. Res. (IJAER)*, vol. 6, no. 3, pp. 15–28, Mar. 2022.

[75] J. Yogapriya, V. Chandran, M. G. Sumithra, P. Anitha, P. Jenopaul, and C. Suresh Gnana Dhas, "Gastrointestinal tract disease classification from wireless endoscopy images using pretrained deep learning model," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–12, Sep. 2021. [Online]. Available: https://www.hindawi.com/journals/cmmm/2021/5940433/

[76] C. Kumar A and D. M. N. Mubarak, "Classification of early stages of esophageal cancer using transfer learning," *IRBM*, vol. 43, no. 4, pp. 251–258, Aug. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1959031821001172

[77] P. Komorowski, H. Baniecki, and P. Biecek, "Towards evaluating explanations of vision transformers for medical imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 3726–3732.

[78] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4768–4777. [Online]. Available: https://papers.nips.cc/paperf

PREETI BISSOONAUTH-DAIBOO received the B.Sc. degree in computer science from the University of Mauritius, in 2004, and the M.Sc. degree in computer science from the University of Bath, U.K., in 2007. She is currently pursuing the Ph.D. degree with the Department of Software and Information Systems, University of Mauritius. Her research interests include artificial intelligence, computer vision, and medical image processing.

MUHAMMAD MUZZAMMIL AUZINE received the B.Eng. degree in mechatronics engineering from the University of Mauritius, in 2021, where he is currently pursuing the Ph.D. degree in artificial intelligence in medical imaging. He is also a Research Assistant with the University of Mauritius. His research interests include artificial intelligence, computer vision, and medical image analysis, with a particular focus on explainable deep learning methods for disease diagnosis.

MUHAMMAD I. KHAN received the master's degree in engineering from the University of Warwick, U.K., and combines technical expertise with business insight. His research focuses on machine learning, artificial intelligence, and the Internet of Things (IoT), targeting smart systems, automation, and data-driven business solutions. He emphasizes practical AI/ML deployment in industrial, healthcare, and IoT contexts to improve efficiency, security, and scalability. Passionate about AI-driven business transformation, he contributes to the tech community through research, projects, and innovation. Currently he is associated with AIDA Laboratory, School of Engineering, University of Warwick, Coventry, U.K., as a Researcher.

FATIMA ALSHANNAQ received the B.Sc. and M.Sc. degrees in computer information systems from Yarmouk University, Jordan, in 2006 and 2009, respectively, and the Ph.D. degree in computer science from the University of Jordan, in 2022. She has served as an IT Instructor at Princess Nourah bint Abdulrahman University, Riyadh Saudi Arabia, and as an Instructional Designer at Jordan University of Science and Technology, Jordan. She is currently an Assistant Professor with the Department of Computer Science, Prince Sultan University, Riyadh Saudi Arabia. She has published in peer-reviewed journals and international conferences, and is involved in collaborative research focused on Arabic NLP and ethical AI systems. Her research interests include evolutionary computation, machine learning, and natural language processing, particularly in the analysis of social media and online communities.

TANZILA SABA is currently a Professor with Prince Sultan University, Riyadh, Saudi Arabia, where she teaches courses related to data analytics and machine learning, and bioinformatics. She is working in medical imaging, and cancer diagnosis has had a profound impact on the field, advancing both the technology and its application in cancer diagnosis. Her dedication to research and education continues to inspire and drive progress in these critical areas of healthcare. She is also with the AIDA Research Laboratory, as the Leader, that focuses on developing innovative solutions for cancer diagnosis, involving the creation of advanced diagnostic tools that can detect cancer at early stages using AI tools. She has worked on various types of cancer, including breast cancer, lung cancer, and skin cancer. She has published numerous research papers in high-impact journals and has presented her findings at international conferences, with more than 400 publications that have attracted 18 000 citations and 80 H-index. Her publications are widely cited and have contributed to advancements in health risks diagnostics using AI innovative models. Additionally, she has received several awards for her outstanding contributions to the field, including research excellence awards from Prince Sultan University. She is also leading and involved in funded research projects. She is a renowned expert in the field of medical imaging and cancer diagnosis. Her extensive research and contributions have positioned her as a leading figure in these areas.

XIAOHONG GAO is currently a Full Professor of computer vision and imaging with Middlesex University London, London, U.K. Her research interests include the application of machine learning and deep learning to medical applications. She has worked on early detection of oesophagus cancer in the last four years and has been taken part a few competitions for detection of lesions and artefacts from oesophageal videos.

MALEIKA HEENAYE-MAMODE KHAN (Senior Member, IEEE) received the B.Sc. degree (Hons.) in computer science and engineering, in 2006, and the Ph.D. degree in computer science and engineering from the University of Mauritius, in 2015. She received a scholarship from the Tertiary Education Commission (TEC), Mauritius, in 2007, to purse the Ph.D. degree. She is currently an Associate Professor with the Department of Software and Information Systems, University of Mauritius. She is the Dean of the Faculty of Information, Communication and Digital Technologies. She has 13 years of experience in academia and research. Her research interests include computer vision, medical image processing, artificial intelligence, biometrics, and data analytics. She is an active reviewer of various reputed international journals in her research areas. She has secured various research funds at the national and international levels. She has also initiated the setting up of a specialized Artificial Intelligence and Data Analytics (AIDA) Laboratory, University of Mauritius.