IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received 9 September 2023, accepted 20 September 2023, date of publication 25 September 2023,
date of current version 29 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3319068

RESEARCH ARTICLE

Wireless Capsule Endoscopy Image Classification: An Explainable AI Approach

DARA VARAM[1], ROHAN MITRA[1], (Member, IEEE), MERIAM MKADMI[1], RADI AMAN RIYAS[1], DIAA ADDEEN ABUHANI[1], SALAM DHOU[1], (Member, IEEE), AND AYMAN ALZAATREH[2]

[1]Department of Computer Science and Engineering, American University of Sharjah, Sharjah, United Arab Emirates
[2]Department of Mathematics and Statistics, American University of Sharjah, Sharjah, United Arab Emirates

Corresponding author: Salam Dhou (sdhou@aus.edu)

This work was supported in part by the Open Access Program from the American University of Sharjah under Award OAPCEN-1410-E00194.

ABSTRACT Deep Learning has contributed significantly to the advances made in the fields of Medical Imaging and Computer Aided Diagnosis (CAD). Although a variety of Deep Learning (DL) models exist for the purposes of image classification in the medical domain, more analysis needs to be conducted on their decision-making processes. For this reason, several novel Explainable AI (XAI) techniques have been proposed in recent years to better understand DL models. Currently, medical professionals rely on visual inspections to diagnose potential diseases in endoscopic imaging in the preliminary stages. However, we believe that the use of automated systems can enhance both the efficiency for such diagnoses. The aim of this study is to increase the reliability of model predictions within the field of endoscopic imaging by implementing several transfer learning models on a balanced subset of Kvasir-capsule, a Wireless Capsule Endoscopy imaging dataset. This subset includes the top 9 classes of the dataset for training and testing. The results obtained were an F1-score of 97%±1% for the Vision Transformer model, although other models such as MobileNetv3Large and ResNet152v2 were also able to achieve F1-scores of over 90%. These are currently the highest-reported metrics on this data, improving upon prior studies done on the same dataset. The heatmaps of several XAI techniques, including GradCAM, GradCAM++, LayersCAM, LIME, and SHAP have been presented in image form and evaluated according to their highlighted regions of importance. This is in an effort to better understand the decisions of the top-performing DL models and look beyond their black-box nature.

INDEX TERMS Deep learning, explainable AI, gastrointestinal diseases, machine learning, vision transformer, wireless capsule endoscopy.

I. INTRODUCTION

Wireless capsule endoscopy (WCE) is a popular procedure within the medical domain due to its non-invasive nature in screening the digestive tract. It aids in the early detection and classification of diseases and potentially cancerous cells within the intestinal tract and in stunting the development of high-risk illnesses [1]. Many indicators can be classified as pre-cancerous, as they have not yet developed into a state of malignancy. The removal of these abnormalities before their transformation into malignancy can significantly reduce the risk of ailment. However, examinations of the colon and digestive tract are usually a lengthy and invasive process, which is the main advantage that comes from WCEs. With WCEs, images can be collected and screened remotely from within the intestinal tract, and diagoses can be drawn regarding the state of a patient's intestinal tract. The accurate diagnosis and classification of these endoscopies is critical for medical diagnosis.

With the development of new technologies in the field of medical image processing, a recent rise in the use of Machine learning (ML) and specifically Deep Learning (DL)

The associate editor coordinating the review of this manuscript and approving it for publication was Cristian A. Linte.

applications has paved the way for researchers to explore novel techniques in abnormality detection and classification [2]. The improvements in Computer-Aided Diagnosis (CAD) through advances in ML algorithms has proven substantial for effective diagnosis. The utilization of CAD in cancer diagnosis has allowed physicians to serve as an additional level of confirmation. Therefore, following a recent rise of DL in medical image analysis, artificially intelligent models can aid in the classification of abnormalities for the prevention of colorectal cancer.

Previous iterations of CAD models were viewed as inefficient due to their heavy emphasis on image feature extraction. This slowed down potential advancements in visual classification, since the models look for visual features of the image as opposed to specific attributes of objects within the image. In other words, by focusing on extracting the features of an image, such ML algorithms fail to characterize the patterns that allow for classification. To overcome this, DL algorithms employed convolutional operations within hidden layers, overruling the need for traditional feature extraction. In addition to significantly increasing the efficiency of these artificially intelligent models, DL algorithms also outperformed traditional ML models, making them the alternative for functional usage, including medical image processing applications.

A popular DL framework is that of the Convolutional Neural Network (CNN). As opposed to traditional feature extraction mechanisms, CNNs learn features in an automatic fashion and can classify from a range of diverge images accordingly. For this reason, CNNs are favorable to traditional ML algorithms specifically for CAD applications. The current state-of-the-art sees optimized DL algorithms tailored specifically for medical image processing (DL-CAD), but the extent of its utilization is still limited due to the concerns of unreliability when the system is put into place [3]. It is, however, important to note the benefits of employing DL-CAD systems for the detection and classification of gastrointestinal (GI) tract diseases.

The primary limitation faced when attempting to train such deep learning algorithms for medical applications is the need for large-scale datasets. This proves to be difficult within the field of medicine, as the collection of reliable, properly labeled data can be challenging and time-consuming. Similarly, unlike other fields, the use of simulated and synthetic data is unreliable, as precision and the mitigation of risks are of utmost importance in the field [2].

The focus of this paper is the evaluation of several novel DL architectures for the classification of gastrointestinal diseases and causes for concern in an effort to prevent and stunt the development of colorectal cancer. The data that is used is the Kvasir-capsule dataset, which is publicly accessible [4]. The dataset contains images gathered from Wireless Capsule Endoscopies (WCEs). In this work, multitude of available DL models are evaluated to come to a conclusion on the models, parameters and metrics that work best. Then, several ML models are evaluated to get a baseline understanding of how classification is conducted on the dataset considered in this work. In addition to the CNN-based models, novel transformer-based models are considered, namely, the Vision Transformer (ViT). Upon presenting a complete analysis of the different DL models used, the dataset and the models are analyzed by applying clustering algorithms and evaluating how well the models cluster unseen data.

The contributions of this paper are summarized below:

1) Apply popular DL architectures on a balanced sample of the Kvasir-capsule image dataset, including the Vision Transformer, which has not previously been implemented on this data. This is in an effort to comparatively evaluate the best-performing models on our data. To the best of our knowledge, this dataset has not been studied previously in the literature, with the most similar study being on video frame classification of the same nature;
2) Apply, analyze, and compare five Explainable AI techniques to the best-performing classifiers, including GradCAM, GradCAM++, LayerCAM, LIME and a SHAP-based algorithm;
3) Use the interpretations of the XAI models to further evaluate how classifiers detect the existence of different anomalies within the Kvasir-capsule dataset, and by consequence within the field of GI disease detection through WCEs.
4) Aggregate the results of classification, the verification of the best model's performance and the XAI model interpretations to come to a comprehensive understanding of the state-of-the-art WCE classification.

Fig. 1 demonstrates the proposed framework used in this study. We will begin by taking the captured video frames from the capsule (using the Kvasir-Capsule dataset). Then, the images will go through several pre-processing techniques (further highlighted in section III-A). The data will then pass through a deep learning model before being passed through an XAI framework.

The rest of the paper is structured as follows. Section II presents a literature review of the previous research conducted in the field, including classification, object detection and segmentation. The review further elaborates on the Explainable AI approaches done on similar datasets including WCE images. Section III provides an overview of the dataset, models, and the three variants of Explainable AI used in this work. The results and discussion are presented in Section IV, which is divided into two parts: part (A) that displays the results for the models used and their evaluation metrics, and Part (B) that goes into detail regarding the Explainable AI techniques and the features/ regions of importance highlighted by XAI. Finally, the future works and limitations are discussed in the conclusion of the paper.

II. LITERATURE REVIEW

Although a large variety of gastrointestinal diseases exist, a large subset of the literature dedicated to their classification and diagnosis is built on polyp classification. Polyps are a

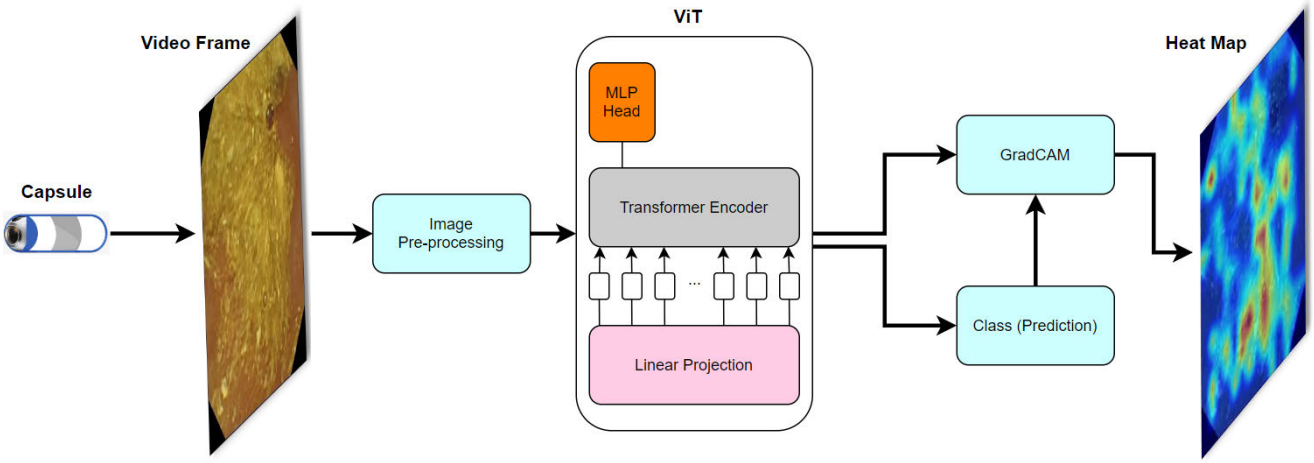


FIGURE 1. The proposed system model demonstrating the process of collecting data through the endoscopy capsule, applying machine learning approach for image classification, and generating heat maps through an appropriate explainable AI (XAI) framework.

significant indicator of colorectal cancer, but they are not the only indicator of potentially harmful diseases within the GI tract. In this section, the classification of GI tract diseases is covered, along with an overview of recent studies done using object detection and segmentation. The review also considers studies performed on datasets that are either similar to the one employed in this study, or the same.

The majority of gastrointestinal indicators are non-malignant. However, hyperplastic and tubular adenomas are also prone to turning cancerous if not treated for an extended period of time. For this reason, their detection, and their further classification, is of significant importance in the diagnostics domain for GI diseases [5], [6].

A. GASTROINTESTINAL DISEASES AS A CLASSIFICATION PROBLEM

The classification of this scope of GI diseases is a deep learning problem that has been tackled in the past. In [7], the authors analyzed a binary dataset extracted from videos of endoscopic examinations. A total of 1, 800 images were collected and classified as adenomatous or non-adenomatous, with each image being of size 256 × 256 pixels. The results when a CNN was trained on the data indicated an accuracy of 0.751 across 10 folds of cross validation. The authors suggest further studying the automated classification of polyps through CAD systems. However, it is important to note the limitations of such a study, given that they treat the classification of polyps as binary as opposed to separating them into different classes. Furthermore, the total number of images in this study were 1, 800, which can be expanded upon to produce more training data. The authors also opted to develop a custom neural network architecture as rather than using a transfer learning model. A similar binary classification approach to polyp detection can be seen in [8], in which polyps are classified as either ''malignant'' or ''non-malignant.'' The total size of the dataset is 600 images, collected from a cohort of 142 patients. However, the original size of the dataset was increased through image augmentation to 3, 600 images. The authors used both their own architecture and also comparatively used the VGG-16 transfer learning model to come to a baseline regarding the results of the polyp classifier. Their results were based on the use of fine-tuning parameters, along with using VGG-16's feature extraction layers instead of the base CNN model. The results, as reported by the authors, are summarized below in Table 1. The best-performing results reported were that of the fine-tuned model with an F1-score of 0.83. The authors described the fine-tuning process of the model as consisting of using a VGG filter, freezing the top layers of the VGG network and using the RMSprop optimizer. The learning rate used for this model in specific was 2×10^{-5}.

TABLE 1. Summary of results in [8], indicating that the best-performing model was the fine-tuned model presented by the authors.

	Accuracy	Precision	Recall	F1-Score
Base CNN model	0.76	0.83	0.66	0.74
VGG-16	0.80	0.84	0.73	0.78
Fine-tuned model	0.83	0.81	0.86	0.83

Similarly, [9] presents a dual-path CNN that classifies colonoscopy images as either polyp or non-polyp. The structure of the proposed method is as such: A colonoscopy image is taken, put through an image enhancement layer (pre-processing), and the features are extracted using the dual-path CNN. Finally, these images are classified into the binary categories. The image enhancement algorithm used takes the images and transforms them into the hue, saturation, and value (HSV) color space. Then, with the help of a Gaussian function, the V value is extracted from the color space, and the brightness is corrected using Gamma correction. The images are then converted back into the RGB color space using image fusion. These images are then fed into the dual-path CNN, consisting of 8 layers. The data was trained on the CVC-ClinicDB, and tested on two other unseen datasets, namely

CVC-ColonDB and ETIS-Larib. The authors were able to report an F1-score of 0.9960 on the former, and an F1-score of 0.9100 on the latter.

Fine-tuning of the models are, however, a significant indicator of better results. In [10], Younas and colleagues presented an ensemble learning model for polyp classification, with their main objective being the optimization of the hypertuning parameters. The models were trained on the PICCOLO dataset [11], discussed later in this section. Another publicly available dataset was used [12], containing 76 images. With three classes (serrated polyps, hyperplastic polyps, adenomatous polyps), the data was first augmented and oversampled to increase from the initial 3, 433 data-points available in the dataset. Authors presented a comparative study of 6 transfer learning models, which included GoogLeNet, ResNet-50, Inception-v3, Xception, DenseNet-201, and SqueezeNet. Based on the performance of these transfer learning models, the best performing model was then selected to conduct further experimentation, namely with regards to the hyper-tuning parameters, including the learning rate and the optimizers used.

In [13], Zachariah et al. and authors presented a pre-trained CNN using ImageNet. The feature extraction layers of this proposed CNN model are based on the Inception-ResNetv2 algorithm. In total, 6, 223 images were used for training and testing, classified into three categories of polyps. The authors used 5-fold cross validation and obtained a negative predictive value of 0.97 for unseen data, and an overall surveillance concordance of 0.94.

In [14], researchers studied the problem of polyp classification by applying a dataset to the family of ResNet transfer learning models. These include the following 5 models: ResNet-18, ResNet-34, ResNet-50, ResNet-101 and ResNet-150. The models were trained on a dataset consisting of four classes: Tubular adenoma, villous adenoma, hyperplastic polyps, and sessile serrated adenoma. The dataset contains whole-slide images - a total of 487 slides. These slides were divided using a sliding-window approach to classify the images. This study comparatively evaluated the performance of the ResNet family of models against annotated images by pathologists. Whilst the pathologists accurately identified the type of polyp with 0.914 accuracy, the ResNet family of models obtained a mean accuracy of 0.935. Similarly, another paper using a whole-slide image dataset is presented in [15], where a total of 2, 074 images were collected through sliding windows, split across five classes of polyps. The authors similarly used a transfer learning approach, using the following architectures: AlexNet, GooLeNet, VGG and the ResNet family of models, including ResNet-50, ResNet-101, and ResNet-152 (two 152 layer models were used, with different projection mappings). Evaluating the models based on the accuracy, precision, recall, and F1-score, it was reported that ResNet performed best out of all, with F1-scores of 0.93, 0.897, 0.883, and 0.888 respectively for each of the four ResNet models tested.

In addition, DL techniques have been employed in the detection and classification of hemorrhages, another indicator of gastrointestinal diseases. In particular, a study conducted in 2016 on the detection of internal bleeding within the gastrointestinal tract produced a DL model trained on a dataset of 10, 000 images [16]. These images, obtained through WCEs, produced a neural network capable of achieving an F1-score of up to 0.9955 in the best-trained instance of the model. This model was built based on the CNN architecture proposed by [17] in 2016 and was able to out-perform previous iterations of deep learning models for the detection of internal bleeding within the gastrointestinal tract [18], [19]. However, the proposed solution is binary in nature - classifying images as either containing internal bleeding or "Normal." A similar study was conducted in 2017 where several transfer learning models were implemented on a dataset containing 12, 090 images adopted from WCEs - 390 of which containing internal bleeding, and 11, 700 of them not [20]. The transfer learning models used in this study were LeNet, AlexNet, GoogLeNet, and VGG-Net. The authors then tested the CNN models on both the original testing images and also an augmented set. The augmentation process involved rotating the images, changing brightness (luminance), and adding noise to the images through blurring and Poisson noise. Their final metrics are reported in Table 2:

TABLE 2. Summary of results in [20], reporting F1-score.

	LeNet	AlexNet	GoogLeNet	VGG-Net
Augmented Images	0.8766	0.9854	0.9770	**0.9887**

In a survey published in 2023, several prior approaches to WCE image classification is discussed as a machine learning problem [21]. The authors summarize the available public datasets, which includes the Kvasir dataset used heavily within the scope of our study. The authors state the lack of publicly available datasets as one of the challenges that researchers face in the field. Other challenges include the nature of the data itself, since certain anomalies are difficult to distinguish from one another, and finally, the reliability of the diagnoses through the machine learning systems. The authors go on to cite the use of Explainable AI for increased trust in back-box models, which is extensively discussed within the scope of our study. Wahab and researchers conclude that the promise of WCE imaging for GI disease diagnosis and localization allows for significant advancement in the field, with a view towards the incorporation of more ML-based systems to aid physicians.

B. KVASIR DATASET(S)

In general, the detection of lesions and diseases within the gastrointestinal tract comes in a few forms, including object detection, segmentation, and classification. Most notably, in [22], a CNN-based model was developed based on the Kvasir dataset, which is an open-source dataset containing images of lesions, polyps, and other indicators of

gastrointestinal diseases through Wireless Capsule Endoscopies (WCEs).

More recently, [23] presented an automatic DL-based hemorrhage detection system in which they trained their own CNN on the Kvasir dataset. This dataset contains 4, 778 images, split into ''Normal'' and ''Bleeding'' classes. By considering different parameters for batch size, number of epochs, and optimizers, the authors were able to draw conclusions on what combinations of parameters work best for the Kvasir dataset. The authors reported that the best parameters for the dataset was a batch size of 32, running for 10 epochs using the ''Adam'' optimizer. However, this study was not able to outperform the study conducted by Sharif et al. and colleagues [24], where the authors used a private dataset obtained through WCE imaging containing three classes for the detection of gastrointestinal diseases. The classes were labeled ''Bleeding,'' ''Healthy,'' and ''Ulcer.'' This study presented the novel use of geometric feature extraction for initially segmenting the lesions within the image and then using that for classification. Their best results were achieved using a KNN classifier, with a classification accuracy of 0.9942 and a precision rate of 0.9951.

In 2023, Padmavathi et al. and researchers published a study on the segmentation and classification of WCE images, specifically using the Kvasir-V2 dataset [25]. For the classification, the authors propose the use of a DL model based on the LeNet-5 architecture. However, the classification is done based on the features extracted from the segmentation of the images. This feature extraction is done through the use of DeepLapV3+. When tested on the Kvasir-V2 dataset, the authors were able to report an F1-score of 98.49%, improving upon prior approaches cited in the work.

C. GRADIENT EXPLANATION TECHNIQUES

Given that deep learning methods are usually presented in a black-box fashion, a significant portion of information is lost regarding the decisions made by the network in coming to a final classification. This has given rise to modern Gradient Explanation techniques, as presented in [26] and [27]. These techniques explain decisions made by DL models using heat maps. An example of a gradient explanation technique is GradCAM [28], which uses gradient information from the final convolutional layer of a CNN to determine the importance of different neurons in the final category assigned by the network. However, a criticism of GradCAM is its shortcomings when localizing multiple instances of an object in a singular image. This led researchers to the development of an improved technique, HiresCAM, which addresses this particular limitation [29]. Other proposed explainable AI techniques exist, such as Score-CAM [30] and Grad-CAM++ [31], which both build on the idea presented by Grad-CAM with improvements to the algorithm used in identifying neuron importance. These explainable techniques often provide researchers with visualizations as to why certain classifiers segment images in the way that they do, which significantly improves the interpretation of the classification. This allows us to see exactly which features the CNN model uses to come up with its decision, highlighting the important regions identified.

In [32], Mukhtorov et al. and authors presented a classification method for endoscopic images based on Explainable Deep Learning. In their 2023 study, they combined a ResNet-152 architecture with GradCAM to obtain the highest accuracy reported of 0.9346 when trained and tested on the Kvasir dataset. The authors undersampled the data and split it as follows: the training data contained 8 balanced classes of 800 image instances each (for a total of 6, 400 images for training), 800 images for validation, and a final 800 images for testing. They also presented their a custom CNN model. The authors were also able to interpret and study the important features of the images based on the different regions identified when applying different Gradient Explanation techniques, including GradCAM, GradCAM++, LayerCAM, HiresCAM, and XGradCAM. The authors were able to out-perform prior classifiers that were trained on the Kvasir dataset, going a step beyond to include XAI interpretations within their work. This includes works conducted in 2019 by Fonollá et al. and colleagues [33] with a highest recorded accuracy of 0.9020 and Pozdeev et al. [34] with an accuracy of 0.88.

Despite the significant advances with DL models and their performance, we still face challenges with regards to how reliable model predictions are within the scope. In [35], Dhar et al. and researchers discuss one of the main issues of the use of DL in the medical field: The lack of trust and explainability of deep learning models to perform accurate diagnosis. The authors suggest a higher degree of transparency within DL models and architectures without having to compromise data privacy, which can be addressed through Explainable AI. The authors conclude with a view towards the co-existence of AI and humans, with a restored awareness and trust for models' predictions.

Within the same domain, we can see an increase in the use of Explainable AI for medical classification and diagnosis. In [36], a new XAI framework is presented for use in ultrasound image classification. The XAI model is trained on an image dataset containing 19341 ultrasound images of the thyroid, annotated by physicians for the existence of anomalies. Similar to what will be presented in this paper, GradCAM has been used as the explainer. More specifically, the authors propose the use of XAI within the medical domain for its incorporation in the ultrasound screening process. This is because of their provision of heatmaps which provide explanations for the classifier's predictions.

As presented in sections II-B and II-C, several studies exist in the literature that use ML and DL techniques for the analysis of endoscopic images. In 2023, [37] used two transfer learning models - namely ResNet-50 and Inception-V4, for the feature extraction of images obtained through the KID 2 dataset. They were able to successfully identify lesions in the GI tract with an accuracy of 0.9809. Simiarly in 2023,

the authors of [38] proposed their own CNN architectures and trained on the WCE Curated Colon Disease Dataset [39]. Their best results were from a custom CNN based on the MobileNet-V1 architecture, with an F1 Score of 0.9933. Other works such as [40] and [41] propose novel methods of classifying WCE images, but are purpose-built for binary classification. In particular, [40] uses the Seeker Optimization algorithm and the Elman Neural Network (ENN) for classification, whilst [41] uses the Water Strider Optimization algorithm and long short-term memory (LSTM) for classification. Both these works used data obtained from multiple sources. Table 3 summarizes the most recent and relevant studies conducted in the field of WCE image classification and analysis, particularly focusing on the Kvasir datasets and other datasets of similar nature.

It is important to note that the works presented in this section is either conducted on different datasets, different variations of the Kvasir datasets, or is on imbalanced data. Therefore, it must be acknowledged that there is no specific baseline to compare our results to. Furthermore, the goal of this paper is to provide highly accurate results for use within the medical domain, supported by the deployment of XAI techniques for validation.

III. MATERIALS AND METHODS

In this section, the dataset used for the study is described in Section III-A, while the DL models that we train on the dataset is described in detail in Section III-B. Similarly, the Explainable AI techniques employed in this study are covered in Section III-C. The main techniques used are LIME, Shapley and Class Activation Mapping (CAM) based methods, which can further analyze the models' decision-making process and identify the discriminative features that it has learned for classification.

A. DATASET

The Kvasir-Capsule dataset is the largest publicly released WCE dataset to date. This dataset contains data from real world clinical examinations. The dataset was collected, analyzed and labelled by a trained clinician. More detail on the collection of this dataset can be found in [4]. The dataset is comprised of 47, 238 labeled images and 117 videos where anatomical landmarks, pathological, and normal findings are detected [4]. However, only the labeled images are considered for the purposes of this study. The dataset contains 14 different classes of anatomical and luminal findings, and samples of the top 9 classes are shown in Fig. 2. The dataset is heavily imbalanced as indicated in table 4. It contains around 34, 000 images of the ''normal mucosa'' class while other classes samples range between 500-4200 images. Hence, under-sampling techniques is applied to enhance the models' performance. We use the threshold of 500 instances to select the top 9 classes. All classes with over 500 images are then under-sampled. These classes, along with the number of instances in the original dataset, are shown in table 4. The under-sampling technique was chosen in this study as opposed to other balancing methods since it has proven to work well for highly imbalanced image classification [42].

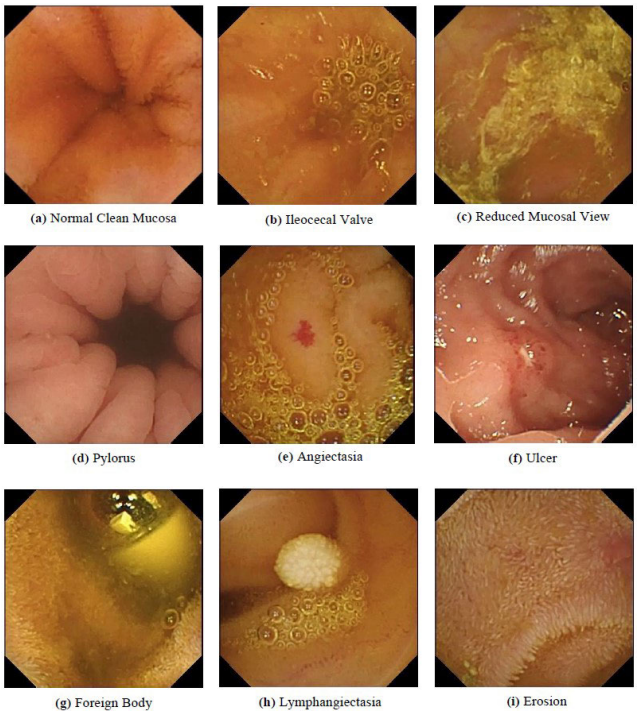


FIGURE 2. A sample image from each of the 9 classes selected for this work.

In [43], the authors compared several CNN models on the Kvasir-Capsule dataset and managed to achieve a weighted average F1-score of 0.673. This was achieved using a custom CNN developed by the authors, FocalConvNet. An important fact to note is that this study used the top 12 classes of the 117 videos in Kvasir-Capsule, meaning that the authors were classifying video frames instead of images. This indicates the lack of comparison between this study and what is presented in this paper as a comparative baseline, since the nature of the datasets used are different.

We apply pre-processing techniques to the dataset based on each transfer learning model used and their requirements. This includes image resizing, normalization, etc. Moreover, we use image augmentations with vertical and horizontal flips of the image because the capsule could be flipped as it travels along the digestive track, to reduce over-fitting.

B. MODELS

We explore the use of several deep learning and machine learning models for image recognition and classification. For this study, we will be using CNN-based models as well as transformer-based models. We first perform 10-fold cross validation testing to accurately compare the performance of each of the models. We then use statistical tests, such as the One Way Kruskal-Wallis test, to tell if the different distributions of F1-Scores for each model are statistically different from each other. Furthermore, we use Wilcoxon Signed-Rank

TABLE 3. Summary of studies presented in section II.

Paper	Dataset	Purpose	Approach	Results
Sharif et al. [24]	Privately Collected WCE images	Multi-class classification	Proposed CNN based on features extracted from fusion of VGG-16 and VGG-19	Accuracy: 0.9942
Borgli et al. [22]	HyperKvasir	Dataset proposal	ResNet-152 and DenseNet-161 Averaged	F1 Score: 0.91
Latha et al. [23]	Kvasir	Binary classification	Proposed Binary Classifying CNN	Accuracy: 0.96
Padmavathi et al. [25]	Kvasir-V2	Multi-class classification	Proposed DL model based on LeNet-5 Architecture	F1 Score: 0.9849
Caroppo et al. [37]	KID 2	Multi-class classification	ResNet-50 and Inception-V4 for feature extraction, mRMR for feature selection	Accuracy: 0.9809
Dey et al. [38]	WCE Curated Colon Disease	Multi-class classification	Proposed CNN models	F1 Score: 0.9933
Amirthalingham et al. [40]	Multiple sources	Binary classification	Proposed model w/ Seeker Optimization algorithm and Elman NNN (ENN)	F1 Score: 0.9778
Amirthalingham et al. [41]	Multiple sources	Binary classification	Proposed model w/ Water Strider algorithm and LSTM	F1 Score: 0.9829

TABLE 4. Distribution of images in the top 9 classes of the Kvasir-Capsule dataset, including all classes with more than 500 instances.

Class	Instances In Original Dataset
Normal Clean Mucosa	34,338
Ileocecal Valve	4,189
Reduced Mucosal View	2,906
Pylorus	1,538
Angiectasia	866
Ulcer	854
Foreign Body	776
Lymphangiectasia	592
Erosion	507

Test to determine exactly which models performed better or worse.

After determining the best deep learning model, we further explore their ability to extract useful features. This is done by using the extracted features and training a classical machine learning based classifier. Once again, we perform 10-fold cross validation and compare the machine learning models' ability to distinguish between the classes based on the features extracted by the best model.

1) CNN-BASED MODELS

CNNs are a type of artificial neural network that is specifically designed for processing image data [17]. CNNs are the most used method for many computer vision tasks, including image and video recognition and natural language processing (NLP) [17]. They are the basis on which the following architectures and models are built upon. In this paper, we explore the use of 6 different pre-trained CNN models as outlined below.

a: INCEPTIONV3

The InceptionV3 architecture is a CNN designed for image classification and object recognition [44]. The architecture improves upon the InceptionV1 and InceptionV2 models and uses several new ideas to increase accuracy and efficiency. The network is made up of repeated blocks of "inception modules", which are composed of parallel convolutions with varying kernel sizes and pooling operations. They are put together to capture the local and global features of the image. Moreover, batch normalization and residual connections are also employed to improve gradient flow and decrease the possibility of overfitting. InceptionV3 uses factorized 7×7 convolutions to reduce the number of parameters and computational cost. A new module called "grid reduction" is introduced, which uses max pooling to reduce the spatial size of the feature maps. Overall, the InceptionV3 architecture achieves exceptional performance on the ImageNet dataset. Due to its low computational cost and high efficiency, InceptionV3 has particularly grown in popularity in the field.

b: EFFICIENTNETV2

EfficientNetV2 is an image classification model which builds upon the success of the previous EfficientNet models. It introduces several new components to make the model more compact and faster to train [45]. The general EfficientNet structure is made up of a stem, a series of repeating blocks, and a head. However, EfficientNetV2 incorporates new concepts, such as the "Squeeze-Excite" block and "Conv-BN-Act" block. It also contains a new type of regularization called the "Stochastic Depth" along with a new scaling method called "Compound Scaling." The key blocks are explained below:

- The "Squeeze-Excite" block that makes the model focus on the most important features. It has two main components: a squeeze operation that decreases the dimensionality of the input tensor and an excitation operation that selectively enhances the important features.
- The "Conv-BN-Act" block, consisting of a convolutional layer followed by batch normalization.

- "Stochastic Depth" is a type of regularization that randomly drops some of the blocks during training, which helps prevent overfitting.
- "Compound Scaling" is a scaling method that allows the model to be scaled efficiently across multiple dimensions such as depth, width, and resolution. It helps in creating smaller and more efficient models that can perform equally as well as more complex models.

Overall, the EfficientNetV2 architecture improves upon previous models by increasing in efficiency and speed whilst decreasing in size. Despite the reduction in size, the architecture is still able to achieve noteworthy performances on image classification tasks.

c: VGG-16 AND VGG-19

Both VGG-16 and VGG-19 were introduced in the paper [46], that proposed these two models which have a simple architecture yet different from most popular transfer learning models.

Trained on the ImageNet dataset, VGG16 consists of 16 layers which inspires the name. It contains 13 convolutional and 3 fully connected layers used for classification. The novelty of the algorithm lies in its use of a small receptive field during processing, which means using filters of smaller sizes with a smaller stride. This is for both convolutional and max pooling layers.

Similarly, the VGG19 architecture consists of 19 layers, this time with 16 convolutional layers instead of 13. Each of the convolutional layers are grouped into 5 groups, with max pooling layers in between. VGG19 adds an extra convolutional layer to each of the last three groups as comapred to VGG16. Both architectures proprosed in the paper are characterized by their small receptive fields which allow the model to identify and learn more complex features. Moreover, the VGG set of models are designed to be simple and avoid unnecessary computation using layers like batch normalization. However, it still prevents overfitting by using max pooling layers after every group of convolutions to reduce the size of the feature maps.

d: MOBILENETV3LARGE

The MobileNetV3 is a family of lightweight convolutional neural networks designed for mobile applications [47]. It uses an inverted residual structure made up of a stem layer and several bottleneck layers. Each bottleneck layer uses a singleton pointwise convolutional layer, followed by a depth-wise convolutional layer, and another singleton point-wise convolutional layer. The MobileNet architectures make use of hard swish and squeeze-and-excitation techniques to boost performance and improve efficiency. The model achieves SOTA performance on several image classification tasks while still keeping its model size small and computational costs low. In fact, [47] boasts about a 25% faster detection for MobileNetV3 while matching MobileNetV2's accuracy, showing the impact of the latest MobileNet version.

e: RESNET152V2

ResNet152V2 is an extension of the original ResNet architecture [48]. It has a very deep architecture consisting of 152 layers with residual blocks, and shortcut connections that improve gradient flow through the network. These shortcut layers work to resolve the vanishing gradient problem. It is designed to also minimize the effect of degradation or the reduction of accuracy when a network is deeper. The residual blocks in ResNet152V2 are made up of a group of convolutional and batch normalization layers. In the residual block, the shortcut connections skip one or more convolutional layers and directly connect the input to the output, preserving the input information. The use of shortcut or skip connections is the primary characterization of the ResNet architecture, which allows for the use of deeper neural networks without exploding or vanishing gradients, leading to improved performance.

2) TRANSFORMER-BASED MODELS

Transformer based models [49] are based on the "self-attention mechanism" which allows them to produce a representation of different positions of the input sequence. They adopt an encoder-decoder architecture, with blocks of replicated layers in each half. The encoder maps an input sequence into a sequence of hidden states based on the sequence embedding and the positional encoding of the data. Each layer of the encoder has two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. The self-attention mechanism allows the model to learn from a longer range of the input sequence, while the feed-forward network generates a sequence of output tokens by acting as a non-linear transformation to each position in the sequence. The decoder operates on the output of the encoder, and attempts to generate the sequence one token at a time. It uses a stack of redundant layers and a self-attention layer. The multi-head attention mechanism in the decoder helps the model to focus on the relevant parts of the input sequence for each output token. Both the encoder and decoder layers are connected with residual connections and layer normalization.

Vision Transformer (ViT): The ViT is one of the most popular transformer models for image classification. The ViT architecture [50] is based on the idea that an image can be represented as a sequence of flattened 2D patches, and that these patches can be treated as a sequence of tokens that can be processed by a transformer network. The ViT architecture consists of several layers of transformers. To adapt the ViT model to images, the input image is first divided into a grid of fixed-size patches, and each patch is flattened into a vector. These flattened patches are then taken as a sequence of tokens, which are fed into the transformer layers. The first token in the sequence is always an embedding representing the entire image. For image classification, the ViT architecture uses a classification head to map the output of the final transformer layer to a set of class probabilities.

In our implementation, we use the ViT pre-trained self-attention layers followed by the following hyperparameters: A single dense layer with 1024 neurons and ReLU activation, a dropout layer with a dropout rate of 0.3, and an output layer of 9 neurons with softmax activation for classification. The model was trained with Adam optimizer with a learning rate of 0.001.

3) CLASSICAL MACHINE LEARNING ALGORITHMS

Several classical machine learning algorithms were used to classify the images based on the features extracted from the best performing deep learning model, as explained in the pipeline. This is better illustrated in Fig. 3, where the input images of any class is sent to the ViT model, which provides a feature vector of size 1024. This feature vector is used as an input for different classical machine learning models for classification. Hence, this essentially replaces the dense layer classifier at the end of ViT with a classical ML algorithm. For comprehensive testing, we employ a variety of methods. In particular, we explore Decision Trees, K-Nearest Neighbors, Naive Bayes, Support Vector Machines, Logistic Regression, Random Forest, Gradient Boosting Classifiers, and AdaBoost.

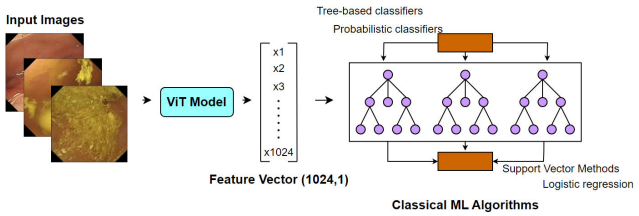


FIGURE 3. The general pipeline for applying the classical ML algorithms. First, the input images are taken and feature vectors are extracted from them using the ViT model. Then, these feature vectors are fed into classical ML algorithms.

The input of each of these classifiers will be the feature vector, obtained from the output of the final layer in the deep learning model. This allows us to use the feature extraction of the deep learning model, and thus classifying the image based on the separation of the feature space. We perform 10-fold cross validation testing to accurately compare the performance of each of the models and use these classical machine learning models' performance to comment on the usefulness of the extracted features by the best model. This includes examining the t-Distributed Stochastic Neighbor Embedding (t-SNE) projection of the feature space and the separation of the different classes to explain the performance of the classical machine learning models.

C. APPLYING EXPLAINABLE AI TECHNIQUES

In the medical domain, model prediction explainability and interpretability are necessary in assisting clinical research and decision making. The primary issue with endoscopy is that it requires a doctor or trained nurse to constantly watch the video feed for up to 30 minutes to observe any anomalies in the patients body. Having an intelligent machine learning-driven system that complements the doctors findings can be beneficial in uncovering otherwise hidden ailments. However, the lack of explainability of most deep learning models makes them unsuitable for implementation in the medical field. Hence, eXplainable AI (XAI) techniques have been developed to provide explainability to deep learning models to semantically understand the criteria they use for classification, which improves trust and reliability of the model.

Typically, there are two main sets of methods to develop such explainable systems – ante-hoc and post-hoc techniques. Ante-hoc methods incorporate explainability into a model from the very beginning. Post-hoc techniques generally make decisions after the model has been trained. In this study, we apply five of the latest post-hoc machine learning methods - SHapley Additive exPlanations (SHAP) [51], Local Interpretable Model-agnostic Explanations (LIME) [52], Gradient-weighted Class Activation Mapping (GradCAM) [28], GradCAM++ [31], and Layer-wise Class Activation Mapping (LayerCAM) [53]. Fig. 1 in the introduction provides the reader with the general XAI pipeline and experimental structure used in this paper. The model itself and its predictions are fed into both categories of XAI techniques (Feature-based and Propagation-based), resulting in the feature importances and the heatmaps.

As endoscopic imaging belongs to the risk-sensitive field of medicine, it is not adequate to provide global explanations of the model's decision making. Each patient requires diagnosis and localization on a local level. For that, the XAI methods that would provide case-specific explanations were used. In addition, model-agnostic technique with post-hoc implementation level were considered for generalization purposes. As a result, we applied three different XAI methods, two of which are feature based; SHAP [51] and LIME [52], and a third which is propagation-based, GradCAM [28]. Other close XAI techniques include rule-based methods such as Anchors [54], Bayesian Rule Lists (BRL)s [55], and Generalized Additive Models (GAM)s. Nonetheless, we argue that Anchors follow a very similar approach to that of LIME as both provide explanations through a local region in the feature space, however, anchors follow a slightly different procedure by generating if-then rules while LIME relies on a linear model around an instance [56]. On the other hand, GAM and BRL do not provide local-level explanations. It is worth mentioning that GradCAM has different variances and thereby we decided to analyze our images using GradCAM, GradCAM++, and LayerCAM for illustration purposes.

The following subsections detail each of the XAI techniques used for the purposes of this study.

1) LIME

Local Interpretable Model-agnostic Explanations or LIME was proposed by in Ribeiro et al. and colleagues [52], to explain the predictions of any classifier or regressor by approximating it locally with some interpretable model.

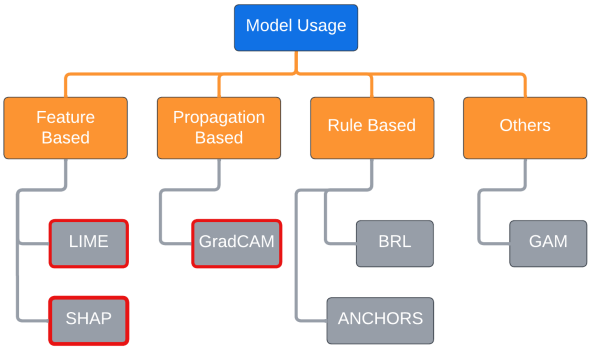


FIGURE 4. Explainable AI (XAI) Approaches considered per model, divided into feature-based, rule-based, propagation-based, and other approaches.

Consider the blackbox AI model to be explained as a function f that maps $\mathbb{R}^d \rightarrow \mathbb{R}$. So, the probability that an input x belongs to a particular class is denoted by $f(x)$. Now, from a class G of potentially secondary models, consider $g \in G$ to be a possible explanation of the prediction made by f. We define a function $\pi_x(z)$ to represent the proximity of x to an instance z. The unfaithfulness of the prediction by our model g in the region within $\pi_x(z)$ is then denoted by $\mathcal{L}(f, g, \pi_x)$. By definition, LIME then evaluates an explanation based on the following equation:

$$\text{explanation}(x) = \text{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (1)$$

An important point to note is that $\Omega(g)$ here is simply a measure of the complexity and not the interpretability of the explanation. Hence, an optimal explanation would require for us to minimize $\mathcal{L}$ whilst keeping $\Omega(g)$ as low as possible for faster computation.

LIME is algorithmically implemented within five main steps.

1) A data point 'x' is selected from the original dataset
2) A synthetic dataset is then generated through sampling the original dataset and artificially producing instances, both close and far to the desired data point x
3) Model is run on the new perturbed dataset to make predictions
4) We assign weights to each feature based on proximity to data point x
5) A surrogate model is used to approximate the model in the region of the selected point x.

The choice of the model is implementation specific and depends on the nature of the dataset.

2) SHAP

Shapley Additive exPlanations (SHAP) is another XAI technique that provides interpretability of the model. According to [51], SHAP is based primarily on the concept of Shapley values, an important concept in Game Theory. Shapley values calculate the average marginal contribution of some feature f towards a model's score. The advantage of using Shapley values is that it is a black-box model agnostic technique, which means that the technique works independently from the model it is assigned to. On that basis, with just the inputs and outputs, we can examine the feature contributions and draw conclusions on the data. What makes SHAP a popular algorithm choice, is its capability to explain the output of a model by attributing importance to each feature, whilst also taking into account the interactions between the features themselves.

Using coalitions, we are able to compute the Shapley values of each feature for a data instance x, as predicted by a model f. The actual Shapley value itself is the average marginal contribution of a feature.

The interpretability of SHAP values are computed through the importance values of each feature for local predictions. In general, the SHAP method calculates the results for $2n$ combinations of pixels within an image and the model's prediction - where n is the number of pixels. Nonetheless, given that images come with a huge number of pixels n, SHAP relies on Shapley sampling values method to approximate estimations for each feature attribute [51].

3) GRADCAM AND GRADCAM++

Another post-hoc explainability method used is the Gradient-weighted Class Activation Mapping (GradCAM) technique proposed by [28]. GradCAM is focused primarily on computer vision, whereas SHAP and LIME can also be applied to tabular and textual data. Built as an improvement to the original CAM technique proposed by Zhou et al. and authors [57], GradCAM is suitable for more complex CNN architectures with connected layers. The layer used for making predictions is the final convolutional layer of each of the models to be tested as they have the most discriminative features. As seen in Fig. 5, a heat-map is generated, highlighting areas that the model considers relevant to the final prediction.

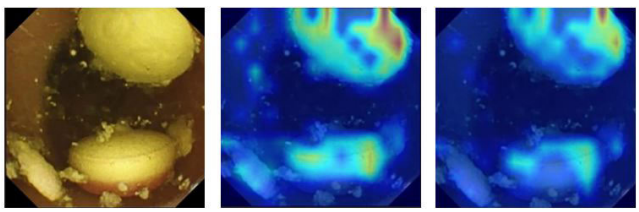


FIGURE 5. Original image from Foreign Body class, along with the generated heat-map from GradCAM and GradCAM++

A crucial limitation of GradCAM is its inability to properly identify objects in an image if the image has multiple occurrences of the same class. In the Kvasir-capsule dataset, multiple occurrences are not uncommon and hence any improved explainability is beneficial. GradCAM++ was developed to account for such limitations [31]. Though they are both based on the CAM architecture, the techniques used by GradCAM and GradCAM++ are different. GradCAM generates visual explanations by computing the gradient of the output class score with respect to the feature maps of the

last layer of the model. GradCAM++ takes this a step further by also computing the second-order gradients to capture more fine-grained information about the feature maps. The fundamental difference in all 3 models lies in the way the weights are calculated and the inclusion of backpropagation in GradCAM and GradCAM++.

According to the authors, Y^c represents the score of a particular class c. w_k^c denotes the weights for a particular feature map A^k and class c. Then, using the equation below, we can calculate the area of interest L^c in each location in the location (i, j) as follows:

$$L_{ij}^c = \sum_k w_k^c * A_{ij}^k \tag{2}$$

As each location (i, j) directly correlates to the map L_{ij}^c, we can thus obtain a heat map indicating how the black box AI model made the prediction. The fundamental difference in the models are the way the weights, are calculated and the inclusion of backpropagation in GradCAM and GradCAM++.

4) LAYERCAM

In [53], the authors proposed a more advanced version of GradCAM called Layer-wise CAM (LayerCAM). Unlike GradCAM which only uses the final convolutional layer, LayerCAM uses multiple convolutional layers to generate the class activation maps. Furthermore, LayerCAM also attempts to spatially weight the activations by positive gradients solely. It computes the weighted sum of GradCAM maps from several intermediate layers of the CNN. Since the last convolutional layer tends to have a lower spatial resolution than the previous layers, the inclusion of multiple layers improves the accuracy of the algorithm by several orders of magnitude.

Let the image classifier be denoted by a function f for any parameter θ. The following equation then predicts a target score y^c when some image I is inputted into the function f

$$y^c = f^c(I, \theta) \tag{3}$$

Assuming the black box model being explained is a CNN, let the feature maps of the convolutional layers be denoted by A. For a network with n layers, the ith feature map can be obtained by A^k. Similar to GradCAM, LayerCAM also includes the derivatives of the prediction score obtained [53]. So for any feature map A^k, we can obtain the gradient at any point of interest (i, j) by the following equation:

$$g_{ij}^{kc} = \frac{\partial y^c}{\partial A_{ij}^k} \tag{4}$$

The weight now assigned to the chosen point of interest can be represented as:

$$w_{ij}^{kc} = \text{ReLU}(g_{ij}^{kc}) \tag{5}$$

In order to produce a heat-map for any layer, we obtain the class activation map by multiplying the weight in the equation above with each point of the feature map [53].

$$\hat{A}_{ij}^k = w_{ij}^{kc} A_{ij}^k \tag{6}$$

In order to produce the final heat-map of the image passed, we find the class activation map by simply finding the linear combination of the values obtained in A^k. The equation below represents the final expression:

$$M^c = \text{ReLU}(\sum_k \hat{A}^k) \tag{7}$$

Thus, the XAI techniques explained above are used in this work in order to provide explainability and interpretability to the deep learning models considered in this work and enhance their reliability and acceptance among physicians and healthcare professionals.

IV. RESULTS AND DISCUSSION

A. PHASE 1

In this section, the results of applying the following transfer learning models are presented: InceptionV3, EfficientNet, VGG16, Vision Transformer (ViT), VGG19, MobileNetV3Large, and ResNet152v2. K-Fold Cross Validation Was conducted with $K = 10$ for each model. The F1-scores, accuracy, precision and recall scores, along with the confusion matrix for each of the models were obtained. Since the F1-scores are the primary metrics to explain the performance of the model, the distribution of F1-scores across the 10-Fold CV was explored. Table 5 shows the average and standard deviation of the F1-scores across the 10-Folds of cross validation for all the models applied in this work. A more visual representation of the distribution of F1-scores can be seen in a box-plot of the distribution of 10 F1-scores per model in Fig.6.

TABLE 5. Average and standard deviation of F1-scores across 10 folds of Cross Validation.

Model Name	F1 Score
InceptionV3	0.26 ± 0.04
EfficientNet	0.95 ± 0.01
VGG16	0.94 ± 0.02
ViT	0.97 ± 0.01
VGG19	0.38 ± 0.06
MobileNetV3Large	0.95 ± 0.02
ResNet152v2	0.94 ± 0.01

We can see that across 10-fold CV, EfficientNet, VGG16, ViT, MobileNetV3Large, and ResNet152V2 all boast F1-scores of higher than 0.90, with ViT performing the best with 0.97 ± 0.01. On the other hand, InceptionV3 and VGG19 perform the worst with F1-scores significantly lower than the other models.

Since InceptionV3 and VGG19 performed poorly on this dataset, they were dropped from the plot leaving the top 5 models to be examined as in Fig.7. As can be seen, the top 5 models performed exceedingly well, achieving an F1-score of over 90% consistently, with some achieving F1-scores of over 95% too. The distribution of the top 5 models were examined further.

We observe that ViT and MobileNetV3Large have the highest median performances which are well over 96% F1-score. Moreover, as shown in Table 5, ViT has one of the

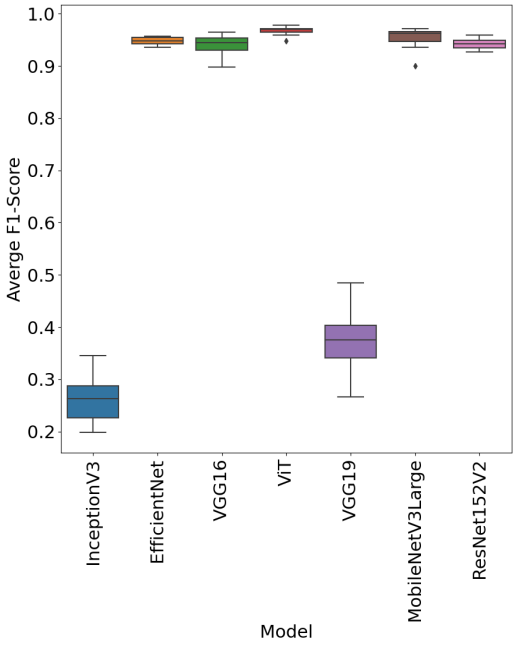


FIGURE 6. Box-plot showing statistical distributions of the F1-scores for all deep learning models tested.

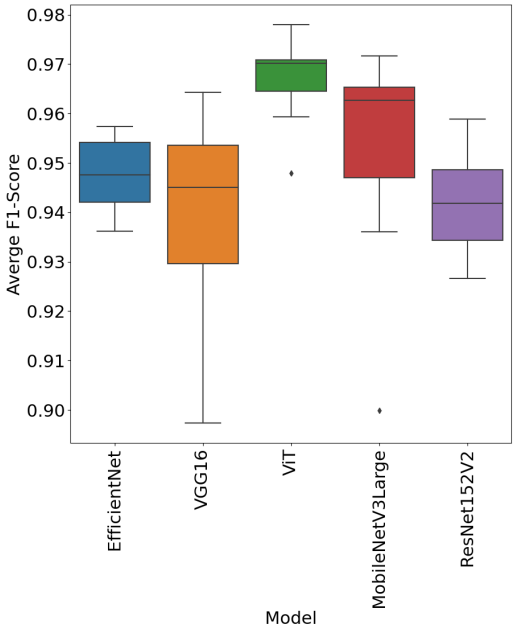


FIGURE 7. Box-plot showing statistical distribution of the top 5 performing models' F1-scores.

lowest standard deviations of F1-score across the 10 folds, which indicates a stable model. In fact, other models such as EfficientNet and ResNet152v2 have also low standard deviations of F1-scores which indicate stable models as well.

To further confirm whether all the models tested have statistically different distributions, we first perform the Kruskal-Wallis test to test whether all distributions have similar medians ($M_i = M_j\, i, j = 1, \ldots, 7$). The Kruskal-Wallis' p-value is 0.0000 which allows us to reject the null hypothesis that all the distributions are equal. Hence, we employ all pairwise tests using the Wilcoxon Signed-Rank test with the hypotheses:

$$H_0 : M_i = M_j,\ i \neq j \in \{1, \ldots, 7\}$$
$$H_a : M_i \neq M_j,\ i \neq j \in \{1, \ldots, 7\}$$

where M_i is the median for the F1-score distribution of model i. This test allows us to examine exactly which distributions differ from each other by performing a pair-wise comparison between every pair of models to confirm whether they have an equal median or not. The column "Statistically Different?" allows us to determine which models are statistically different from each other. If H_0 is rejected, then that means that the distributions are statistically different. Otherwise, (False) means that we fail to reject H_0, suggesting that the models' distributions are not statistically different from each other. Fig. 6 shows clearly that VGG-19 and InceptionV3 perform poorly when compared to the other models.

In table 6, the Wilcoxon Signed-Rank Test is performed with the top 5 models. It can be noticed that ViT always leads to the null hypothesis being rejected. Examining the boxplot shows us that it is because ViT has a superior performance compared to the other models. Hence, the Vision Transformer is the best model for solving the problem.

TABLE 6. Pair-wise comparison of distributions for top 5 models using Wilcoxon Signed-Rank Test.

Model 1	Model 2	Statistically Different?
EfficientNet	VGG16	False
EfficientNet	ViT	True
EfficientNet	MobileNetV3Large	False
EfficientNet	ResNet152V2	False
VGG16	ViT	True
VGG16	MobileNetV3Large	False
VGG16	ResNet152V2	False
ViT	MobileNetV3Large	False
ViT	ResNet152V2	True
MobileNetV3Large	ResNet152V2	False

We examine the average confusion matrix and classification report for the ViT across the 10 folds of testing to gain a better insight into its successes and points of failure. Table 7 shows the precision, recall, F1-Score of the ViT model along with the number of instances in each class during testing in each iteration of K-Fold (support). The support of 450 shown for accuracy, macro average, and weighted average shows that they are taken over all 450 images in the testing data.

The corresponding confusion matrix for the model can be seen in Fig. 8. The value in each cell indicates the percentage of instances classified correctly from that class.

As seen in the classification report presented in Table 7, the ViT model achieved a macro average precision, recall, and F1 score of around 97% across all the 10 folds, which indicates the success of the ViT in classifying the images. Further examination of the confusion matrix shows that it is able to classify all classes correctly, with only very rare confusions

TABLE 7. Complete model metrics for ViT, including the metrics per-class and the overall macro and weighted scores (For Precision, Recall, and F1-score).

	Precision	Recall	F1-Score	Support
Angiectasia	0.953129	0.966	0.958454	50
Erosion	0.943111	0.94	0.940911	50
Foreign Body	0.971149	0.97	0.970076	50
Ileocecal Valve	0.991905	0.974	0.982552	50
Lymphangiectasia	0.973554	0.96	0.965578	50
Normal Clean Muc.	0.974958	0.99	0.982172	50
Pylorus	0.96332	0.948	0.954964	50
Reduced Muc. View	0.990379	0.986	0.987995	50
Ulcer	0.963602	0.98	0.971331	50
Acc			0.968222	450
Macro-Avg	0.969456	0.968222	0.968226	450
Weighted-Avg	0.969456	0.968222	0.968226	450

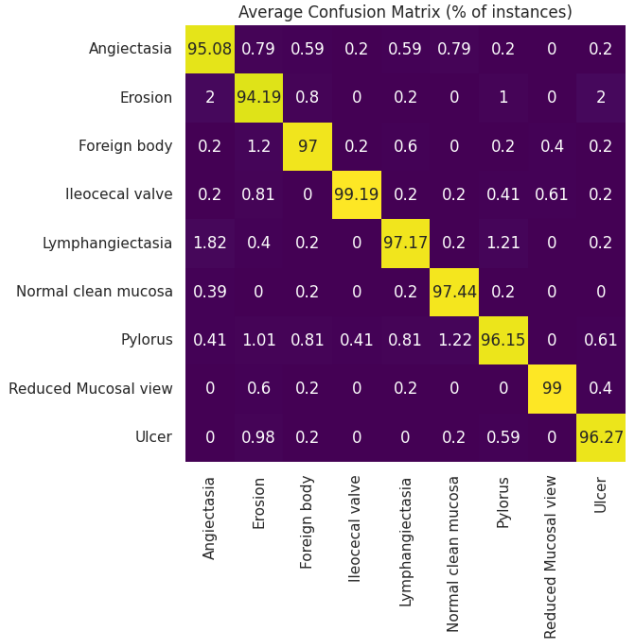


FIGURE 8. Confusion matrix of percentage of predictions of each class across 10-fold cross validation for ViT Model.

with other classes since all the percentages are very close to a 100%. It seems Erosion and Angiectasia classes are the most confused with other classes, since the model only got 94.19% and 95.08% of them correctly, but the misclassifications are still a rare occurrence.

Based on the achieved results, it can concluded that ViT has been conclusively shown to be the best model. Looking at ViT architecture, it can be noticed that the main difference between ViT and the other CNN methods used is the features extracted by the transformer layers with the multi-head self-attention. Thus, the idea of using the features extracted by ViT as input to some classical machine learning models to classify the diseases was explored. Hence, a series of machine learning models trained with the features extracted from ViT were tested. The goal of such testing is to better understand the usefulness of the extracted features from ViT. The models

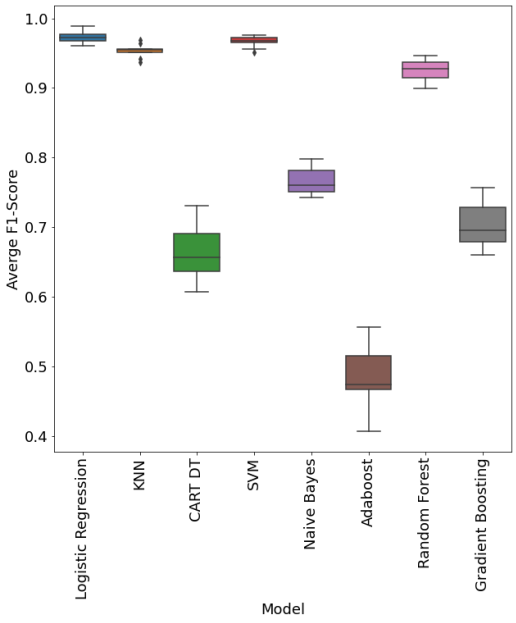


FIGURE 9. Box-plot showing the distribution of F1-scores for each of the ML models used.

TABLE 8. Results of GridSearch for best 3 performing models, showing best parameters for each.

Model			
LR	c: 100	Penalty: l1	Solver: liblinear
SVM	c: 10	Gamma: 0.01	Kernal: rbf
KNN	n_neighbors: 1	Metric: euclidean	Weights: uniform

tested are: Decision Tree, Random Forest, KNN, XGB, Naive Bayes, SVM, Logistic Regression, and AdaBoost. 10-fold cross validation was performed. The box plot of the distribution of the F1-scores of the machine learning models is presented in Fig. 9.

As in Fig. 9, we can easily see Logistic Regression, KNN, SVM, and Random Forest perform significantly better than the rest. We can examine these 4 best models more closely as in Fig. 10 to see that they all achieved an F1 score over 90% and in this case are comparable to the results obtained through deep learning. In fact, Logistic Regression and SVM perform as well as the ViT model with dense layers for classification.

We will take the top 3 best performing models and perform GridSearch to hypertune the parameters. In this case, we are referring to Logistic Regression, SVM and KNN. When running GridSearch, we obtain the following tuned hyperparameter results, shown in Table 8.

Observing the high performance of classical machine learning models using the features extracted by the vision transformer, indicates that the features extracted are of high quality and separate the data well. This can be examined by viewing the distribution of the extracted features, using a high dimensional visualization technique called t-SNE, as done in Fig. 11. Viewing the data in two dimensions provides us with a better understanding of how each class is separated in the

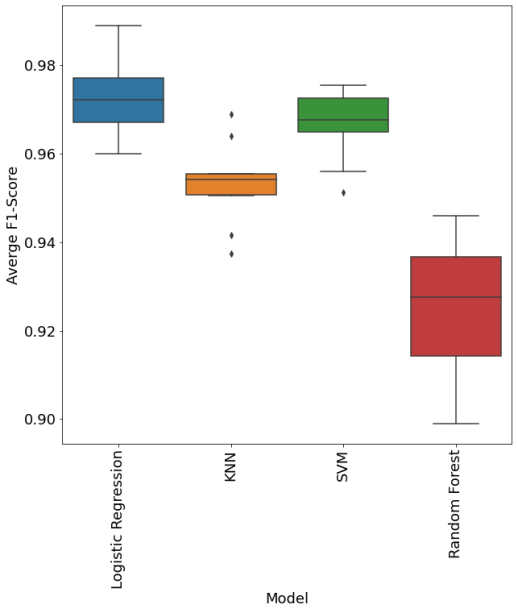


FIGURE 10. Box-plot of F1-score distributions for only the ML models that achieved an F1-score above 90%.

TABLE 9. Clustering metrics.

Metric	Score
Silhouette Score	0.02425
Calinski-Harabasz Score	77.51718
Davies-Bouldin Score	4.55432

high dimensional space. This can be clearly seen in Fig. 11 since the different colors (different classes) are well separated and in distinct clusters.

Clearly, the features form distinct clusters based on the class, which is why even simple models such as logistic regression performs exceedingly well. However, there is still slight overlap between a few classes. This may explain why random forest performs slightly worse than logistic regression and SVM, but still achieves over 90% F1-scores. The high performance can be attributed to the feature space being well separated, allowing it to easily segment the space into the different classes. Moreover, this is also why simple methods such as KNN perform exceedingly well since the data is already highly clustered based on the classes.

Examining the clusters quantitatively, we use three metrics - Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score to help determine how well different clusters are separated, mostly based on the inter-cluster and intra-cluster distances. As in Table 9, we see the data is well separated in the latent space.

As can be seen in Table 9, the silhouette score is close to 0 which indicates that the decision boundaries are relatively close to each other. This is likely because one class spans multiple clusters which are far apart. Hence, this metric may seem misleading when used individually. The Davies-Bouldin Score is fairly low, which indicates good clustering between similar clusters. The Calinski-Harabasz Score is high, which indicates well separated and dense clusters. Hence, the ViT features performs useful feature extraction which is why most of the classical machine learning models perform so well.

Based on the experiments performed in Section IV-A, it can be concluded that ViT performed better that the rest of the deep learning models explored. Moreover, the use of the features of ViT with machine learning models have also been examined which have also proven to be successful. However, despite clustering analysis of the extracted features of ViT, the models are still not explainable due to the nature of the feature extraction from ViT, which is the issue to be addressed in Section IV-B.

B. PHASE 2

This section covers the results from the various XAI techniques. Fig. 4 summarizes the different approaches based on the model usage and the XAI methods implemented in this paper.

1) FEATURE-BASED EXPLANATIONS

SHAP values and LIME methods, as feature-based XAI techniques, are by far the most comprehensive and dominant across XAI methods for visualizing feature interactions and feature importance [58]. In this section the results of applying both SHAP and LIME methods on a variety of endoscopic images are presented and discussed.

a: SHAP

In this section, the results of applying SHAP on the dataset of endoscopic images are presented. Fig. 12 shows the SHAP results on the erosion and ulcer classes respectively along with the highest three predictions. The results showed that SHAP values provided clear explanations for many classes within endoscopic images. The red pixels represent features that contribute towards the corresponding prediction, whilst the blue pixels represent the features that do not contribute to the prediction.

SHAP provides adequate explanations in many classes with notable areas of distinction. However, it fails to provide reasonable classifications for classes with more complex inputs (See Fig. 13). SHAP's approximations result in a reduction in the input images' quality. Hence, it reduces SHAP's ability to effectively interpret models' results. It also clears out an area of improvement as both the classifications in Fig. 13 are, in fact, incorrect. Both images belong to the foreign body class, whilst it is predicting them as "Ileocecal Valve." SHAP clearly shows that the model considered foreign body as the third and second highest probabilities respectively, yet it was unable to detect regions of interest to support this prediction.

b: LIME

In this section, the results of applying LIME on the dataset of endoscopic images are presented. Fig. 14 shows LIME results

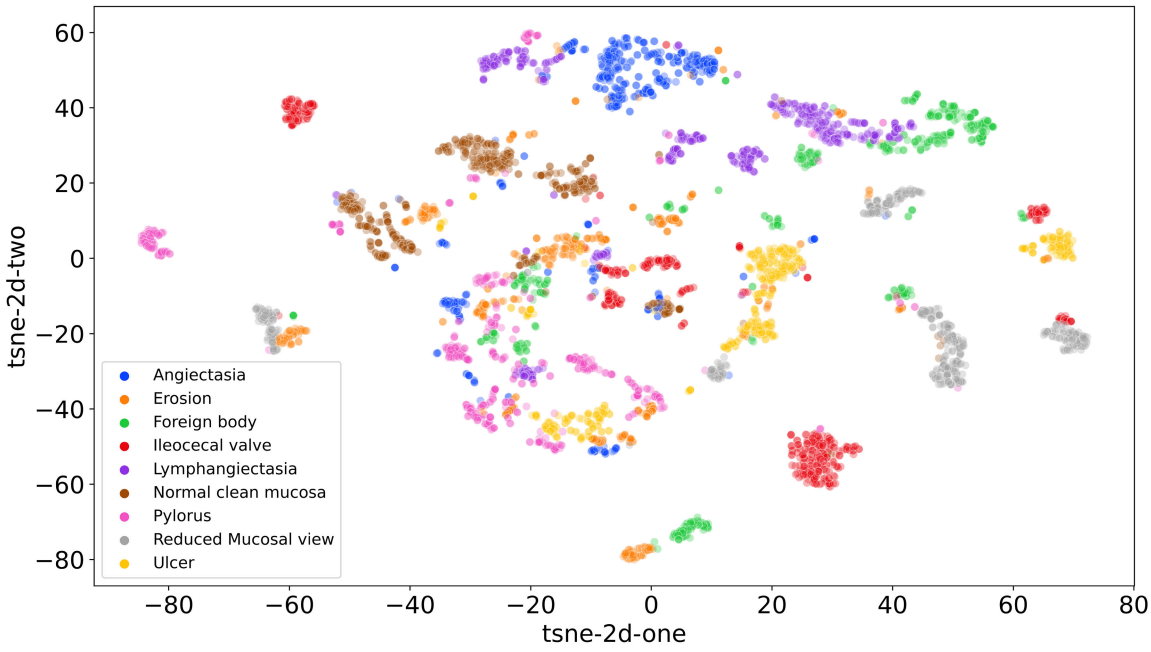


FIGURE 11. tSNE map showing the clustering of classes when projected onto a 2-dimensional plane. Classes are indicated according to their color.

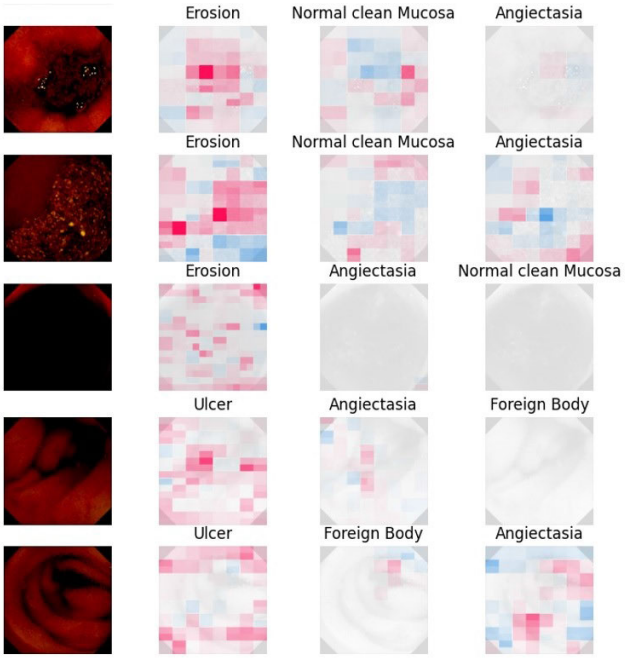


FIGURE 12. SHAP results on three highest predictors for Erosion and Ulcer classes.

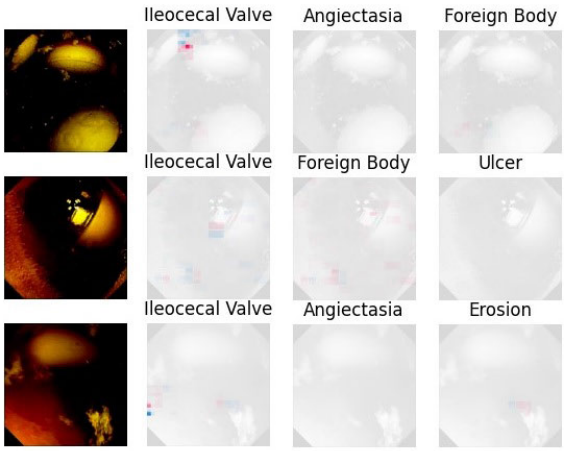


FIGURE 13. SHAP results showing inadequate predictions. Input images are both under "Foreign Body" class, but SHAP cannot identify regions of importance.

on Lymphangiectasia, Pylorus, and Foreign Body classes respectively.

As evident, LIME provided more sensible explanations to the model predictions in many cases, especially for the Foreign Body class, which SHAP could not explain. Nonetheless, LIME could easily miss out on important features, since the surrogate model may produce inaccurate generalizations [28]. A good indication of this inaccurate generalization can be seen in some of the sample images (case 1 and case 2) in which the black edges of the endoscopic images are considered important for classification.

This can be clearly seen in many cases in which LIME considers the black edges of endoscopic images as important (See case 1 and case 2 in Fig. 14), which is a limitation in LIME technique.

2) PROPAGATION-BASED EXPLANATIONS

Propagation-based explainers such as GradCAM are improved versions of class activation maps (CAMs) that aim

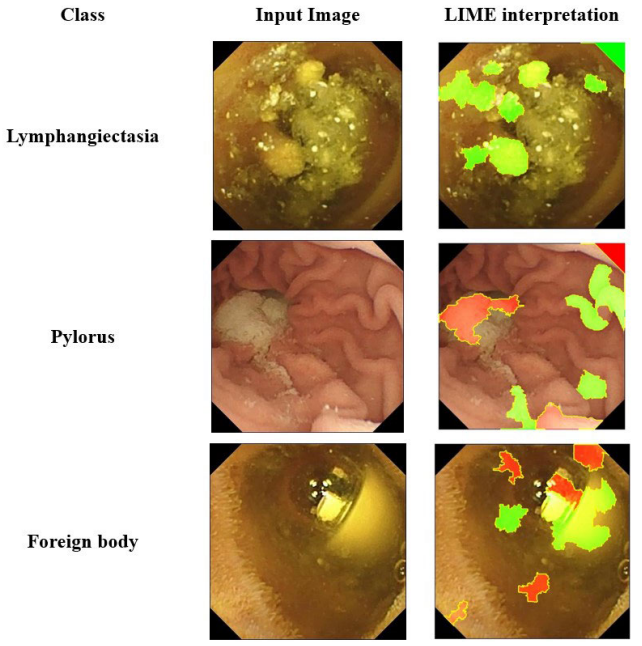


FIGURE 14. LIME interpretations for Lymphangiectasia, Plyorus, and Foreign Body. For the first and second images, LIME interprets regions in the top right of the images (green and red respectively) as important, although they are camera faults.

to reduce the time complexity of the latter. As this study utilized a ViT model with no convolutional layers, we decided to fit the last token computed in the very last attention layer of the model. This is a safe assumption as the gradient of the output will remain zero along the channels in the last layer [59]. GradCAM weighs the 2-d activations by the average gradient to generate an image for the selected features. Two variants of GradCAM were also tested for comparison purposes which includes GradCAM++ and LayerCAM. Fig. 15 shows GradCAM results on three different classes: Ulcer, Reduced Mucosal View, and Foreign Body respectively.

From the results obtained, we noticed that GradCAM with the first order of gradients produces the best results for the dataset. Further, GradCAM showed the best overall results across all three XAI methods implemented. This can be clearly observed with troublesome classes such as Ileocecal Valve. Fig. 16 shows a comparison between the three different XAI techniques results on the class of Ileocecal valve. As discussed earlier, these important areas are highlighted based on the features obtained after training the ViT model in the case of LIME and SHAP. On the other hand, GramCAM method depends on the back propagation weights obtained from the last prediction layer of the model to generate a heat map highlighting the important areas. A more detailed description of the procedure followed for each technique can be found in III-C.

In general, propagation-based models tend to extract better visual explanations from neural networks as the weights of the last output layer of the feature map tends to carry more information about the important regions within an image.

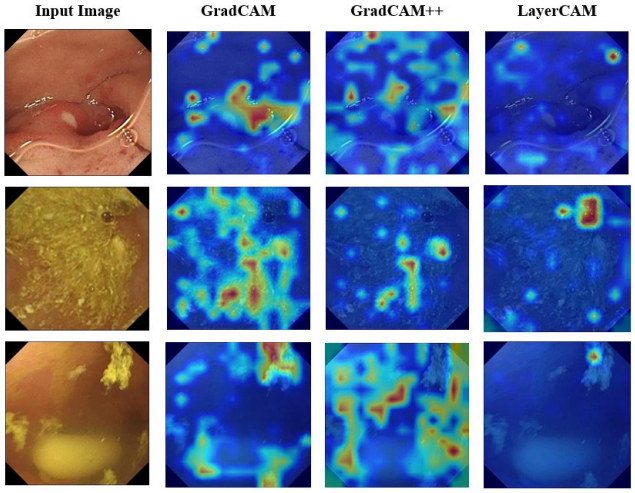


FIGURE 15. Given an input image, the GradCAM, GradCAM++ and LayerCAM interpretations and results showing regions of importance.

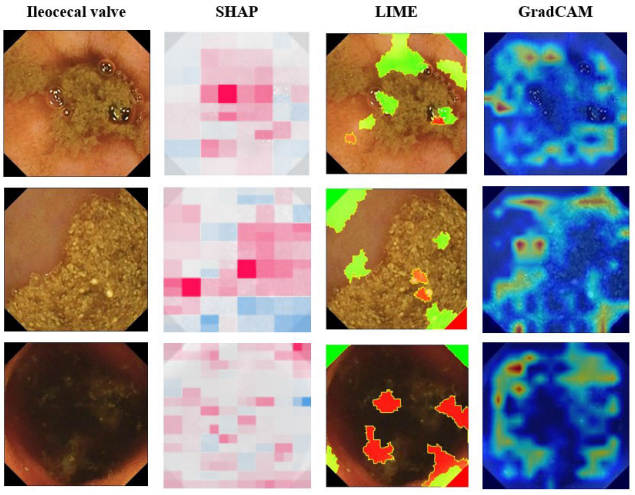


FIGURE 16. For the Ileocecal Valve class, the interpretations of three variants of XAI techniques, namely SHAP, LIME and GradCAM and their regions of importance.

This indicates that we expect GradCAM to remain superior to other explanatory methods regardless of the domain of the problem at hand. Nevertheless, when the model use does not utilize a back-propagation mechanism like decision trees or k-nearest neighbours, a mathematical approach such as SHAP and LIME techniques is the only way to go. Additionally, One limitation of GradCAM is that it fails in the localization of objects with multiple occurrences of the same class [60]. Consequently, the quality of GradCAM results can be highly reduced in details dense applications such as bacterial microscopic images.

V. CONCLUSION

In this paper, different deep learning models were applied on a dataset of endoscopy images, from which the top 9 classes in terms of the number of instances we extracted. In summary, the images were first processed through state-of-the-art DL

models, with the results showing that the Vision Transformer (ViT) achieves the most promising outcome with an average accuracy as high as 96.8% and an F1-score of 97%. The performance of different surrogate classifiers were further investigated using the last feature maps of the vision transformer as an input. The results showed that logistic regression, KNN, SVM, and Random Forest all managed to achieve an average F1-score above 90% where both logistic regression and SVM achieved results as good as that of ViT dense layers. Finally, a thorough comparison between three variants of XAI techniques that can be used in the field of endoscopic imaging were provided. It can be concluded that GradCAM yielded the best outcome across all three methods due to its mechanism of relying on the back-propagation gradients.

The work presented in this paper show the promise of using ML and DL models for the purposes of medical diagnosis. In fact, this is particularly evident through our highest accuracy of 96.8%, showing a clear trend of increased reliability within such models. Especially within the field of medicine and more specifically for WCE images, this paper helps to promote the reliability of DL models, as verified and demonstrated further through the implementation of XAI. However, it is important to note the limitations of using machine intelligence for complete diagnosis, since even a small margin of error as little as 3% can cause a significant misdiagnosis. We aim to continue improving the performance of DL models in the future. Other future work includes verifying the interpreted XAI results through medical practitioners in the field. This could include a verification process that authenticates whether a highlighted region of importance by the model actually identifies a GI tract disease, which can further improve upon our study. Collaborative efforts between ML experts and medical professionals will not only improve the reliability of such studies, but can also work towards the end-goal of full implementations within practice. We also strongly encourage researchers to look further into XAI for other medical applications, especially for screening and classification purposes.

ACKNOWLEDGMENT

The authors are grateful for the comments and suggestions by the referees and the editor. Their comments and suggestions have greatly improved the article. This article represents the opinions of the authors and does not mean to represent the position or opinions of the American University of Sharjah.

REFERENCES

[1] S. Soffer, E. Klang, O. Shimon, N. Nachmias, R. Eliakim, S. Ben-Horin, U. Kopylov, and Y. Barash, "Deep learning for wireless capsule endoscopy: A systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 831–839, Oct. 2020.

[2] J. K. Min, M. S. Kwak, and J. M. Cha, "Overview of deep learning in gastrointestinal endoscopy," *Gut Liver*, vol. 13, no. 4, pp. 388–393, Apr. 2019.

[3] O. Attallah and M. Sharkas, "GASTRO-CADx: A three stages framework for diagnosing gastrointestinal diseases," *PeerJ Comput. Sci.*, vol. 7, p. e423, Mar. 2021.

[4] P. H. Smedsrud et al., "*Kvasir-Capsule*, a video capsule endoscopy dataset," *Sci. Data*, vol. 8, no. 1, p. 142, 2021.

[5] Y. Hao, Y. Wang, M. Qi, X. He, Y. Zhu, and J. Hong, "Risk factors for recurrent colorectal polyps," *Gut Liver*, vol. 14, pp. 399–411, Jul. 2020.

[6] N. Shussman and S. Wexner, "Colorectal polyps and polyposis syndromes," *Gastroenterol. Rep.*, vol. 2, pp. 1–15, Feb. 2014.

[7] Y. Komeda, H. Handa, T. Watanabe, T. Nomura, M. Kitahashi, T. Sakurai, A. Okamoto, T. Minami, M. Kono, T. Arizumi, M. Takenaka, S. Hagiwara, S. Matsui, N. Nishida, H. Kashida, and M. Kudo, "Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: Preliminary experience," *Oncology*, vol. 93, no. 1, pp. 30–34, 2017.

[8] S. Patino-Barrientos, D. Sierra-Sosa, B. Garcia-Zapirain, C. Castillo-Olea, and A. Elmaghraby, "Kudo's classification for colon polyps assessment using a deep learning approach," *Appl. Sci.*, vol. 10, no. 2, p. 501, Jan. 2020.

[9] J. S. Nisha, V. P. Gopi, and P. Palanisamy, "Automated colorectal polyp detection based on image enhancement and dual-path CNN architecture," *Biomed. Signal Process. Control*, vol. 73, Mar. 2022, Art. no. 103465.

[10] F. Younas, M. Usman, and W. Q. Yan, "A deep ensemble learning method for colorectal polyp classification with optimized network parameters," *Appl. Intell.*, vol. 53, no. 2, pp. 2410–2433, May 2022.

[11] L. F. Sánchez-Peralta, J. B. Pagador, A. Picón, Á. J. Calderón, F. Polo, N. Andraka, R. Bilbao, B. Glover, C. L. Saratxaga, and F. M. Sánchez-Margallo, "PICCOLO white-light and narrow-band imaging colonoscopic dataset: A performance comparative of models and datasets," *Appl. Sci.*, vol. 10, no. 23, p. 8501, Nov. 2020.

[12] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli, "Computer-aided classification of gastrointestinal lesions in regular colonoscopy," *IEEE Trans. Med. Imag.*, vol. 35, no. 9, pp. 2051–2063, Sep. 2016.

[13] R. Zachariah, J. Samarasena, D. Luba, E. Duh, T. Dao, J. Requa, A. Ninh, and W. Karnes, "Prediction of polyp pathology using convolutional neural networks achieves 'resect and discard' thresholds," *Amer. J. Gastroenterol.*, vol. 115, no. 1, pp. 138–144, Oct. 2019.

[14] J. W. Wei, A. A. Suriawinata, L. J. Vaickus, B. Ren, X. Liu, M. Lisovsky, N. Tomita, B. Abdollahi, A. S. Kim, D. C. Snover, J. A. Baron, E. L. Barry, and S. Hassanpour, "Evaluation of a deep neural network for automated classification of colorectal polyps on histopathologic slides," *J. Amer. Med. Assoc. Netw. Open*, vol. 3, no. 4, Apr. 2020, Art. no. e203398.

[15] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour, "Deep learning for classification of colorectal polyps on whole-slide images," *J. Pathol. Informat.*, vol. 8, no. 1, p. 30, Jan. 2017.

[16] X. Jia and M. Q.-H. Meng, "A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 639–642.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[18] Y. Fu, W. Zhang, M. Mandal, and M. Q.-H. Meng, "Computer-aided bleeding detection in WCE video," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 2, pp. 636–642, Mar. 2014.

[19] Y. Yuan, B. Li, and M. Q.-H. Meng, "Bleeding frame and region detection in the wireless capsule endoscopy video," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 2, pp. 624–630, Mar. 2016.

[20] P. Li, Z. Li, F. Gao, L. Wan, and J. Yu, "Convolutional neural networks for intestinal hemorrhage detection in wireless capsule endoscopy images," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1518–1523.

[21] H. Wahab, I. Mehmood, H. Ugail, A. K. Sangaiah, and K. Muhammad, "Machine learning based small bowel video capsule endoscopy analysis: Challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 143, pp. 191–214, Jun. 2023.

[22] 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.

[23] R. S. Latha, G. R. Sreekanth, G. Murugeasan, S. Aruna, B. Inbaraj, S. Kanivel, and S. Karthikeyan, "Deep learning based automatic detection of intestinal hemorrhage using wireless capsule endoscopy images," *Natural Volatiles Essential Oils*, vol. 8, no. 5, pp. 92–103, 2021.

[24] M. Sharif, M. Attique Khan, M. Rashid, M. Yasmin, F. Afza, and U. J. Tanik, "Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images," *J. Exp. Theor. Artif. Intell.*, vol. 33, no. 4, pp. 577–599, Feb. 2019.

[25] P. Padmavathi, J. Harikiran, and J. Vijaya, "Effective deep learning based segmentation and classification in wireless capsule endoscopy images," *Multimedia Tools Appl.*, vol. 82, pp. 1–25, May 2023.

[26] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Int. Conf. Learn. Represent. Workshop track*. Banff, Canada: Held in Rimrock Resort in Banff, Apr. 2014.

[27] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet," in *Proc. Int. Conf. Learn. Represent. Workshop track*, San Diego, CA, USA, 2015.

[28] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, pp. 1–23, Oct. 2016.

[29] M. Esmaeili, R. Vettukattil, H. Banitalebi, N. R. Krogh, and J. T. Geitung, "Explainable artificial intelligence for human-machine interaction in brain tumor localization," *J. Pers. Med.*, vol. 11, no. 11, p. 1213, Nov. 2021.

[30] H. Wang, M. Du, F. Yang, and Z. Zhang, "Score-CAM: Improved visual explanations via score-weighted class activation mapping," *CoRR*, vol. abs/1910.01279, pp. 1–11, Oct. 2019.

[31] 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.

[32] D. Mukhtorov, M. Rakhmonova, S. Muksimova, and Y.-I. Cho, "Endoscopic image classification based on explainable deep learning," *Sensors*, vol. 23, no. 6, p. 3176, Mar. 2023.

[33] A. A. Pozdeev, N. A. Obukhova, and A. A. Motyko, "Automatic analysis of endoscopic images for polyps detection and segmentation," in *Proc. IEEE Conf. Russian Young Researchers Electr. Electron. Eng. (EIConRus)*, Jan. 2019, pp. 1216–1220.

[34] R. Fonollá, F. van der Sommen, R. M. Schreuder, E. J. Schoon, and P. H. N. de With, "Multi-modal classification of polyp malignancy using CNN features with balanced class augmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 74–78.

[35] T. Dhar, N. Dey, S. Borra, and R. S. Sherratt, "Challenges of deep learning in medical image analysis—Improving explainability and trust," *IEEE Trans. Technol. Soc.*, vol. 4, no. 1, pp. 68–75, Mar. 2023.

[36] D. Song, J. Yao, Y. Jiang, S. Shi, C. Cui, L. Wang, L. Wang, H. Wu, H. Tian, X. Ye, D. Ou, W. Li, N. Feng, W. Pan, M. Song, J. Xu, D. Xu, L. Wu, and F. Dong, "A new xAI framework with feature explainability for tumors decision-making in ultrasound data: Comparing with grad-CAM," *Comput. Methods Programs Biomed.*, vol. 235, Jun. 2023, Art. no. 107527.

[37] A. Caroppo, P. Siciliano, and A. Leone, "An expert system for lesion detection in wireless capsule endoscopy using transfer learning," *Proc. Comput. Sci.*, vol. 219, pp. 1136–1144, Jan. 2023.

[38] R. K. Dey, M. E. Rana, and V. A. Hameed, "Analysing wireless capsule endoscopy images using deep learning frameworks to classify different GI tract diseases," in *Proc. 17th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2023, pp. 1–7.

[39] J. Bernal et al., "Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge," *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.

[40] M. Amirthalingam and R. Ponnusamy, "Wireless capsule endoscopic image classification using seeker optimization algorithm with deep learning model," in *Proc. 7th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2023, pp. 110–116.

[41] M. Amirthalingam and R. Ponnusamy, "Improved water strider optimization with deep learning based image classification for wireless capsule endoscopy," in *Proc. 3rd Int. Conf. Artif. Intell. Smart Energy (ICAIS)*, Feb. 2023, pp. 851–857.

[42] D. Lehmann and M. Ebner, "Subclass-based undersampling for class-imbalanced image classification," in *Proc. VISIGRAPP*, 2022, pp. 493–500.

[43] A. Srivastava, N. K. Tomar, U. Bagci, and D. Jha, "Video capsule endoscopy classification using focal modulation guided convolutional neural network," in *Proc. IEEE 35th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2022, pp. 323–328.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016.

[45] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015.

[47] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

[49] 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," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[51] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, pp. 1–10, May 2017.

[52] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, pp. 1–10, Aug. 2016.

[53] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. Image Process.*, vol. 30, pp. 5875–5888, 2021.

[54] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. 32nd AAAI Conf. Artif. Intell. 13th Innov. Appl. Artif. Intell. Conf. 8th AAAI Symp. Educ. Adv. Artif. Intell. (AAAI/IAAI/EAAI)*. Palo Alto, CA, USA: AAAI Press, 2018, pp. 1–9.

[55] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *Ann. Appl. Statist.*, vol. 9, no. 3, pp. 1350–1371, Sep. 2015.

[56] G. Alicioglu and B. Sun, "A survey of visual analytics for explainable artificial intelligence methods," *Comput. Graph.*, vol. 102, pp. 502–520, Feb. 2022.

[57] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *CoRR*, vol. abs/1512.04150, pp. 1–10, Dec. 2015.

[58] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[59] J. Gildenblat. *Advanced AI Explainability for Computer Vision. Support for CNNs, Vision Transformers, Classification, Object Detection, Segmentation, Image Similarity and More*. Accessed: Jun. 4, 2023. [Online]. Available: https://github.com/jacobgil/pytorch-grad-cam

[60] M. Mehta, V. Palade, and I. Chatterjee, *Explainable AI: Foundations, Methodologies and Applications*. Germany: Springer, Nov. 2022.

DARA VARAM was born in Tehran, Iran, in 2001. He is currently pursuing the dual degree in computer engineering and mathematics with a minor in data science with the American University of Sharjah (AUS), United Arab Emirates. Since 2021, he has been an Undergraduate Research Assistant with the Department of Computer Science and Engineering and the Department of Mass Communication. His research interests include artificial intelligence, deep learning, computer vision, and optics. His awards and honors include his membership in the Eta Kappa Nu Honor Society and the Engineering Honors Society, AUS.

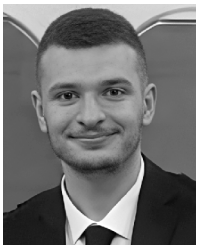
ROHAN MITRA (Member, IEEE) was born in Kolkata, India, in 2000. He is currently pursuing the dual degree in computer science and mathematics with a minor in data science with the American University of Sharjah (AUS), United Arab Emirates. He has been an Undergraduate Research Assistant in machine learning with the Department of Computer Science and Engineering, the Department of Mathematics, and the Department of Mass Communication, since 2020. His research interests include artificial intelligence, deep learning, computer vision, and reinforcement learning. His awards and honors include his membership in the Eta Kappa Nu Honor Society, the Engineering Honors Society, and the Alpha Lambda Delta Honor Society, AUS. Moreover, he received a 100% scholarship for the entirety of the double degree from AUS, as well as been awarded an Undergraduate Research Grant for further research.

MERIAM MKADMI was born in Mississauga, Canada, in 2001. She is currently pursuing the B.S. degree in computer science and the B.S. degree in mathematics with a minor in data science with the American University of Sharjah (AUS), United Arab Emirates.

Her research interests include artificial intelligence, deep learning, and computer vision. She is a member of the Engineering Honors Society and the IEEE Eta Kappa Nu Honor Society. She was a recipient of many scholarships, as well as an undergraduate research grant to support her research interests. She was the Chair of the IEEE-CS Student Chapter, AUS.

RADI AMAN RIYAS was born in Dubai, United Arab Emirates, in 2002. He is currently pursuing the dual degree in computer science and mathematics with the American University of Sharjah (AUS), United Arab Emirates.

He is a Machine Learning Research Assistant with the Department of Industrial Engineering. His current research interests include deep learning, control theory, and origin-of-life simulations. He is the Chair of the IEEE Computer Society Chapter, AUS.



DIAA ADDEEN ABUHANI was born in Amman, Jordan, in 2001. He is currently pursuing the B.S. degree in computer engineering with a double minor in computer science and data science with the American University of Sharjah (AUS), United Arab Emirates. His current research interests include deep learning, explanatory artificial intelligence (XAI), the Internet of Things (IoT), computer vision, and precision farming using UAVs. He is a member of the Engineering Honors Society and the IEEE Eta Kapa Nu Honors Society.

SALAM DHOU (Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from the Jordan University of Science and Technology, Jordan, in 2004 and 2007, respectively, and the Ph.D. degree in electrical and computer engineering from Virginia Commonwealth University, USA, in 2013. She was a Postdoctoral Research Fellow with the Department of Radiation Oncology, Harvard Medical School, USA, from 2013 to 2016. She is currently an Assistant Professor with the Department of Computer Science and Engineering and the Biomedical Engineering Graduate Program, American University of Sharjah (AUS), United Arab Emirates. Her research interests include medical imaging and informatics, machine learning, and data mining.

AYMAN ALZAATREH joined the Department of Mathematics and Statistics, American University of Sharjah (AUS), in August 2017. Previously, he was with institutions, including Austin Peay State University, USA, and Nazarbayev University, Kazakhstan. His current research interests include generalizing statistical distributions arising from the hazard function, statistical inference of probability models, characterization of distributions, bivariate and multivariate weighted distributions, and data mining.