Computer Science and Information Systems 19(1):277–307 https://doi.org/10.2298/CSIS201030049S

Explainable Information Retrieval using Deep Learning for Medical images ⋆

Apoorva Singh[1], Husanbir Singh Pannu[1], and Avleen Malhi[2]

[1] Thapar Institute of Engineering and Technology
Patiala India 147004
apoorvasingh.singh1993@gmail.com, hspannu@thapar.edu
[2] Bournemouth University
Fern Barrow, Poole BH12 5BB, UK
amalhi@bournemouth.ac.uk

Abstract. Image segmentation is useful to extract valuable information for an efficient analysis on the region of interest. Mostly, the number of images generated from a real life situation such as streaming video, is large and not ideal for traditional segmentation with machine learning algorithms. This is due to the following factors (a) numerous image features (b) complex distribution of shapes, colors and textures (c) imbalance data ratio of underlying classes (d) movements of the camera, objects and (e) variations in luminance for site capture. So, we have proposed an efficient deep learning model for image classification and the proof-of-concept has been the case studied on gastrointestinal images for bleeding detection. The Explainable Artificial Intelligence (XAI) module has been utilised to reverse engineer the test results for the impact of features on a given test dataset. The architecture is generally applicable in other areas of image classification. The proposed method has been compared with state-of-the-art including Logistic Regression, Support Vector Machine, Artificial Neural Network and Random Forest. It has reported F1 score of 0.76 on the real world streaming dataset which is comparatively better than traditional methods.

Keywords: machine learning, explainable AI, image processing, medical images, capsule endoscopy.

1. Introduction

Machine learning (ML) and artificial intelligence (AI) systems imitates the humans way of learning by associating the cognitive ability and pattern recognition. AI systems are quite complex and intend to mimic human intelligence and automatic learning; ML is about automatic decision making and future predictions through given data distribution and pattern recognition and without explicit coding. Image classification using ML for commercial purposes is good but is still needs improvement in complex images such as medical imaging including cancer cells, endoscopy, x-ray and MRI images. Thus human capabilities and expertise is still superior in these fields as compared to ML.

ML models are flexible, efficient and well-generalized but they are opaque and obscure to understand about how they work. Its power of reasoning is thus limited due to

⋆ This is an extended version of a conference paper [44]: Explaining Machine Learning-Based Classifications of In-Vivo Gastral Images

inability to layout the road map of the decision making phenomenon in case of testing dataset. Thus machine learning model must be able to give justification about the model rationale which can be evaluated by experts to audit the decision making factors. There should be a quantified phenomenon to see how the machine reasons for an outcome in contrast to a human expert for potential conflicts and legal norms. Explainable artificial intelligence (XAI) provides such a formal explanation by the model agnostic interpretation against action taken or decision made by ML model, given the test data and features involved. Figure 1 shows the basic XAI framework with valid questions to be answered

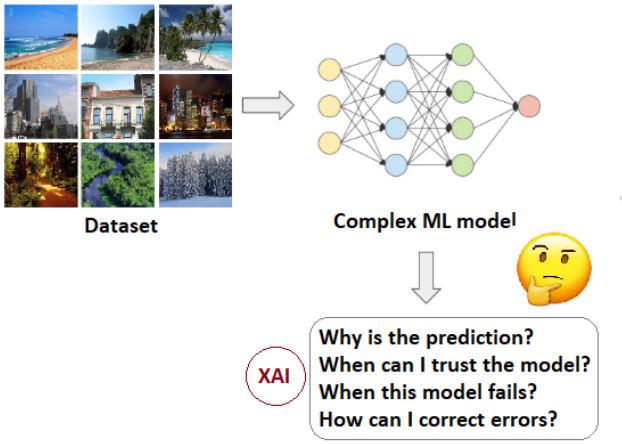

Fig. 1. Basic motivation of explainable artificial intelligence (XAI) is to answer the four questions. The inputs to XAI are the trained model and test dataset.

by the underlying ML model. They include questions like (a) What is the prediction (b) What is the credibility of model (c) Conditions of model failure (d) How to correct errors if any? Afterwards, it justifies the recommended decision through explanation interface. Now the motivation question to consider is that why to use CNN, gastral images and XAI?

1. Actually, CNN imitates human brain to learn by automatic features extraction with numerous layers and neurons as said by father of deep learning Professor Geoffrey Hinton. Professor believed that neural networks are not stuff of science fiction or toil in obscurity. Neural networks are simplified model of how the brain works [41].
2. Gastral images obtained from capsule endoscopy for example, are complex due to movement of camera, organs, noise, compression for transmission. Thus if a model can learn these obscure images then it would also work well on other real world application involving similar constraints and assumptions.
3. XAI explains the test results in regards to feature proportions involved in training just like a psychologist who explores the reasons behind humans' way of thinking. Otherwise the human mind is invincible to traverse for how it works and processes information.

1.1. Convolutional Neural Network (CNN)

Deep learning is a subdivision of machine learning involving neural networks that work similarly to the human brain and are capable of learning from unstructured data such as images [58]. The strength of deep learning is that the low-level features of an image (edges or textures) are compared and connected to the higher-level features (shapes, objects) automatically and autonomously by the model using enormous amount of training data. It involves the hierarchy of concepts for information extraction in form of image features. In traditional machine learning algorithms like SVM, Linear regression, Random Forest, the features have to be extracted from the images manually, whereas in deep learning like CNN, features are learned from the raw data automatically by the network [31].

A CNN is a feed-forward neural network that is used to identify the complicated features in the dataset. It can devise and derive features from unprocessed data automatically and can work on massive amounts of data. CNN performs remarkably good in the fields of images-analysis, pattern-detection, edge-detection, image/object recognition, powering vision in robots, for self-driving vehicles, etc. It also reduces computational burden by offering the automatic feature extraction and briefly explained in the subsection below. There are a variety to deep learning models available for CNN architecture. Similar to the CNN model proposed by Jia et al. [37], each of the model CNN involved in the ensemble has been comprised of eight-layers that involve three convolutional layers (C1-C3), two fully-connected layers (FC1, FC2) and three pooling layers (MP1-MP3). In our execution, rectified linear units (ReLUs) are used as the activation function in convolutional layers (C1-C3) and the first fully-connected layer (FC1). Max-pooling is used in the pooling layers (MP1-MP3) to detect the maximal activation over input patches. Lastly, the output of the second fully-connected layer (FC2) consists of two neurons (bleeding and normal) and can be activated by a soft-max regression function, which is defined as:

$$f_\theta(x^{(i)}) = \begin{bmatrix} Q(y=1|x^{(i)};\theta) \\ Q(y=2|x^{(i)};\theta) \\ \vdots \\ Q(y=M|x^{(i)};\theta) \end{bmatrix} = \frac{c}{b} \tag{1}$$

where

$$c = \begin{bmatrix} \exp(\theta^{(1)\top} x^{(i)}) \\ \exp(\theta^{(2)\top} x^{(i)}) \\ \vdots \\ \exp(\theta^{(M)\top} x^{(i)} \end{bmatrix} \tag{2}$$

and

$$b = \sum_{k=1}^{M} \exp(\theta^{(k)\top} x^{(i)}) \tag{3}$$

So

$$f_\theta(x^{(i)}) = \frac{c}{b} \tag{4}$$

where $x^{(i)} \in \Re^n$ are the input attributes with the corresponding labels $y(i)$. M is the number of classes. The model parameters $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)} \in \Re^n$ are trained to minimize the loss function:

$$L(\theta) = -(1)\left[\sum_{m=1}^{t}\sum_{n=1}^{K} 1\left\{y^{(m)} = n\right\} \log \frac{\exp(\theta^{(n)\top} x^{(m)})}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x^{(m)})}\right] \tag{5}$$

where t denotes the size of the training set. Particularly, in the binary classification setting, we have $y^{(i)} \in \{0, 1\}$ and $K = 2$. CNN derives the features from images automatically, which results in a reduced computational burden and reducing the semantic gap between humans way of perceiving and algorithmic approach. Image features are learned during the training of the network on the image dataset. One more benefit of using CNN is that only the number of filters and the filter size is required to be defined, whereas the values of the filters are determined by CNN automatically during the training phase. Unlike most of the other ML techniques, object detection in images is carried out by CNN regardless of the location of the object to be recognized. Pooling feature of CNN also prevents overfitting of the network.

1.2. Explainable Artificial Intelligence

Explainable artificial intelligence is getting a lot of attention nowadays. Machine learning algorithms have been used for medical imaging but these models do not explain the assessment they make. Humans cannot trust these models since they do not understand the reason of their assessment. Although there is an increasing number of works on interpretable and transparent machine learning algorithms, they are mostly intended for the technical users. Explanations for the end-user have been neglected in many usable and practical applications. Many researchers have applied the explainable framework to the decisions made by model for understanding the actions performed by a machine. There are many existing surveys for providing an entry point for learning key aspects for research relating to XAI [6]. Anjomshoae et al. [10] gives the systematic literature review for literature providing explanations about inter-agent explainability. The classification of the problems relating to explanation and black box have been addressed in a survey conducted by Guidotti et al. [32] which helped the researchers to find more useful proposals. Machine learning models can be considered reliable after integration of explainability feature for the expert analysis and retraining of the model. Contextual Importance and Utility has a quite significance in explaining the machine learning models by giving the rules for explanation [26]. Framling et al. provides the black box explanations for neural networks with the help of contextual importance utility [25][27].

There are many methods used for providing the explanations for example; LIME (Local Interpretable Model-Agnostic Explanations) [3], CIU (Contextual Importance and Utility) [26], ELI5 [2], Skater [5], SHAP (SHapley Additive exPlanations) [4] etc. Most of them are the extensions of LIME which is an original framework and approach being proposed for model interpretation. These techniques provide model prediction explanations with local interpretation, model prediction values with shape values, building interpretable models with surrogate tree based models and much more. Contextual Importance (CI) and Contextual Utility (CU) explains the prediction results without transforming the model into an interpretable one. These are numerical values represented as visuals and natural language form for presenting explanations for individual instances [26]. The CIU has been used by Anjomshoae et al. [9] to explain the classification and prediction results made by machine learning models for Iris dataset and Car Pricing dataset where the authors have CIU for justifying the decisions made by the models. The prediction results are explained by this method without being transformed into interpretable model. It yields the explanations for linear as well as non linear models demonstrating the felexibility of the method.

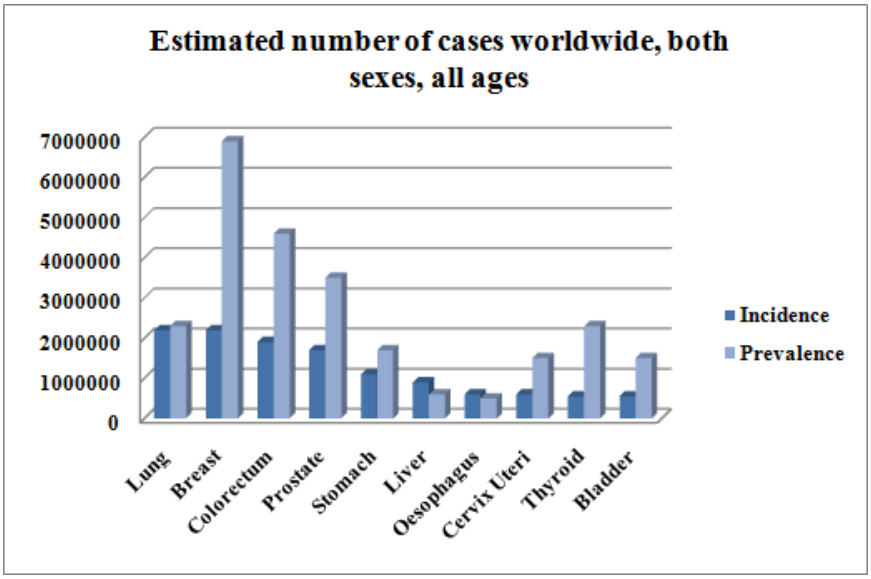

Fig. 2. The estimated incidence and prevalence of worldwide cancer cases [17], [49]. Most of them are related to gastrointestinal organs

1.3. Capsule Endoscopy

In Gastroenterology (GI), gastric cancer is the fifth most common cancer worldwide and seventh most prevalent in accordance to the GLOBOCAN 2018 as shown in Figure 2. In few states of India such as Tamil Nadu, Assam, Kerala, and Karnataka, the malignant

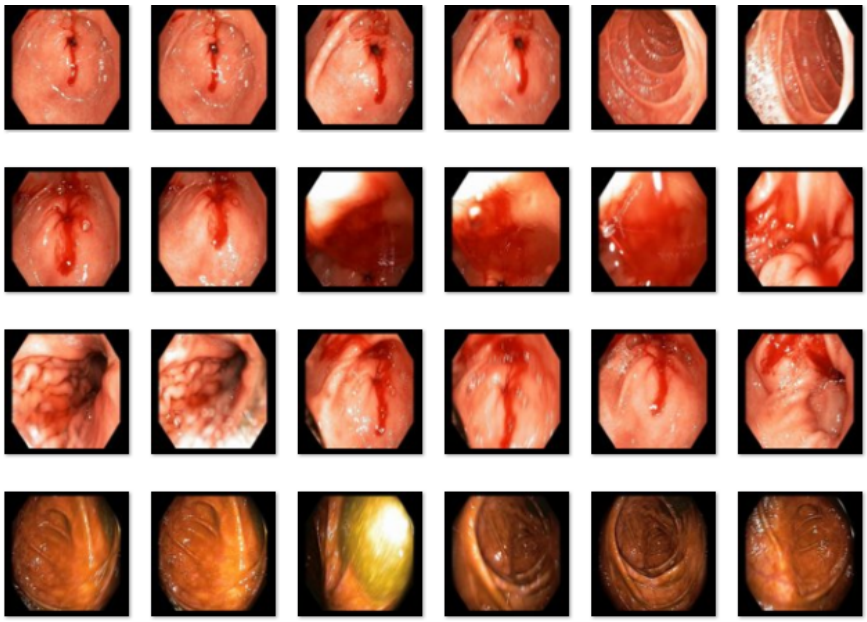

Fig. 3. Sample video shots of GIngastrointestinal tract with bleeding symptoms. Capsule endoscopy images collected from PSRI hospital, New Delhi - www.psrihospital.com

tumors or cancers are most commonly present in the squamous cells of the esophageal carcinoma section of the digestive tract [18]. For males, 10% of the total cancers is Colorectal cancer and that makes it the third most prevalent cancer in males with cases of 663,000 all over the world. Whereas, it is the other most frequent cancer in females with cases of 570,000, which is 9.4% of the total cases globally [49], [17].

GI tract comprises of several organs such as digestive canal, throat or esophagus, liver, duodenum, bile ducts, pancreas, gallbladder, small intestine, and large intestine, colon, rectum. Gastroenterology also addresses the complications that may harm these organs including polyps, ulcers, cancer, and esophageal reflux. Wireless capsule endoscopy (WCE) technique was first introduced in year 2000 [36]. The small intestine is one of the complex organs to diagnose and heal without conducting surgery. CE supports physicians to see inside the regions of our body that are not readily reached with traditional endoscopy. The video obtained by the pill-sized camera used in WCE is carefully observed by the doctor for irregularities inside the digestive system. The manual review of WCE video is not time-efficient, with an average reading time of 45–120 minutes approximately [21], [**?**]. RAPID software is available with the PillCam kit to automatically detect the bleeding frames out of all the frames from video captured by the camera. The efficiency obtained by this software is not satisfactory [36], [22], [50]. It may also skip the frames with inactive bleeding or frames with blood spots of very small sizes. As the results obtained by PillCam's RAPID software algorithm are not efficient, there is a need for a better algorithm to detect the anomalies in WCE frames [55].

Our goal is to propose an automated soft-computing technique for the detection of presumed frames that can have the appearance of bleeding. This may significantly lessen the evaluation time while the ultimate judgment is still left to the endoscopy experts. The paper has been organised as follows: next section 2 is about the contemporary literature survey for research motivation and gaps; section 3 is the proposed methodology; section 4 is performance metrics; section 5 is about results of the case study; section 6 is about conclusion and future directions.

2. Literature Review

This section studies the motivational state-of-art work done in the field of image segmentation to detect plausible abnormalities like polyp, tumor, ulcer and bleeding. Table 1 gives the comparison of the existing machine learning approaches proposed for the endoscopic bleeding detection whereas figure 4 gives the classification of the different approaches used for the anomaly detection in the digestive tract. Most of the investigated modern techniques based on this comparison have worked on the automated detection of abnormalities seen in the GI Tract Endoscopy.

Table 1. State-of-the-art comparison for the bleeding detection in endoscopic images

Type	Study	Year	Dataset	Method	Results
ML	[48] Obukhova et al.	2019	Bleeding frames in KVASIR (open), 8000 images	Block-based segmentation and color characteristics	95% accuracy
ML	[53] Tuba et al.	2019	Bleeding framesinF. Deeba, "Bleeding images and corresponding ground truth of CE	Texture and color features (HSI, CIE ULBP), GrowCut	0.85 Dice similarity coefficient, 0.092 misclassification error images, 50 images
ML	[56] Jia et al.	2018	Bleeding frames in 1000 WCE images from random patients.	Superpixel-color histogram, KNN	0.9922 accuracy
ML	[30]Ghosh et al.	2018	Bleeding zones in Kid: Koulaouzidis-iakovidis database for capsule	Semantic segmentation, SegNet, CNN	94.42% accuracy Endoscopy, 335 images
ML	[13] Bchir et al.	2018	Multiple bleeding frames in Imaging PillCam, 1275 frames	Fuzzy C-means clustering, KNN	90.92% accuracy
ML	[28] Ghosh et al.	2018	Bleeding frames in CE (Online), 2350 images	Color Histogram of Block Statistics	97.85% accuracy

ML	[52]. Sivakumar et al.	2018	Bleeding frames	Superpixel segmentation, Semi-Naïve Bayesian classifier	N/A
ML	[29] Ghosh et al.	2017	Bleeding frames in The capsule endoscopy website (public), 2350 frames from 32 WCE videos	Cluster based statistical feature extraction	97.05% precision
CNN	[33] Hajabdollahi et al.	2019	Bleeding regionsinF. Deeba, "leeding images and corresponding ground truth of CEimages	Multi-layer perceptron, CNN	AUC-ROC 0.97, DICE for CNN 0.869 for MLP = 0.831
CNN	[38] Jia et al.	2017	Bleeding frames in 1500 WCE images	Handcrafted features based CNN	F1 score 0.9285
CNN	[37] Jia et al.	2016	Bleeding frames in 10,000 WCE images	Deep CNN with SVM	Recall 99.2%, F1 Score 99.5%
Other	[34] He et al.	2018	Hookworm frames in West China Hospital, 440K images	CNNs (Edge extraction and Hookworm classification network)	88.5% accuracy
Other	[54] Vieira et al.	2019	Small bowel angioectasiasinKID (public), 27 images, and PillCam, MiroCam in Hospital of Braga (Portugal), 300 frames	Maximum a Posteriori, Expectation-Maximization	Sensitivity 96%, Specificity 94.08%, Accuracy 95.58%
Other	[7] Alaskar et al.	2019	Ulcer in Dr. Khoroo'sMedical Clinic (Online available), 1875 images	AlexNet and GoogleNet CNN	100% accuracy with learning rate 0.0001
Other	[12] Aoki et al.	2019	Erosions and Ulcer frames in The University of Tokyo Hospital, Japan, 15800 images	Deep CNN with a Single Shot Multibox Detector	91.5% accuracy
Other	[47] Nawarathna et al.	2019	Mucosal abnormality in MiroCam WCE images	Filter bank, local binary patterns, Textons histogram	Recall 92%, Specificity 91.8%

Other	[42] Leenhardt et al.	2018	GI Angiectasia during small bowel in 6360 frames from pre-med students	CNN-based semantic segmentation	Sensitivity 100%, Specificity 96%
Other	[23] Diamantis et al.	2019	GI Abnormalities in Endovis challenge, 10,000 images, and KID, 2352 images	Look-Behind Fully CNN (LB-FCN)	AUC 93.5%
Other	[14] Billah et al.	2017	Polyp frames in Endoscopic Vision Challenge, more than 14,000 images	Color wavelet, CNN and SVM	Accuracy 98.34%, Sensitivity 98.67%, Specificity 98.23%
Other	[20] Deeba et al.	2017	Bleeding frames in PillCam SB1 and PillCam SB2, 8872 images	SVM ensemble and exhaustive feature selection	Accuracy 95%, Specificity 95.3%, Sensitivity 94%

2.1. Machine Learning Techniques

Various machine learning techniques have been used for automation of the bleeding detection in WCE images. In [48], authors proposed a method for automatic feature extraction and detection of bleeding in endoscopy images. The endoscopy images are segmented using block-based segmentation. The local features are discovered using color characteristics. Different algorithms are investigated in this approach to classify the bleeding and non-bleeding images. The performance parameters are also calculated to test the effectiveness of these algorithms. The accuracy obtained by the proposed approach is 95%. In [53], this approach presented bleeding detection in WCE images based on region-based feature extraction. This method extracts features from HSI and CIE color spaces. Authors used a uniform library binary pattern to label each region. The secondary set of features can be extracted from the grayscale image. Classification of regions is done by support vector machine (SVM) into three categories, namely, non-bleeding region, bleeding region, and background. GrowCut algorithm is used for the concluding segmentation of CE images.

Xing et al. [56] introduced a three-step algorithm for automated detection of bleeding in endoscopy images. Authors have done Key-frame extraction and edge removal as the first step of preprocessing. In the second step, they separated the bleeding images from the dataset of all the frames using the KNN classifier, applying the concept of principle color spectrum by utilizing the superpixel color histogram feature. In the last step, the segmentation of bleeding regions from the various color spaces is executed by securing a 9-D color feature vector at the superpixel feature. The accuracy attained is 0.99.

The system proposed by [13], focused on identifying the multiple blood specks in the several frames captured from the WCE video. To overcome the performance degradation due to the small size of the region of interest, the authors suggested an unsupervised ML

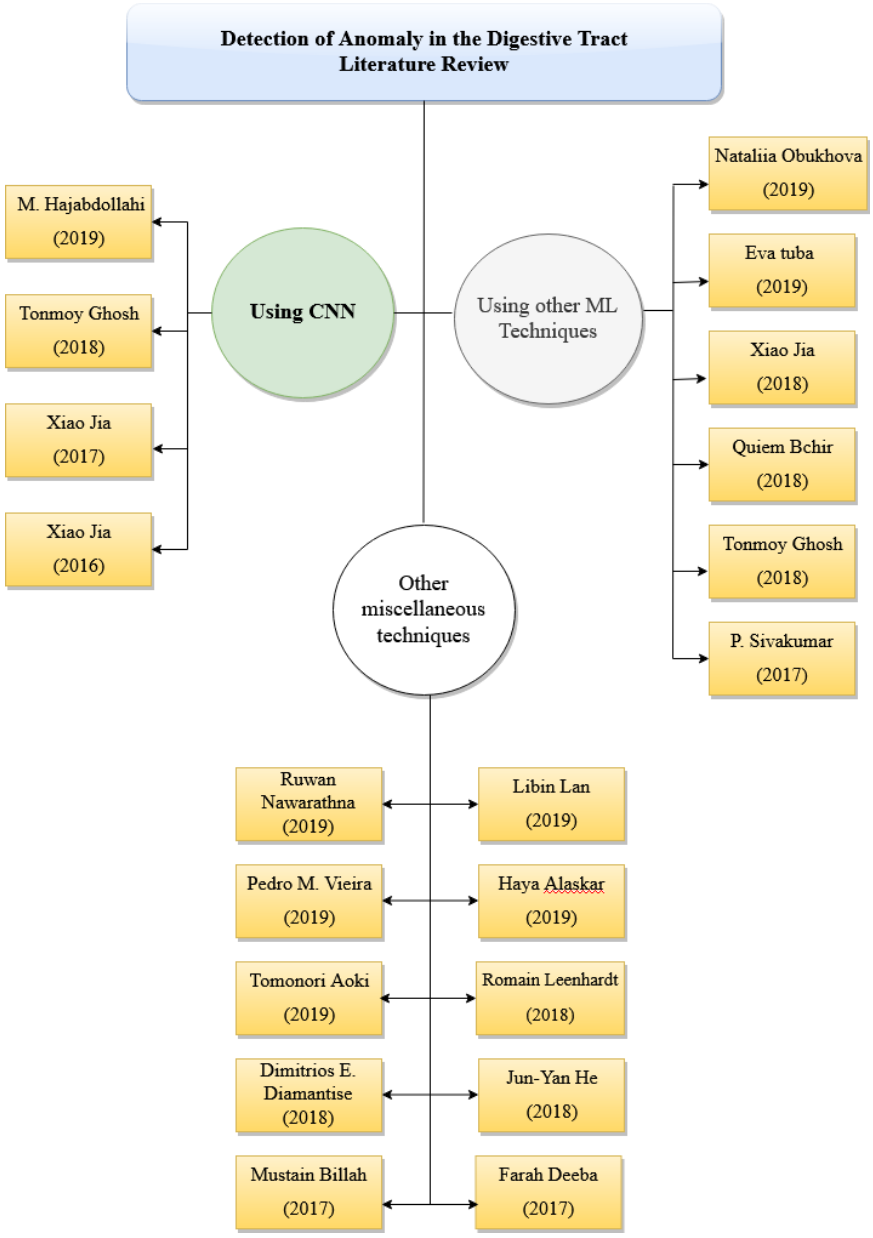

Fig. 4. Literature review organization

technique. It breaks the principal classification query into many confined classification queries. This technique cluster the training set using fuzzy C-means, then implement improvised KNN on the determined centers rather than the entire training dataset. Authors have also analyzed the performance and results of the proposed algorithm as opposed to the typical KNN and SVM.

In the paper [52], Naive Bayes classifier and superpixel segmentation are used for automated obscure bleeding disclosure in WCE video dataset. They used the color histogram for region discovery and feature extraction. In the final step, they adopted an improvised semi naive bayesian classifier. In [28], authors have classified bleeding and non-bleeding frames in wireless CE images utilizing color histogram of block statistics. Local feature extraction is done using a WCE image block, preferably of an individual pixel. For the contrasting color panels of RGB color space, index values are defined. So, the authors used index values to extract the color histogram. Color histogram is useful in securing distinct color texture characteristics. Feature reduction using color histogram and principal component analysis is adopted to decrease the dimension of these local features. Extracted local features that do not result in any computational strain provides the blocks with bleeding regions. Authors used a public dataset of 2350 images that renders 97.85% accuracy.

In [29], authors have worked on the system that distinguishes the bleeding images and regions from the WCE dataset. WCE images are preprocessed to convert into a color space defined by green to red pixel-ratio. These transformed images are used to obtain various analytical features from the overlying spatial blocks. These various blocks are then clustered into two clusters using K-means based clustering (unsupervised). These clusters are used to obtain cluster-based features. These features combined into a global feature is used along with differential cluster-based features to detect blood zone frames using an SVM (supervised learning classifier). The proposed system achieved 97% precision.

2.2. CNN based segmentation techniques

Many authors have proposed CNN based methods for WCE bleeding detection mechanisms. Authors proposed a way in [30] for detection of bleeding regions in CE images employing a semantic segmentation based on deep neural network, called SegNet. CNN is trained using the successive layers of SegNet. Bleeding regions are detected in CE images by segmenting the test images on the trained CNN. The efficiency achieved is 94.42%. In [33], authors advised a simplified neural network (NN) for the detection of bleeding frames by performing automatic bleeding regions segmentation on the CE dataset. Fitting color channels are chosen as inputs to the neural network. An multi-layer perceptrons and CNN are applied to conduct image classification individually. They decreased the number of computational operations. The performance of the recommended systems is assessed using the DICE score. The area under the receiver operating curve (AUC-ROC) is 0.97. Due to significantly fewer computations, CNN is proved to be more beneficial than multi-layer perceptron.

Authors in the paper [37] presented a high-level detection system for the bleeding frames in the WCE image dataset. This system is implemented using a deep CNN that detects both active and inactive frames. Authors have designed CNN to have 8 layers. The network composes 3 convolutional layers, 3 pooling layers, and two fully connected layers. The ReLU or the rectifier function is implemented at the convolutional layers and

the first FC layer to increase non-linearity in the network. Images are made of different objects that are not linear to each other. Without applying this activation function, the image classification is treated as a linear problem, while it is in actual a non-linear one. Pooling layers implement Max-Pooling to preserve the main features while also reducing the size of the image. This helps reduce overfitting. One of the main causes for the overfitting to occur is that too much information fed into CNN. Especially if that information is not relevant in classifying the image. SVM is more beneficial in the case where the user wants to depreciate the entropy loss for prediction. So, CNN is devised by substituting the SoftMax regression function at the secondary FC layer with an SVM classifier. But this proposed system resulted in complex computations and required a large dataset of designated images for training the CNN. So, the authors presented a way in the paper [38] that implements the deep-learning technique of CNN with handcrafted features that are obtained using a k-means clustering technique. This model is focused on detection of frames with active and inactive bleeding. This approach reduces the computational cost incurred in the training of CNN.

2.3. Other techniques

Authors in [47] have worked on a computerized way of finding abnormalities in the endoscopy images. These abnormalities can be erosion, erythema, ulcerations, polyp, bleeding, etc. This proposed method examines images for varied textures so that, it can differentiate abnormal images from the normal ones effectively. The distribution of different textures in an endoscopy image can be captured using a textons histogram. It is done by applying a FB (filter bank) and LBP (local binary patterns). This proposed approach gives 92% recall and 91.8% specificity on WCE images. In [40], authors have worked on a computer-aided system that uses various approaches to detect the abnormalities in the images obtained from CE. Authors have employed CNN, region recommendation, transfer learning. In the first step, they have used a CascadeProposal to recommend high-recall regions and abnormal frames. In the second step, the authors used a multi-regional combination technique to detect the regions of interest and have also operated a salient region segmentation approach to catch certain region spots. For object boundary filtration, a dense-region fusion algorithm is applied. And lastly, to increase the efficiency of the proposed model, transfer learning tactics are exercised in CNN.

Authors of [23] presented a computer-aided look-behind fully CNN (LB-FCN) algorithm to automatically catch the anomalies in CE images. It uses blocks of parallel convolutional layers with varied filter dimensions to derive the multi-scale features from WCE images. All the LB linked features are combined with the features deduced from prior layers. As LB-FCN has fewer free parameter as compared to conventional CNN, it makes it much easier to train the network on smaller datasets. The AUC performance of LB-FCN achieved is 93.5%. In [20], they have aimed at reducing the analysis time of the WCE video frames by presenting a computer-aided approach that automatically identifies the abnormal frames. They have worked on an ensemble of two SVMs that are based on HSV and RGB color spectrums. Feature selection and parameter tuning are done by using a nested cross-validation approach. For the betterment of performance, exhaustive analysis is carried out to decide the best feature sets. The dataset used comprises of 8872 WCE frames. This fusion system renders an accuracy of 95%, specificity of 95.3% and sensitivity of 94%. A CNN is suggested in [42] for the GI angioectasia detection during

small bowel in CE images. Local features are extracted through deep feature extraction using an approach of segmentation of images based on semantics. Authors created a semantic segmentation-based CNN for classification of GI angioectasias. The sensitivity and specificity achieved is 100% and 96% respectively.

The work done in [54] aims for an automated way for the detection of angioectasias in WCE image dataset. This approach depends on the automatic separation of a region of importance. That region is chosen by applying a module for the task of image segmentation based on the approach of Maximum a Posteriori where a new hastened variant of the Expectation-Maximization is also advised. This proposed method attained sensitivity and specificity values of 96% and 94.08% respectively with 95.58% accuracy in a database comprising 800 WCE frames designated by two gastroenterologists. In this paper [7], they have conducted ulcer and lesion detection and classification in WCE dataset employing two pre-trained CNN, GoogleNet and AlexNet. These two networks perform object classification to obtain ulcer and non-ulcer frames. Due to a huge number of layers in GoogleNet, AlexNet resulted in double the efficiency of GoogleNet for training. The efficiency of both networks is enhanced by tuning the parameters. It is also found in this study that higher the learning rate of the network, higher is the resulting accuracy. The learning rate of 0.0001 renders adequate results for both the networks.

AlexNet attained 100% accuracy with the rate of 0.001. Authors in [12] trained a deep CNN to distinguish ulcers and erosions in small bowel CE images automatically. This CNN is based on an single shot multiBox detector that holds 16 layers. The CNN is trained using SSD on the 5,360 images. For the testing phase, 10,440 WCE images are fed to CNN, out of which, 440 are of erosions or ulcers. This system renders an accuracy of 91.5%. Whereas, in the paper [14], they have focused on decreasing the misidentification rate for a polyp in CE images. This will support the professionals in finding the most significant regions to pay consideration. Features are deduced using color wavelet and CNN. These extracted features are then fed to a train an SVM. SVM will classify the CE frames into the polyp region and normal frames classes. They achieved 98.34% accuracy. The study performed by [34] has focused on detecting a hookworm abnormality in wireless capsule endoscopy images. They have adopted the deep learning algorithm to recognize the tube-like pattern of hookworm. For the better activity of the classification, two neural networks are employed, edge extraction CNN and hookworm classification CNN. Both the CNNs are seamlessly integrated into the recommended system to evade edge feature caching. The edge extraction CNN provides the tubular regions and the hookworm CNN gives the feature maps. Both the results are integrated into the pooling layers to produce an intensified feature map accentuating tubular region and achieved 88.5% accuracy.

Our research concentrates on the automatic bleeding detection in capsule endoscopy videos using a convolutional neural network. Literature review organization in terms of techniques has been shown in Figure 4 while highlighting the CNN based approaches. CNN is fast, efficient, and it needs limited preprocessing of the images [23], [7], [12]. We have used CNN in the proposed method for the detection of the bleeding frames in WCE images along with explanation of test results and the impact of involved features.

3. Methodology

A computerized system for bleeding detection in WCE images is proposed to catch the presence of a threat in the digestive tract that caused the bleeding. A dataset of WCE videos with frames holding both bleeding and non-bleeding frames is collected (from PSRI hospital, New Delhi) and used for the proposed approach.

3.1. Proposed classification approach

The basic layout of the proposed methodology is shown in Figure 5. It shows preprocessing, denoising and image learning. The obtained WCE video dataset is fed to the VLC software to extract images. Figure 3 shows the snapshot of the extracted image dataset. The number of WCE images extracted for the dataset is 2,621 with 505 bleeding and 2,116 normal frames. All the images extracted are resized consistently to prepare them for the proposed model. High-resolution images involve more computations and higher memory specifications. If the input is a scaled-down variant of the bigger images, then determining key features in the initial layers will be easier for the network. So, we have resized our WCE images to a size of 100×100 pixels for a scaled-down CNN. Stationary Discrete Wavelet Transform tool in MATLAB (SWT) is used as a de-noising algorithm to smoothen the images, remove noise/artifacts and undesired distortions present in the images. Processed images are then fed to the convolutional neural network (CNN). The complexity of the model depends upon the complexity of the data. We can start by adding only one hidden layer in the network with neurons and then check the quality of trained network using cross-validation. Subsequently, deepen the network by adding more layers and neurons until the validation becomes stable. CNN works by automatically extracting features from the image using the training set of WCE images. CNN parameters and options are tuned to obtain better performance on the empirical trail basis. The methods for tuning the options of the CNN is discussed later in this section in detail. The trained network is applied to the test set of images for classification for comparative analysis. The accuracy, sensitivity, and specificity of the network are measured from confusion matrix and precision-recall graph curve is also plotted. Performance of the proposed model is compared with traditional machine learning methods such as Linear classifier, Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Random Forest (RF) algorithms.

Dataset preparation The WCE dataset used in this research is collected from Pushpawati Singhania Research Institute, (PSRI) Delhi, India of gastroenterology through a known gastroenterologist. A set of 2,621 WCE images is extracted from the video using VLC software at a rate of 2 frames per second. This dataset comprises of 505 bleeding frames and 2,116 normal frames. For video to image sampling using the VLC software, one can set the properties like image dimensions, bit depth, frame extraction rate, etc. in the preferences tab of the VLC software. The WCE images have color homogeneity problem as the color to be detected has a wide range of shades. The blood shade may fluctuate extensively from bright red to deep red, brownish, also containing redness of the normal skin tissues.

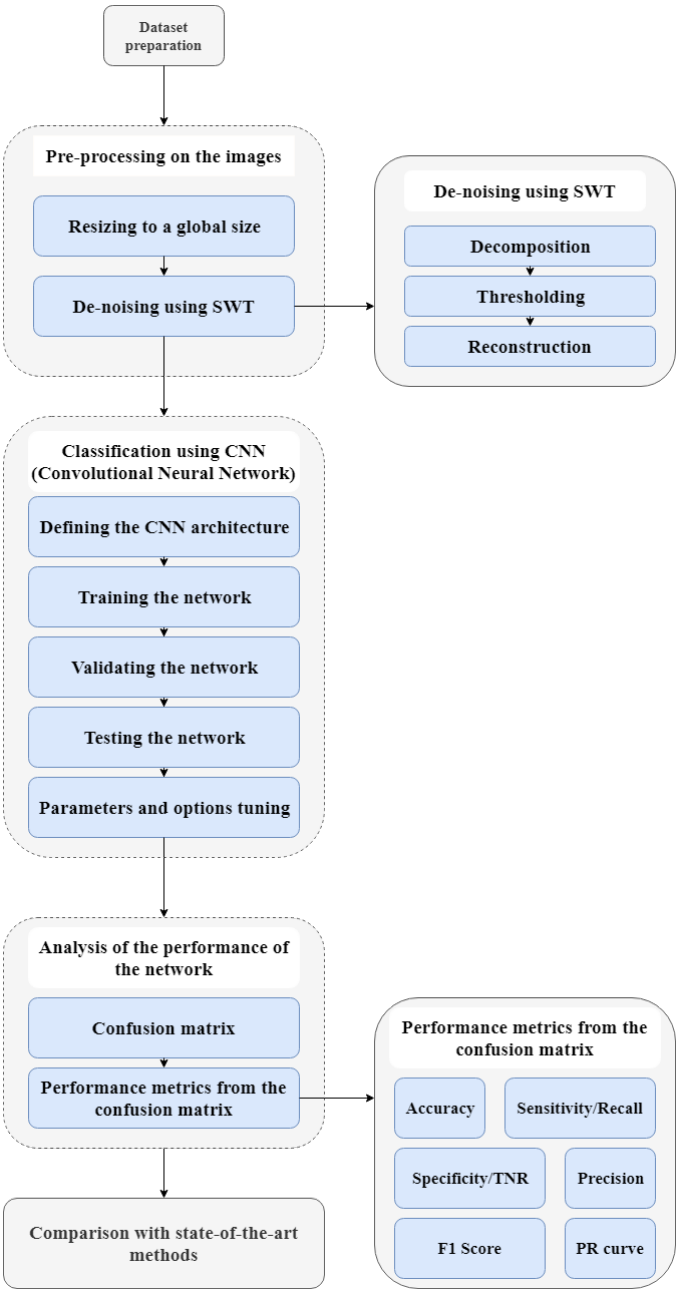

Fig. 5. Pipeline of the proposed technique

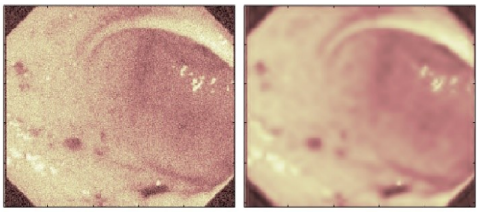

Fig. 6. (a) Input image (b) De-noised image

Pre-processing After acquiring the WCE image dataset, image preprocessing steps are done to enhance the performance of the network model to which the processed images are being fed. Pre-processing steps of image processing techniques are applied to the acquired image dataset for getting enhanced images to contrast the binary classification. The steps for preprocessing are as follows.

The first step of the pre-processing is to resize all the images to the a specific size for example 100×100. Deep learning network need consistent sizes and optimal size images for efficient and faster performance. High-resolution images involve more computations and higher memory specifications. If the input is a scaled down variant of the bigger images, then determining key features in the initial layers will be easier for the network [15].

De-noising using SWT We have implemented stationary wavelet transform (SWT) for noise removal and examine the statistically non-predictable signals, particularly at the region of discontinuities [45]. Since images have discontinuities at the edges, they can be presented spatially in a multi-resolution manner by using the SWT technique [8]. SWT is a wavelet transform that is used to transform signals or images to derive valuable information for the analysis as well as saves the computational cost and reduce required memory space. As we are working with medical images, the loss in information can adversely affect the result. SWT de-noising requires decomposition, level thresholding, and inverse transforming for reconstructing the image. For the step of decomposition, SWT algorithm decomposes an image into coefficients to get details about the image like contrast, correlation, energy, homogeneity, entropy, etc. It is done by a choosing a specific wavelet for image decomposition such as Haar, Daubechies, SymN, etc.

We have chosen Daubechies for our work by experimental analysi and also suggested by [11] for the image decomposition and de-noising. Daubechies wavelet being an orthogonal wavelet does not unnecessarily color the white noise, preserves the energy and relatively offer longer support [46]. The SWT decomposition of the image results into four sub-images to get the coefficients. These sub-images are obtained by applying vertical and horizontal filters that are low-pass filter (LPF) and high-pass filter (HPF). The resultant four sub-images of varying contrast, orientations, sharpness, and resolutions are called as approximations (average components) and detail components (horizontal, vertical and diagonal). After de-noising the images with these threshold limits, we apply an inverse transform to reconstruct the original image from the sub-images but without noise. The reconstruction is the reversed course of decomposition and known as inverse wavelet

transform. Figure 6 shows the smoothing effect by the SWT denoising which is easier for the CNN to perform binary classification.

Classification using CNN Classification of frames into bleeding and non-bleeding category is done by employing a convolutional neural network (CNN). De-noised images are fed into ConvNet (CNN) for detection of the bleeding in the processed set of images.

1. **Defining the CNN architecture:** ConvNet takes in images as three-dimensional objects. Neurons in the layers of CNN are organized in 3 dimensions: width, height, depth. Depth is same as number of color channels in the input image. Determining the optimal number of hidden layers and neurons that are incorporated into the network is a crucial part. Underfitting arises as a result of using too few neurons in the layers to sufficiently recognize the signals in a complex data set, whereas using too many neurons can give rise to time complexity and overfitting issues. Overfitting happens when the network has excessive data processing capability such that the confined amount of data comprised in the training set is not sufficient to train all of the neurons in the hidden layers.
   Another issue that can arise is the increased time taken to train the network with an overly huge number of neurons in the layers. Some adjustment needs to be made between an abundant and inadequate number of neurons in the hidden layers. We can start by adding only one hidden layer in the network with neurons and then check validation accuracy of the network through cross-validation. To optimize the network, we deepen the network gradually so that it can deal with more complex data and avoid underfitting. According to Heaton research in [1], more than 2 hidden layers in the network generally result in enabling the model to learn complex representations and extract features. We can try to optimize the performance by adding more neurons in the existing hidden layers or adding new hidden layers. Heaton [1] suggests that we can use a thumb rule instead of hit-and-trial which can be time-consuming and laborious. For defining a satisfactory number of neurons in the hidden layers, following steps should be used:
   - The number of neurons should fall within the range of the size of the input layer and the output layer.
   - It shall be 2/3 the size of the input layer, as well as the size of the output layer.
   - It needs to be less than twice the size of the input layer.

   The very first layer in the network is an image input layer that receives the denoised images scaled to $100 \times 100 \times 3$, for three color channels. We have incorporated 3 Conv layers, 2 pooling layers and, 1 fully connected layer. Due to small input image patch, 3 layers of convolution is enough, two pooling layers to be in between convolutional layers and one fully connected layer for binary classification. A batch normalization layer is always introduced after every Conv layer, followed by a ReLU layer to maintain the non-linearity of the image. Non-linear features of an image are the changes and shifts in pixels, the edges, borders, various colors, etc. Linear images do not appear normal to the human eye as they lack brightness and above-mentioned features. Conventional layers are linear in nature to learn the concept hierarchy and various non-linear activation functions have been incorporated in CNN for efficient feature extraction. A combination of linear inputs cannot generate a non-linear output, so without a non-linear activation function, the network will act like a single-layer

perceptron to accumulate all the layers of the network and follow a the standard of linear function [57]. As the second layer, a 2-D convolutional layer is defined with eight 3×3 convolutions with stride 1 and the same padding as input images. Pooling layer is applied with the arguments of 2×2 max pooling, stride of 2 and same padding. One fully connected layer for the bleeding and non-bleeding classes is defined and followed by a soft-max and classification layer to apply soft-max function and evaluate the cross-entropy loss.

2. **Training the network:** Dataset is split into training, validation and testing sets with standard proportion of 70%, 15%, and 15% respectively. So, CNN is trained on a random 70% training set split while regulating the weights on the network for less training error until the validation criteria is met.
3. **Validating the network:** Validation set is used to minimize over-fitting in the network and to ensure that any increment in accuracy over the training dataset results into an increment in accuracy over a test dataset that has not been yet exposed to the network or the network has not been trained on it such as validation dataset.
4. **Testing the network:** The trained network is then run on the test dataset of images for the classification of the bleeding and non-bleeding frames. This will yield the performance comparison and prediction robustness of the network.
5. **Parameters and options tuning:** For better performance of the network, parameter and options are tuned to optimal values for the underlying data distribution. Stochastic gradient descent with momentum (SGDM) is used as a solver with mini-batches of sizes 10. The total number of training samples in a batch is the batch size. Too small batch size results into gradient descent not being smooth, slow learning of the model and error may oscillate too much, whereas too high batch size results into the longer time required to do one training iteration with relatively small results [51]. SGDM is one of the best optimization algorithms as it helps in preventing oscillations. The number of epochs is the number of passes through the whole training set while training the network. Using a large number of epochs can result in over-fitting of the network and using a very small number of epochs result in an under-fit network [38]. Early stopping process enables us to use a large number of epochs but stops the network training as soon as a validation criterion is met. A smaller learning rate decreases the speed of the learning in the network, but it enables the network to converge smoothly. First, we chose a small learning rate varying between 1e-5 to 1 and then we check the performance of our network. To improve performance, a smaller learning rate is used [24]. The maximum number of epochs used in the proposed network is 6 with a learning rate of 1e-5 for better training after checking the performance of the network with different hit-and-trial values for these parameters.

3.2. Proposed Explanation Approach

Currently, in medical domain, XAI functionality is a necessary requirement for many machine learning-based medical research, education and clinical decision making scenarios. Systems for solving the medical domain explanation problem can be distinguished into two types; post-hoc systems and ante-hoc systems. Post-hoc systems help in providing local explanations for a particular decision made by machine learning so that it can be made interpretable on demand rather than explaining the whole systems behavior. One of

the algorithms that enables post-hoc explainability is LIME. Local interpretable model-agnostic explanations (LIME) [3] is the original Python implementation of this explanation technique. LIME takes two inputs: the neural network as generated by TensorFlow and the result of a specific frame to generate a matrix representation of the regions that triggered the corresponding classification. Ante-hoc systems are interpretable by design and referred to as *glass-box* approaches in the literature [35]; examples are decision trees, linear regression and fuzzy inference systems. In an applied science context, LIME has already been used for explaining machine learning models for the heat failure detection in air handling units [43]. Base idea of the explanation process has been published by Avleen et al. [44] and the proposed work is an extension of the previous work.

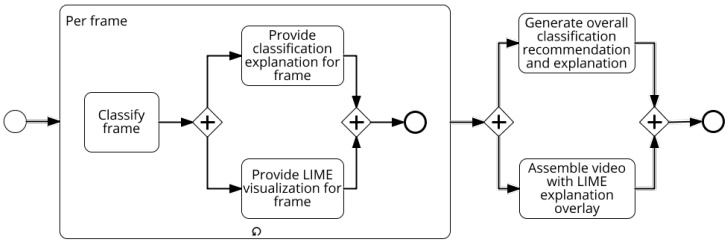

Fig. 7. Classification and Explanation Process

The classification and explanation process has been depicted in Figure 7, which describes the classification procedure by machine learning model as well as explanation and visualization by LIME for each image frame. The image data set is trained with a machine learning model (CNN in our example). The trained model is given to our proposed XAI model for providing classifications and explanations for these binary sick and healthy image classes. The overall explanations for the whole test data can be provided to the medical professionals for assisting them in decision making. The overall recommendation and explanation is provided by the health-care professional by making an aggregate ranking system for providing the severity of the intestinal bleeding in the patient case. The architecture of the proposed model is depicted in Figure 8 where the whole process can be divided into four segments: Pre-processing, applying the CNN model, explanation-generation using LIME, and decision-assistance for healthcare professionals.

4. Performance Metrics

Various performance metrics that have been used for the evaluation of the efficiency of the proposed system are as follows.

4.1. Confusion Matrix

The exactness and reliability of a system are calculated through a confusion matrix which is also termed as an error-matrix. This matrix supports in getting a clear idea of the performance of the system by analyzing the mis-classification rate and accuracy.

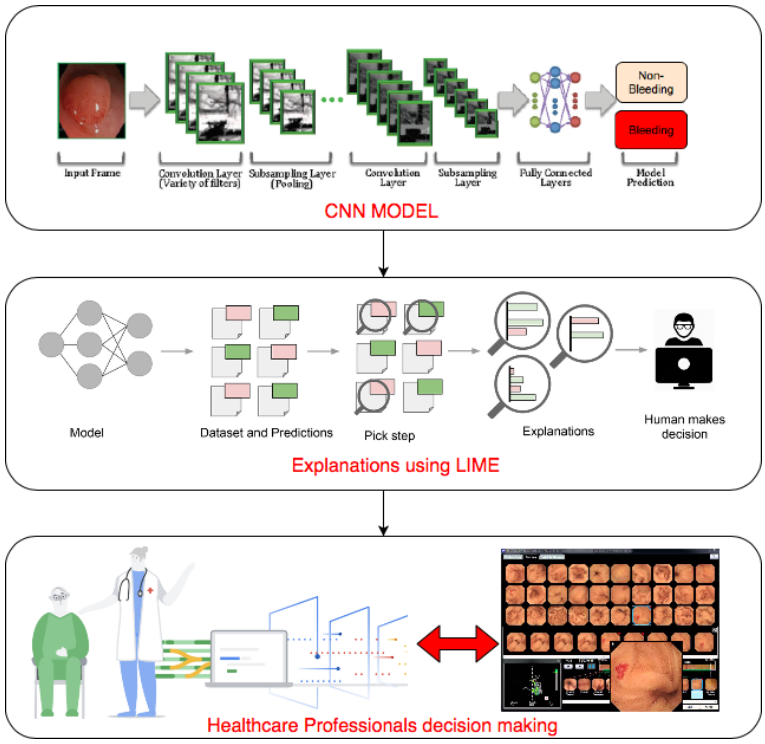

Fig. 8. The Architecture for the proposed model

1. True Positive (TP) is when the predicted and actual classes are identical tp true class. For instance, a frame that has bleeding present in it is getting predicted in the bleeding class.
2. True Negative (TN) is when the predicted and actual element is negative. For instance, a normal image without any bleeding getting predicted in normal class.
3. False Positive (FP): When the system predicts an element to be in true class but in actual it does not. For example a non-bleeding frame getting predicted in the bleeding class.
4. False Negative (FN): When the system predicts that an element does not belong to a false class but in actual it does. For example an actual bleeding frame is predicted as normal by the model.

In our research, the WCE image dataset is slightly unbalanced as it comprises of smaller ratio of bleeding frames as compared to normal frames. Accuracy alone is not considered a good evaluation metric in cases of unbalanced data classification. Both false negatives (FN) and false positives (FP) are important in medical image classification. In the case of endoscopy images, the cost of false negatives is as important as the false positives [19]. The damage of a bleeding frame to not get detected is worse than the damage of detecting a normal frame as bleeding as all the frames predicted as bleeding will be observed by the physician for the final judgment but the frames predicted as normal frames will probably

be overlooked. Nevertheless we do not want a large number of false positives in our prediction as it will reduce the efficiency of the network.

4.2. Accuracy

Accuracy refers to the total number of correct classifications done by the network out of the total number of examples. As shown in equation 6, accuracy is the rate of true predictions by all the true and false predictions combined.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{6}$$

4.3. Sensitivity or Recall

Sensitivity is the rate of accurately predicted positives to genuine positives. Recall gives us an idea about a model's performance proportionate to false negatives. As shown in equation 7, recall focuses on catching all the frames that have "bleeding" with the prediction as "bleeding", not just concerning catching frames correctly. For medical image classification problems, high sensitivity is preferred as it indicated high true positive value and low number of false negatives.

$$Specificity = \frac{TP}{(TP + FN)} \tag{7}$$

4.4. Specificity

Specificity is the rate of accurately predicted negatives to the actual negatives. Specificity is the exact reverse of sensitivity. Equation 8 shows the specificity derived from a confusion matrix.

$$Specificity = \frac{TN}{(TN + FP)} \tag{8}$$

4.5. Precision

Precision is the rate of accurately predicted positives to all the predicted positives. In equation 9, precision shows the proportion of the frames that are detected as having the presence of bleeding, actually had bleeding. Recall provides us an idea about a network's performance concerning false negatives, the frames that the network missed. Precision provides us with the idea of its performance concerning false positives for the frames that were predicted. Precision is about predicting frames correctly, whereas Recall is about prediction all the positive frames correctly. So, for minimizing false negatives, we have to focus on getting Recall as best as possible with a decent and acceptable Precision value. The values of both Precision and Recall can be monitored by a single value performance metric called as F1 score.

$$Precision = \frac{TP}{(TP + FP)} \tag{9}$$

4.6. F1 Score

To consider the role of both precision and recall, the F1 score is computed as in 10 which is simply the harmonic mean of precision and recall. In the case of unbalanced class distribution in the dataset, F1 score is a better evaluation metric than accuracy. Low value of F1 score indicates a problem when one of the Precision and Recall has a low value. In that case, F1 score is closer to the smaller value than the bigger value out of these two.

$$F1\ Score = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \tag{10}$$

4.7. Precision-Recall curve

To demonstrate the trade-off in precision and recall, PR curve gives a more informational depiction of the performance of the network with unbalanced dataset [19], [16]. The area under the PR curve varies from 0 to 1 and also gives an idea about the network's performance. If AUC is close to 1 then the model is considered as good. The closer the curve is to the top-right edge, the more reliable the system. Henceforth, a greater area under the curve (AUC) symbolizes that the system has higher precision and higher recall.

5. Result Evaluation

In this research, we have used MATLAB 2018a for implementing the proposed model. MATLAB offers a good data visualization and it also offers a large number of toolboxes/ apps for processing and ploting image dataset with ease of usage. We have preprocessed the WCE images and trained a convolutional neural network. The performance of proposed model is compared with traditional machine learning methods including linear discriminant model, SVM, ANN, Random Forest. The performances of the models are evaluated from metrics derived from confusion matrix like accuracy, specificity, sensitivity, precision and F1 score.

5.1. System and Data Configuration

ANT PC (10 Cores 20 Threads), Intel C612 Chipset Motherboard with single Socket, 32GB ECC RAM 2400Mhz, Dual Nvidia GeForce RTX 2080TI 11GB, Intel Server Heatsink, 250GB Samsung 860 Evo SATA SSD, 2TB Western Digital HDD, 1000W 80+ Gold Power Supply, Ubuntu operating system. The WCE dataset of $2,621$ images are collected from Pushpawati Singhania Research Institute (PSRI) institute of gastroenterology in Delhi India. It comprises of 505 bleeding frames and 2,116 normal frames.

Various textures, color and contrast types in the image dataset are shown in the Figure 9. These sample images render the frames with both malignant and benign features like bleeding, polyp, wrinkles and contractions. The images with turbid present in the system and images of walls of digestive system are also depicted in the figure 9. In our research, we focused only on extracting the frames that have bleeding present in them from the entire dataset.

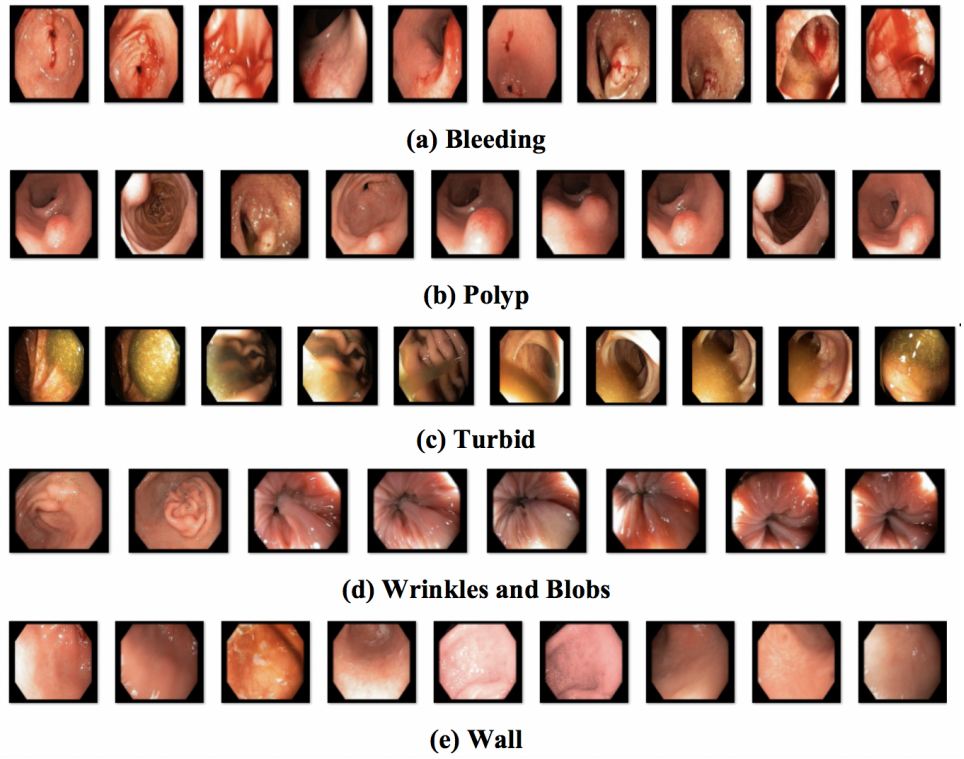

Fig. 9. Various ailments in the dataset. (a) Bleeding (b) Polyp (c) Turbid (d) Wrinkles and blobs (e) Membrane wall

5.2. Classification Results

Confusion matrix of the proposed system is plotted in Table 2 to calculate the evaluation metrics. Metrics like accuracy, sensitivity (recall), specificity, precision, F1 score are evaluated using the values of TP, FP, TN, FN using confusion matrix entries in Table 2. In Table 3, the accuracy ratios have been calculated using confusion matrix (Table 2) of the binary classification has been shown. It is observed that the false negatives are 8 and false positives are 24. The test accuracy obtained is 91.9%. The values taken from confusion plot are the TP = 52, TN = 309, FP = 8, FN = 24, sensitivity (recall or TPR) = 68.5%, specificity (TNR) = 97.5%, precision = 86.7%. Precision and recall are used to calculate F1 score. As we can see in Table 4, test accuracy is better than validation accuracy, so it can be concluded that network is not a result of over-fitting. Recall has slightly lower value and precision is higher. F1 score falls in between both the recall and precision. The recall is affected due to the high number of false negatives.

The precision-recall curve is plotted using the results from the proposed model and shown in Figure 10, the curve depicts the trade-off between precision and recall of the network. The area under the PR curve (AUC-PR) is calculated to be 0.82. The value of

Table 2. Confusion matrix of test dataset which is 15% of (2,116+500) = 393 images. Positive means bleeding and negative means normal images.

		Predicted		
		Positive	Negative	Total
Actual	Positive	52	8	60
	Negative	24	309	333
	Total	76	317	393

Table 3. Performance metric ratios such as sensitivity, specificity, precision & accuracy using confusion matrix in Table 2 for binary classification

		Predicted:Yes	*Predicted:No*	**Overall**
Bleeding	*Actual:Yes*	13.2% (52)	2% (8)	86.7% (Precision)
Normal	*Actual:No*	6.1% (24)	78.6% (309)	7.2%
Overall		68.4% (Sensitivity)	97.5% (Specificity)	91.9 % (Accuracy)

AUC-PR is affected by the low recall value in our experiments. PR Curve depicts the trade-off between precision and recall values of the model.

5.3. Comparative Analysis

In Table 5, we have compared the performance of our model with traditional machine learning models based on standard evaluation metrics. Due to the fact that WCE image dataset being imbalanced in nature, the performance evaluation of the models cannot rely on just accuracy and therefore, advanced metrics are calculated [39]. The F1 score of SVM and Random forest is close to the proposed model. But, these two models require manual feature extraction, segmentation, thresholding, etc. Whereas, the proposed model does not rely on handcrafted features. So, it reduces the human intervention of surveying the feature extraction and ML techniques followed by orchestrating various algorithms together. Hence, the proposed CNN system has outperformed to detect the bleeding frames in the WCE images. The performance of the proposed system on WCE images is slightly better (with F1 score of 0.76) than the other traditional machine learning algorithms shown in compared in Table 5. The performance of a classifier is not always evaluated just by the accuracy but other important metrics like precision, recall, F1 score and ease of implementation must also considered to analyse the efficiency and reproducibility of the model.

Table 4. Performance evaluation of the CNN mode on WCE dataset

Sr.	Metric	Value
1	Validation Accuracy	90.84%
2	Test Accuracy	91.92%
3	Sensitivity	68.42%
4	Specificity	97.48%
5	Precision	86.67%
6	F1 Score	0.7647

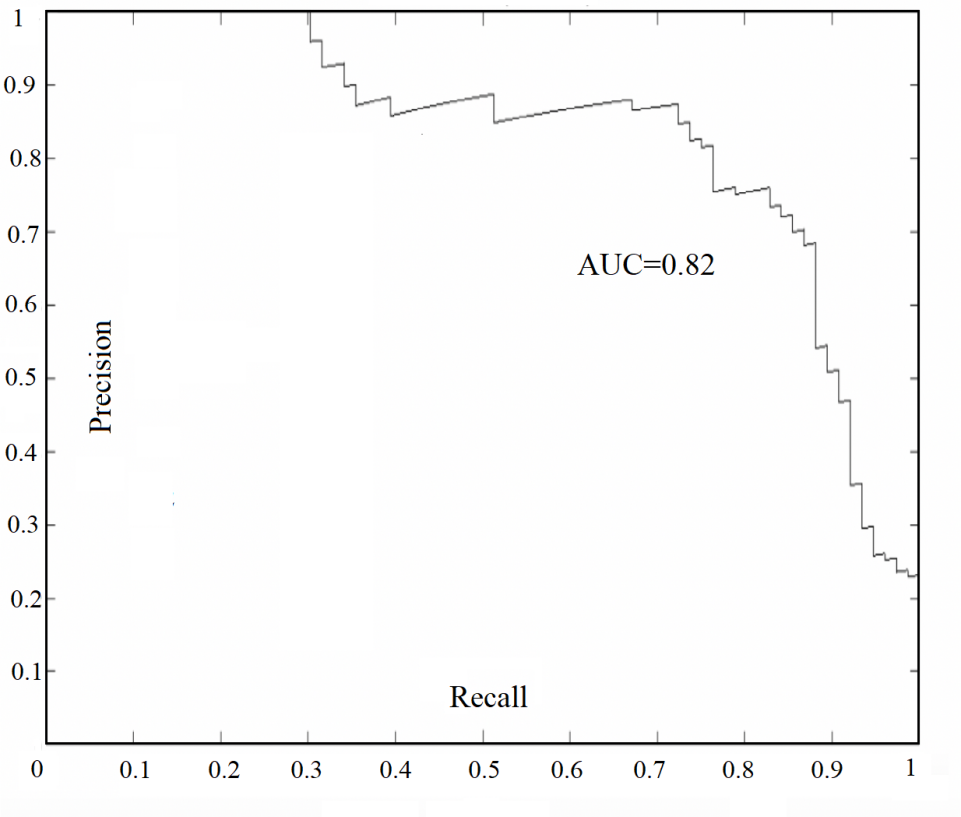

Fig. 10. Precision-Recall curve of CNN model

Table 5. Comparative analysis of the proposed model with state-of-the-art techniques

Sr.	ML Model	Accuracy	Sens.	Spec.	Precision	F1 Score
1	CNN	91.92%	68.42%	97.48%	86.67%	0.76
2	Logistic Regression Model	89.72%	82.27%	90.84%	58.92%	0.68
3	SVM	91.01%	92.19%	91.86%	63.29%	0.75
4	ANN	89.43%	56.34%	97.21%	82.79%	0.67
5	Random Forest	91.11%	87.95%	91.61%	62.30%	0.73

Moreover, the computational burden and human intervention is also decreased in the proposed model as there is no requirement for manual feature extraction as it is in other above stated models. The manual feature extraction by using color histogram and co-occurrence matrix increases the computational steps and hassle for the model application. Moreover it also requires a ground level knowledge of image processing and basic machine learning model. Whereas, using CNN network is trending nowadays for its ease of direct image input for learning and its working strategy is similar to humans way of learning.

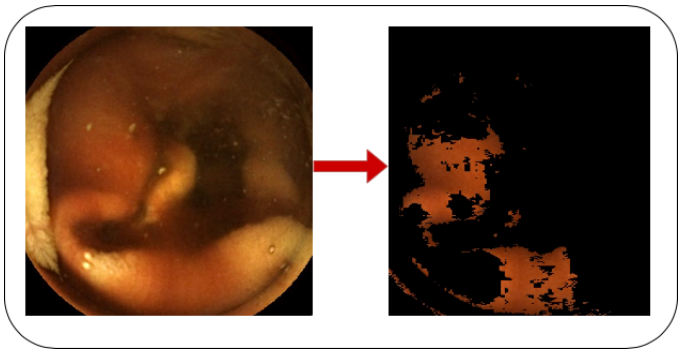

Fig. 11. Annotating the red lesions in capsule endoscopy images

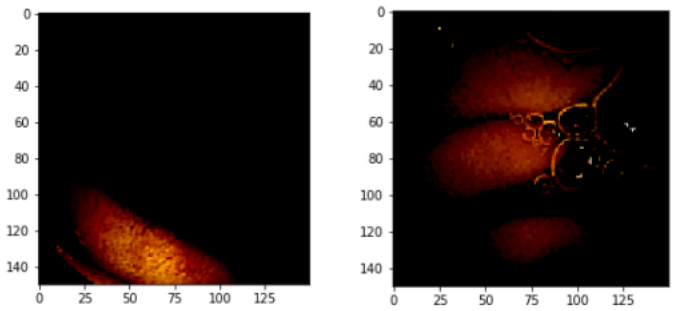

Fig. 12. The bloody regions are shown to give the glimpse of the areas due to which the image is classified as bloody image

5.4. Explanation Results with LIME

LIME provides explanations for bleeding images by drawing the boundaries over the bloody areas in the tested image known as annotations as shown in the Figures 11, 12 and 13. The decision made by black box machine learning model gets justified using LIME XAI model for further analysis by the area experts and adds on the model reliability. Bleeding region is highlighted for features or areas due to which the image is classified as bleeding case. LIME has been tested for all the bleeding images in the validation and test dataset similar to Figure 13. Thus the proposed CNN model is easy to use, efficient and transparent through model agnostic XAI technique for complex medical image applications. Nonetheless, the proposed technique is reproducible and scalable for any image classification application.

6. Conclusion

The proposed CNN model classifies and annotates the bleeding frames in wireless capsule endoscopy (WCE) video dataset. A real time video has been obtained from a known gastroenterologist. Images are sampled from the WCE video using VLC software and

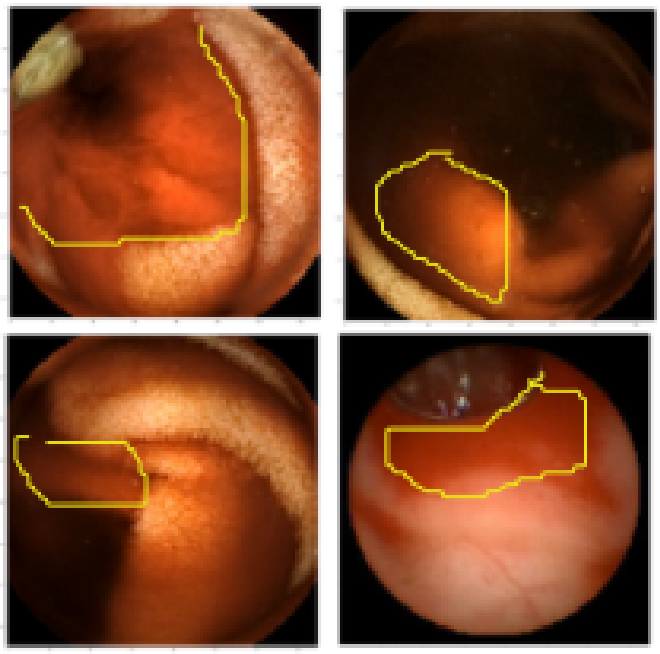

Fig. 13. LIME explanations provided in the form of boundaries for bloody regions

pre-processed to a get standard sized, de-noised images for efficient machine learning model. A convolutional neural network (CNN) is designed and proposed for the classification of WCE frames into the bleeding and non-bleeding categories. The performance of the proposed model is compared with other traditional machine learning models on the basis of evaluation parameters. We have proposed and prototyped an explainable machine learning tool that should be used by medical experts as a decision-support system to detect gastroenterological bleeding faster and in a more reliable manner. The *assumption* for using proposed model is the availability of a GPU system since CNN classification is computationally complex. Traditional machine learning models with handcrafted features could be faster if GPU system is not available.

There is still room for improvement by exploring other deep learning models and variants of explainable artificial intelligence (XAI). Image moments should also be analysed since they yield robust image features which are rotation, scaling and traslation invariants. Other modalities should also be incorporated for a multi-modal machine learning such as free expert text available with the images.

Acknowledgments. This research work has been sponsored by the Seed Money project grant at Thapar Institute of Engineering and Technology Patiala India under Grant TU/DORSP.

References

1. Deep learning project. https://jhui.github.io/2018/02/11/

How-to-start-a-deep-learning-project/ (2018), [Online; accessed 04-June-2019]
2. ELI5. https://github.com/TeamHG-Memex/eli5 (2019), [Online; accessed 04-June-2019]
3. LIME. https://towardsdatascience.com/ (2019), [Online; accessed 04-June-2019]
4. shap. https://github.com/slundberg/shap (2019), [Online; accessed 04-June-2019]
5. Skater. https://github.com/oracle/Skater (2019), [Online; accessed 04-June-2019]
6. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access 6, 52138–52160 (2018)
7. Alaskar, H., Hussain, A., Al-Aseem, N., Liatsis, P., Al-Jumeily, D.: Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images. Sensors 19(6), 1265 (2019)
8. Alwan, I.M.: Color image denoising using stationary wavelet transform and adaptive wiener filter. Al-Khwarizmi Engineering Journal 8(1), 18–26 (2012)
9. Anjomshoae, S., Främling, K., Najjar, A.: Explanations of black-box model predictions by contextual importance and utility
10. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems (2019)
11. Anutam, R.: Performance analysis of image denoising with wavelet thresholding methods for different levels of decomposition. International Journal of & its Applications 6(3), 35–46 (2014)
12. Aoki, T., Yamada, A., Aoyama, K., Saito, H., Tsuboi, A., Nakada, A., Niikura, R., Fujishiro, M., Oka, S., Ishihara, S., et al.: Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. Gastrointestinal endoscopy 89(2), 357–363 (2019)
13. Bchir, O., Ismail, M.M.B., AlZahrani, N.: Multiple bleeding detection in wireless capsule endoscopy. Signal, Image and Video Processing 13(1), 121–126 (2019)
14. Billah, M., Waheed, S.: Gastrointestinal polyp detection in endoscopic images using an improved feature extraction method. Biomedical engineering letters 8(1), 69–75 (2018)
15. Bitenc, M., Kieffer, D., Khoshelham, K.: Evaluation of wavelet denoising methods for small-scale joint roughness estimation using terrestrial laser scanning. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences 2 (2015)
16. Boyd, K., Eng, K.H., Page, C.D.: Area under the precision-recall curve: point estimates and confidence intervals. In: Joint European conference on machine learning and knowledge discovery in databases. pp. 451–466. Springer (2013)
17. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 68(6), 394–424 (2018)
18. Chitra, S., Ashok, L., Anand, L., Srinivasan, V., Jayanthi, V.: Risk factors for esophageal cancer in coimbatore, southern india: a hospital-based case-control study. Indian journal of gastroenterology 23(1), 19–21 (2004)
19. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006)
20. Deeba, F., Islam, M., Bui, F.M., Wahid, K.A.: Performance assessment of a bleeding detection algorithm for endoscopic video based on classifier fusion method and exhaustive feature selection. Biomedical Signal Processing and Control 40, 415–424 (2018)

21. Delvaux, M., Gay, G.: Capsule endoscopy: technique and indications. Best Practice & Research Clinical Gastroenterology 22(5), 813–837 (2008)
22. D'Halluin, P.N., Delvaux, M., Lapalus, M.G., Sacher-Huvelin, S., Soussan, E.B., Heyries, L., Filoche, B., Saurin, J.C., Gay, G., Heresbach, D.: Does the "suspected blood indicator" improve the detection of bleeding lesions by capsule endoscopy? Gastrointestinal endoscopy 61(2), 243–249 (2005)
23. Diamantis, D.E., Iakovidis, D.K., Koulaouzidis, A.: Look-behind fully convolutional neural network for computer-aided endoscopy. Biomedical Signal Processing and Control 49, 192–201 (2019)
24. Drakos, G.: How to select the right evaluation metric for machine learning models: Part 1 regression metrics. Towards Data Science. Saatavissa: https://towardsdatascience. com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0. Hakupäivä 3, 2019 (2018)
25. Främling, K.: Explaining results of neural networks by contextual importance and utility. In: Proceedings of the AISB'96 conference. Citeseer (1996)
26. Främling, K.: Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère. Ph.D. thesis, INSA de Lyon (1996)
27. Främling, K., Graillot, D.: Extracting explanations from neural networks. In: Proceedings of the ICANN. vol. 95, pp. 163–168. Citeseer (1995)
28. Ghosh, T., Fattah, S.A., Wahid, K.A.: Chobs: Color histogram of block statistics for automatic bleeding detection in wireless capsule endoscopy video. IEEE journal of translational engineering in health and medicine 6, 1–12 (2018)
29. Ghosh, T., Fattah, S.A., Wahid, K.A., Zhu, W.P., Ahmad, M.O.: Cluster based statistical feature extraction method for automatic bleeding detection in wireless capsule endoscopy video. Computers in biology and medicine 94, 41–54 (2018)
30. Ghosh, T., Li, L., Chakareski, J.: Effective deep learning for semantic segmentation based bleeding zone detection in capsule endoscopy images. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 3034–3038. IEEE (2018)
31. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
32. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51(5), 93 (2018)
33. Hajabdollahi, M., Esfandiarpoor, R., Soroushmehr, S., Karimi, N., Samavi, S., Najarian, K.: Segmentation of bleeding regions in wireless capsule endoscopy images an approach for inside capsule video summarization. arXiv preprint arXiv:1802.07788 (2018)
34. He, J.Y., Wu, X., Jiang, Y.G., Peng, Q., Jain, R.: Hookworm detection in wireless capsule endoscopy images with deep learning. IEEE Transactions on Image Processing 27(5), 2379–2392 (2018)
35. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)
36. Iddan, G., Meron, G., Glukhovsky, A., Swain, P.: Wireless capsule endoscopy. Nature 405(6785), 417 (2000)
37. Jia, X., Meng, M.Q.H.: A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 639–642. IEEE (2016)
38. Jia, X., Meng, M.Q.H.: Gastrointestinal bleeding detection in wireless capsule endoscopy images using handcrafted and cnn features. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 3154–3157. IEEE (2017)
39. Kaur, H., Pannu, H.S., Malhi, A.K.: A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys (CSUR) 52(4), 1–36 (2019)

40. Lan, L., Ye, C., Wang, C., Zhou, S.: Deep convolutional neural networks for wce abnormality detection: Cnn architecture, region proposal and transfer learning. IEEE Access 7, 30017–30032 (2019)
41. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature 521(7553), 436–444 (2015)
42. Leenhardt, R., Vasseur, P., Li, C., Saurin, J.C., Rahmi, G., Cholet, F., Becq, A., Marteau, P., Histace, A., Dray, X., et al.: A neural network algorithm for detection of gi angiectasia during small-bowel capsule endoscopy. Gastrointestinal endoscopy 89(1), 189–194 (2019)
43. Madhikermi, M., Malhi, A., Främling, K.: Explainable artificial intelligence based heatrecycler fault detection in air handling unit (2019)
44. Malhi, A., Kampik, T., Pannu, H., Madhikermi, M., Främling, K.: Explaining machine learning-based classifications of in-vivo gastral images. In: 2019 Digital Image Computing: Techniques and Applications (DICTA). pp. 1–7. IEEE (2019)
45. Masumdar, R., Karandikar, R.: Comparative study of different wavelet transforms in fusion of multimodal medical images. International Journal of Computer Applications 146(11) (2016)
46. Mortazavi, S., Shahrtash, S.: Comparing denoising performance of dwt, wpt, swt and dt-cwt for partial discharge signals. In: 2008 43rd International Universities Power Engineering Conference. pp. 1–6. IEEE (2008)
47. Nawarathna, R., Oh, J., Muthukudage, J., Tavanapong, W., Wong, J., De Groen, P.C., Tang, S.J.: Abnormal image detection in endoscopy videos using a filter bank and local binary patterns. Neurocomputing 144, 70–91 (2014)
48. Obukhova, N., Motyko, A., Timofeev, B., Pozdeev, A.: Method of endoscopic images analysis for automatic bleeding detection and segmentation. In: 2019 24th Conference of Open Innovations Association (FRUCT). pp. 285–290. IEEE (2019)
49. Rawla, P., Barsouk, A.: Epidemiology of gastric cancer: global trends, risk factors and prevention. Przeglad gastroenterologiczny 14(1), 26 (2019)
50. Signorelli, C., Villa, F., Rondonotti, E., Abbiati, C., Beccari, G., de Franchis, R.: Sensitivity and specificity of the suspected blood identification system in video capsule enteroscopy. Endoscopy 37(12), 1170–1173 (2005)
51. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery 9(2), 283–293 (2014)
52. Sivakumar, P., Kumar, B.M.: A novel method to detect bleeding frame and region in wireless capsule endoscopy video. Cluster Computing pp. 1–7 (2018)
53. Tuba, E., Tomic, S., Beko, M., Zivkovic, D., Tuba, M.: Bleeding detection in wireless capsule endoscopy images using texture and color features. In: 2018 26th Telecommunications Forum (TELFOR). pp. 1–4. IEEE (2018)
54. Vieira, P.M., Silva, C.P., Costa, D., Vaz, I.F., Rolanda, C., Lima, C.S.: Automatic segmentation and detection of small bowel angioectasias in wce images. Annals of biomedical engineering 47(6), 1446–1462 (2019)
55. Westerhof, J., Koornstra, J.J., Weersma, R.K.: Can we reduce capsule endoscopy reading times? Gastrointestinal endoscopy 69(3), 497–502 (2009)
56. Xing, X., Jia, X., Meng, M.H.: Bleeding detection in wireless capsule endoscopy image video using superpixel-color histogram and a subspace knn classifier. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 1–4. IEEE (2018)
57. Yu, F.: A comprehensive guide to fine-tuning deep learning models in keras
58. Yuan, Y., Meng, M.Q.H.: Deep learning for polyp recognition in wireless capsule endoscopy images. Medical physics 44(4), 1379–1389 (2017)

Apoorva Singh got her Master's degree in Computer Science & Engineering from Thapar Institute of Engineering and Technology Patiala India. She did her Bachelor's of En-

gineering degree in Information Technology from Quantum school of technology India. Her research interests are machine learning and image processing.

Husanbir Singh Pannu is an assistant professor in Computer Science & Engineering Department at Thapar Institute India. His research areas are machine learning, text and image analysis. He got his PhD from University of North Texas USA and was postdoc fellow at Trinity College Dublin Ireland.

Avleen Malhi is a senior lecturer in data science and AI at Bournemouth University UK. She received her PhD in autonomous vehicles (2016), ME in Computer Science (2012) from Thapar Institute India and BE in Computer Science Engineering (2010). As part of her PhD research (2012-2016), she was working on an industrial project by Tata Consultancy Services. Between 2016-2018, she was working as an Assistant professor in computer science at Thapar University and from 2019-2020, she was working as postdoctoral researcher at Aalto University, Finland.

Received: October 3, 2020; Accepted: August 18, 2021.