Multimedia Tools and Applications (2024) 83:15825–15844
https://doi.org/10.1007/s11042-023-15347-4

Combining CNNs and 2-D visualization method for GI tract lesions classification

Shima Ayyoubi Nezhad[1] · **Toktam Khatibi**[1] · **Masoudreza Sohrabi**[2]

Received: 17 September 2021 / Revised: 29 March 2023 / Accepted: 15 April 2023 /
Published online: 24 July 2023

Abstract

In recent years, artificial intelligence and its tools are demonstrated enough potential for analyzing medical images. Several deep learning models have been proposed in previous studies for gastrointestinal (GI) tract like ulcers, polyps, bleeding, and other lesions. Hand-operated investigation of these lesions requires time, cost, and an expert physician. Automatic detection and classification of GI tract lesions are vital because misdiagnosis of them can affect the quality of human life. In our study, an effective model is proposed for a GI tract classification with the best performance. The proposed method's main aim is to classify GI tract lesions precisely from endoscopic video frames automatically. The different scenarios are designed, assessed, and compared by implementing 5-fold cross-validation on the KVASIR V1 dataset to achieve this aim. This dataset includes anatomical landmarks (pylorus, z-line, and cecum), pathological findings (esophagitis, ulcerative colitis, and polyp), and polyp removals (dyed lifted polyps, and dyed resection margins) as output classes. Each class includes 500 images, and an image's resolution varies from 750×576 to 1920×1072 pixels. These first and second scenarios are based on deep neural networks (DNNs). However, in the first scenario, a novel approach is proposed for visualizing 2-D data maps from features extracted from the convolutional auto-encoder (CAE). The last one is schemed based on pre-trained convolutional neural networks (CNNs). The experimental results illustrate the average accuracy of the first, second, and third scenarios is 99.87 ± 0.001, 92.07 ± 0.086, and 90.55 ± 0.111, respectively. The first scenario outperforms the compared ones with an average accuracy of 99.87 ± 0.001 and an AUC of 100.00 ± 0.000.

Keywords Medical image analysis · Endoscopy · Deep neural networks · Auto-encoders · Data visualization

✉ Toktam Khatibi
toktam.khatibi@modares.ac.ir

Shima Ayyoubi Nezhad
shima.ayyoubinezhad@modares.ac.ir

Masoudreza Sohrabi
sohrab_r@yahoo.com

[1] School of Industrial and Systems Engineering, Tarbiat Modares University (TMU), Tehran 14117-13114, Iran

[2] Gastrointestinal and Liver Diseases Research Center, Iran University of Medical Sciences (IUMS), Tehran, Iran

1 Introduction

Some illnesses can affect the gastrointestinal (GI) tract, according to GLOBOCAN 2020 estimates of cancer incidence and mortality produced by the International Agency for Research on Cancer [35]. Esophageal, stomach, and colorectal cancer are the most common cancers worldwide [35]. In 2020, 1.9 million new cases and 935,000 deaths from colorectal cancer were estimated, which is the third in incidence and the second in mortality [35]. Stomach cancer, with 1.1 million new cases and 769,000 deaths, has the fifth for incidence and the fourth for mortality, grades globally [35]. Esophageal cancer with 604,000 new cases is the seventh in incidence and with 504,000 deaths is the sixth in mortality in total [35].

With the advent of the minimally invasive surgeries (MIS), like endoscopy for examination of the upper GI tract and colonoscopy for screening the lower GI tract, physicians use MIS techniques for lesions, ulcers, polyps, and other abnormal finding and removal [28]. Failing that diagnoses the mentioned abnormalities during GI screening can lead to growing the diseases such as GI malignancies in patients in the next few years [12]. To reduce the rate of misdiagnosis, previous studies have focused on the automatic identification, classification and localization of the abnormalities used for automated analysis of medical images [8, 11, 14, 20, 30, 32].

Some previous studies have designed and proposed the automatic detection of polyps [32], tumors [36], cancer [19], erosion and ulcer [2], and bleeding [4, 16, 27].

Various previous studies have designed and proposed methods based on conventional machine learning models [19, 23, 36] and deep neural networks (DNNs) [4, 32] as the newer branch of machine learning models [5]. Previous studies have shown that DNNs can extract, analyze, and learn the valuable features from the raw dataset automatically [5, 11, 20, 30]. The principal prerequisite of using and training the conventional machine learning models is extracting handcrafted features [5]. Some previous studies combines handcrafted features and DNNs features to enhance the performance [10, 16].

The different architectures of DNNs are used for the automatic identification GI tract abnormalities in the previous study, like convolutional neural network (CNN) [7, 37], auto-encoders (AE) [13], regionbased convolutional neural networks (R-CNN) [39], generative adversarial networks (GANs) [31].

In [7], the researchers have proposed a CNN to classify bleeding capsule endoscopic video frames from non-bleeding ones. They have used pre-trained AlexNet then trained its last dense layers and also, SegNet which is used for semantic segmentation of the bleeding zones [7].

Another previous study has proposed a method consisting of two sequential convolutional encoder-decoder to extract features from images and detect polyps automatically [13]. The novelty of their proposed approach has been using a hetero-associator (hetero-encoder) in front of the model, which generates labeled images with a specific similarity to the actual image [13].

A study has recommended a feature learning method named stacked sparse auto-encoder with image the manifold of image constraint (SSAEIM), to prepare discriminative explanation of polyps and recognize them into the images of Wireless capsule endoscopy (WCE) [38]. They have assumed that the images with identical classes similar features and the others should be different enough [38].

Our previous study has proposed a semi-supervised deep model for anatomical landmark detection from endoscopic video frames. Our semi-supervised convolutional neural network (SSCNN) has been functional when accessibility to the labeled video frames was difficult. We have examined our previously proposed method on the KVASIR V1 dataset [28].

Previous research is used the representational power of convolutional auto-encoders (CAE) networks for feature extraction. In [21], a CNN, with three convolutional layers, is

displaced with average pooling layers to extract smooth features from Optical Emission Spectroscopy (OES) data. In the other study, the researcher used CAE to learn audio features. This pipeline is for converting source lectures into target ones. This proposed method achieves high-level features that consist of an authentic representation of the audio file [6].

Feature extraction is a necessary step in many machine learning problems. The advent of automatic feature extraction methods has required the conversion and representation of sophisticated data into lower dimensional without losing any information. This approach in feature extraction is the basis of the inspiration and accomplishment of novel technologies in computer vision.

As demonstrated by previous studies, researchers have tended to use DNNs in recent years because of their abilities in various areas. A combination of DNNs and using their advantages of them together to find the lesions in endoscopic frames can be helpful.

Previous studies have shown that discriminating some classes from each other would be difficult, and some automatic models have not demonstrated desirable performance in distinguishing them [2]. Some previous researchers have addressed this issue by working on a new extended version of the dataset [3, 25] or focusing on specific lesions [15, 29].

The prime purpose of the proposed method is to develop a novel approach based on the combination of DNNs for classifying GI tract lesions from endoscopic video frames. The second is to provide a novel method that extracts high-level features from the endoscopic video frames and depicts them into a 2-D data map.

The main novelties of the proposed method can summarize as the following:

- Proposing a new approach combining DNNs to classify the GI tract lesions from endoscopic video frames.
- Utilizing the benefits of CAE and 2-D visualization together.
- Extracting high-level features with CAE and converting them into 2-D data maps.
- Training CNN with novel 2-D visualization data maps.

2 Materials and methods

In the proposed method, three different scenarios are designed, proposed and compared for classifying GI tract lesions from endoscopic video frames as shown by Fig. 1.

As illustrated by Fig. 1, the first proposed scenario is composed of encoding N patches of each image using CAE, visualizing the features into a 2-D data map and considering it as an image, and feeding 2-D data map image into CNNs as their input. Since the first proposed scenario combines using Convolutional Auto-Encoders, the patches extracted from each image feed into Convolutional Auto-encoder as input. 2-D visualization and CNNs, we named this scenario the CAE-2DV-CNN scenario. The second scenario consists of training the end-to-end CNN for classifying the KVASIR V1 dataset, so we named this CNN-Sc. The last scenario uses the Inception-V3 as transfer learning, so we called that Incept_TL.

More details of the main steps of each scenario have described in the following subsections.

2.1 Dataset description

Several studies have used the KVASIR V1 dataset for designing and examining their proposed methods [2, 28, 29]. We use this dataset. It consists of 4,000 images based on anatomical landmarks, pathological findings and polyp removal [28]. These images have been

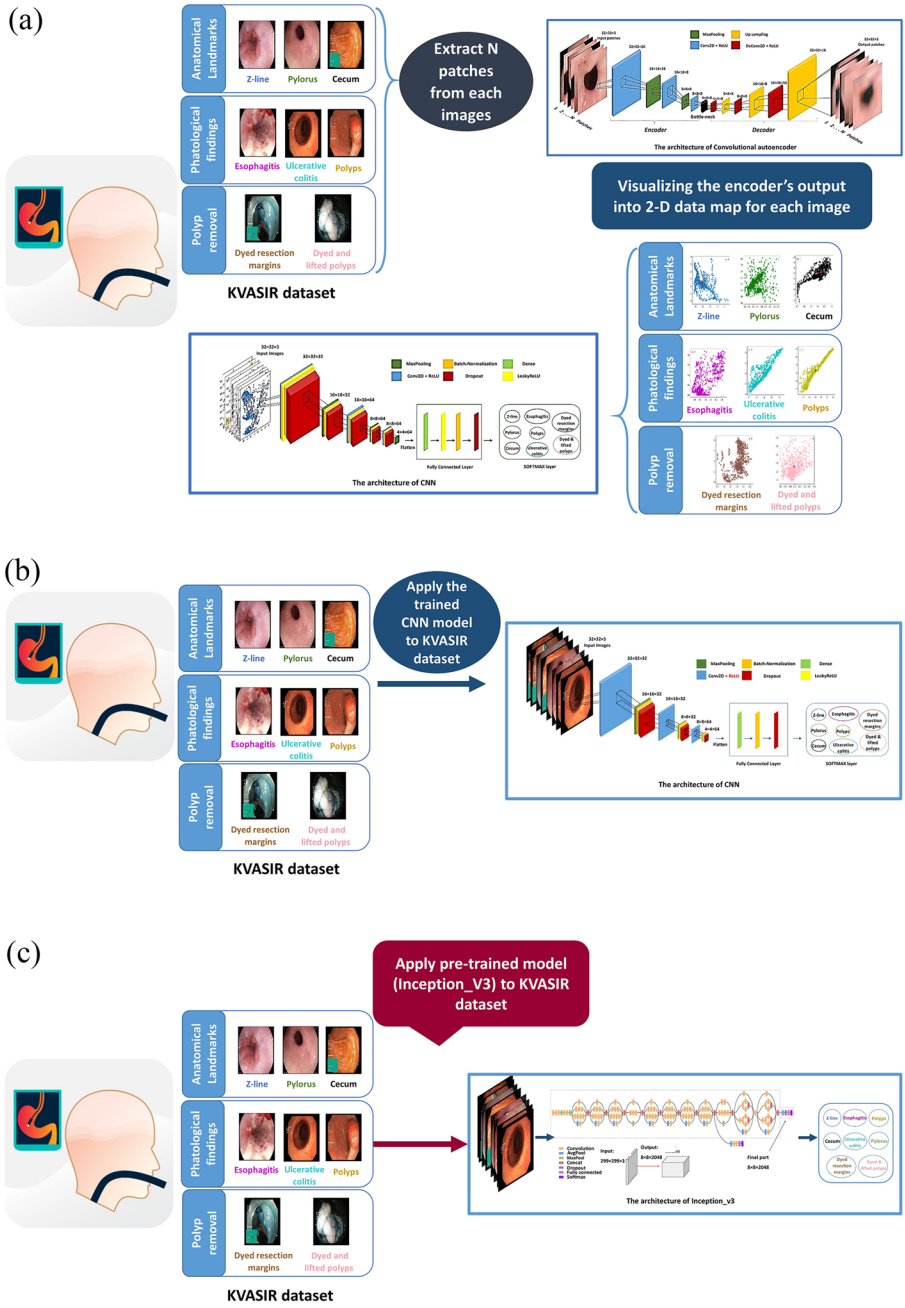

Fig. 1 Three designed and proposed different scenarios for classifying GI tract lesions from endoscopic video frames. **a** The first scenario for classifying GI tract lesions from endoscopic video frames (AE-2DV-CNN). **b** The second scenario for classifying GI tract lesions from endoscopic video frames (CNN-Sc). **c** The third scenario for classifying GI tract lesions from endoscopic video frames (Incept_TL)

listed into eight classes. Each class includes 500 images with different resolutions ranging from 720×576 to 1920×1072 pixels [28].

K-fold cross validation (C.V.) is a sampling technique that is portioned into k-equal subsets [34]. For each k, one of these k subsets is considered the test dataset, and the others are considered the training dataset. In all scenarios of the proposed method, five-fold cross validation is used for sampling from endoscopic video frames to build the training and test datasets. In each fold, 80%, or 3,200 images used for training, and 20%, or 800 images used for testing.

2.2 The first proposed scenario (CAE-2DV-CNN) for classifying GI tract lesions from endoscopic video frames

The first scenario, as shown in Fig. 1, includes image processing and patch extraction, encoding image by CAE, 2-D visualization of the feature map as a data map, training CNN classifier based on data map and evaluation and validation.

2.2.1 Image preprocessing and patch extraction

At first, images were resized into 64×64 pixels. Then N patches ($N=1089$) were extracted from each image with a size of 32×32 pixels.

2.2.2 Encoding image by convolutional auto-encoder (CAE)

An auto-encoder (AE) is a specific type of Artificial Neural Network (ANN) that aim is to regenerate the inputs under the unsupervised learning fashion [22]. A typical AE consists of two main blocks, including an encoder block compressing the inputs into the low dimensional representation and a decoder block that is trained to reconstruct the inputs from the features extracted with the encoder block [22]. The encoder block is a strong feature extractor that can be designed by suitable output layers and then fine-tuned to receive the eligible features [21]. Minimization of the error calculated from the regenerated input images is achieved by optimizing Eq. (1).

$$J(\theta) = \sum (x,\ z)\ \theta = \left\{w,\ w',\ b,\ b'\right\} \tag{1}$$

In the proposed method, the encoder block takes images as inputs. Therefore, the first layers of the encoder block should be convolutional. Thus, this type of AEs is called a convolutional auto-encoder (CAE). CAE has been used to exploit the power of CNN in feature extraction [21]. The patches which are extracted from each image are fed into CAE as input. The best architecture of CAE among the examined architectures in the proposed method for classifying Gastrointestinal (GI) tract lesions from endoscopic video frames is shown in Table 1.

AE is trained by Adam optimizer [18] with learning rate of 0.001 and a Mean Squared Error (MSE) loss function. At last, for each image, the features which are produced by the encoder's layer of our designed CAE will be saved. A list of different hyper-parameter values for each proposed scenario is shown in Table 4.

2.2.3 Visualizing data into two dimensions and generating a 2-D data map

In this section, for constructing creative inputs for training by CNN, the features generated by encoder's block is reshaped into (N, 4*4*8) size. Next, the similarity relationship between the distinctive patches are extracted from each image is visualized by drawing a scatter plot into a

Table 1 CAE architecture for classifying GI tract lesions from endoscopic video frames

Layer	Output Shape	#Parameters
Input Layer	(None, 32, 32, 3)	0
Conv2D	(None, 32, 32, 16)	448
MaxPooling	(None, 16, 16, 16)	0
Conv2D	(None, 16, 16, 8)	1160
MaxPooling	(None, 8, 8, 8)	0
Conv2D	(None, 8, 8, 8)	584
MaxPooling2D (Bottle-neck)	(None, 4, 4, 8)	0
Conv2DTranspose	(None, 4, 4, 8)	584
Upsampling2D	(None, 8, 8, 8)	0
Conv2DTranspose	(None, 8, 8, 8)	584
Upsampling2D	(None, 16, 16, 8)	0
Conv2DTranspose	(None, 16, 16, 16)	1168
Upsampling2D	(None, 32, 32, 16)	0
Output Layer (Conv2D)	(None, 32, 32, 3)	512

2-dimensional data map. The resolution of new data maps is 378×248 pixels. The samples of the 2-D data map for the patches of each image are illustrated in Fig. 2.

2.2.4 Designing and training end-to-end CNN classifier

The images of the 2-D data map are fed to a CNN in an end-to-end fashion to be classified as different classes of GI tract lesions. CNN is a type of deep neural network that can learn hierarchical features from low to high dimensions. The grid search method is c for tuning the hyper-parameters. The hyper-parameters of the examined CNNs tuned by this method include the learning rate, the activation function, dropout rate, batch size, the optimizer,

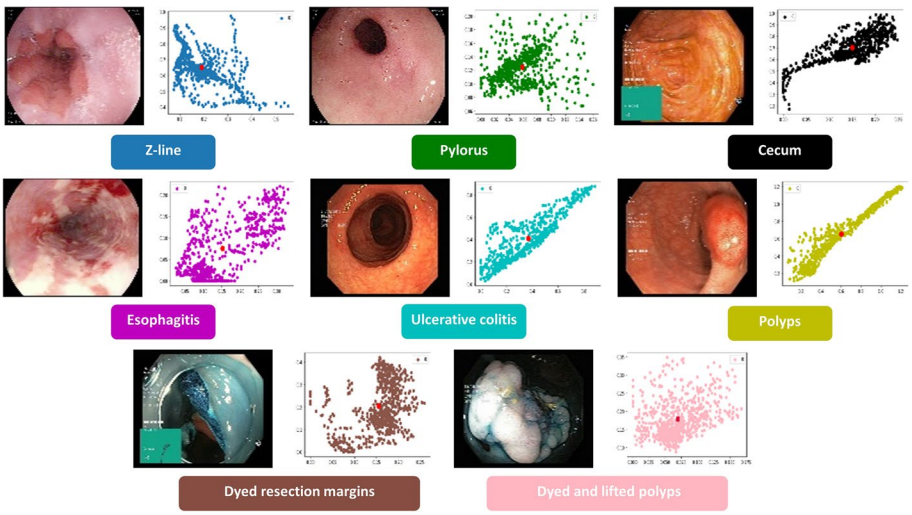

Fig. 2 The 2-D data map for the patches of each image

the number of convolutional blocks, and the number of neurons in convolutional layers. Therefore, the architecture of CNN among the compared and examined ones having the best performance on the validation dataset is shown in Table 2.

According to Table 2, CNN architectonics consists of many layers namely convolution, pooling, dropout, fully connected layer, and dense. Any change in the structure of CNN leads to the creation of new architectonics. The convolutional layers consist of filters of different sizes which slide over the input images. By this layer, the feature of the image is learned and saved into a feature map [26]. W is known as the height of the filters, so, the filter will be $W \times W$ as the multiple of width to height. For the sake of computing pre-non-linearity input in each layer (x_{ij}^{l}), the filter works the same as Eq. (2).

$$y_{j}^{l} = f\left(\sum_{j} x_{j}^{l-1} \otimes w_{ij}^{l} + b_{i}^{l}\right) \tag{2}$$

In Eq. (2), x is the input, y is the output, w is the convolution filter, and b indicates the bias. Sliding of the convolution filters is extracted features from the input images and reduces the parameters. In addition, the pooling layer is used to reduce the parameters. This layer is used for transitioning the proper features to other layers and is included max-pooling, average-pooling and sum-pooling. The max-pooling that is worn in this study is deliberated by Eq. (3).

$$P_{jm} = {}_{k=1}{}^{r}\max\left(x_{j(m-1)n+k}\right) \tag{3}$$

In Eq. (3), x is known as the input matrix. In our method, to overcome the vanishing and the exploding gradient after the convolution layer's operation used the activation functions, ReLU, which is introduced the feature of non-linearity to the DNNs. ReLU is calculated by $max(0.x)$.

During training the model, ReLU can die, so, the Leaky ReLU, which is indicated by Eq. (4), is used to overcome this problem.

$$Irelu(x) = \begin{cases} ax & if\ x \leq 0 \\ x & if\ x > 0 \end{cases} \tag{4}$$

Table 2 CNN architecture of the first scenario for classifying GI tract lesions from endoscopic video frames

Layer	Output shape	#Parameters
Input Layer	(None, 32, 32, 3)	0
Conv2D	(None, 32, 32, 32)	896
MaxPooling	(None, 16, 16, 32)	0
Conv2D	(None, 16, 16, 64)	18496
Dropout	(None, 16, 16, 64)	0
MaxPooling2D	(None, 8, 8, 64)	0
Dropout	(None, 8, 8, 64)	0
Conv2D	(None, 8, 8, 256)	147712
MaxPooling2D	(None, 4, 4, 256)	0
Flatten	(None, 4096)	0
Dense	(None, 128)	524416
Dense	(None, 8)	1032

In the last dense layer to classify the inputs, is used the SOFTMAX activation functions which is calculated as Eq. (5).

$$SOFTMAX(y_i) = \frac{y_i}{\sum_j y_j} \tag{5}$$

CNN is trained for 100 epochs with Adam optimizer [18] with learning rate of 0.001 and batch size of 512. The activation function for all layers except last layer is ReLU [24]. The last layer uses Softmax activation function.

As shown in Fig. 1c, the main steps of our proposed and designed AE-2DV-CNN are described in Algorithm 1.

2.2.5 Evaluation and validation

Different scenarios should be assessed by precise metrics to weigh their strength generalizability. These measures include accuracy, precision, recall, F1-Score and Area under Receiver Operating Characteristics (ROC) curve (AUC) [9].

The value of accuracy shows the strength of the model in classifying the data Eq. (6) [9]:

$$Accuracy = \frac{TP + TN}{N} \tag{6}$$

(TP) is the abbreviation of True Positives, (TN) is the abbreviation of True Negative, and N is the all numbers of data records.

Precision denotes what portion of data is predicted exactly like its actual label [9]. This measure calculates as Eq. (7).

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall is known as *true positive rate* denoted in Eq. (8) and it shows the portion of positive classes which is predicted right.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

(FP) is the abbreviation of False Positives, (FN) is the abbreviation of False Negative.

The F1-measure is the harmonic mean of precision and recall, as is shown in Eq. (9) [9].

Algorithm 1 The steps for training CAE-2DV-CNN

Input: Endoscopic video frames for GI tract
Output: Class labels describing anatomical landmarks, pathological findings, and polyp removal shown in the input video frames
Main Steps of Algorithm A:

1- Resize video frames into 64×64 pixels.
2- Extract N=1089 patches from each image.
3- Add patches to CAE.
4- Visualize the encoder's output into a 2-D data map.
5- Train the CNN model with a 2-D data map for K epochs (in this study, K=100) with an optimizer (in this study, Adam optimizer is used) and batch size of B (in this study, B=512).

$$F1-measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

In the following, Eqs. (10)-(15) shows the measures which are estimated the performance of the multi classes classification [33]:

$$micro-averaged\ precision = \frac{\sum_{c=1}^{NOC} TP_c}{\sum_{c=1}^{NOC} TP_c + \sum_{c=1}^{NOC} FP_c} \tag{10}$$

$$micro-averaged\ recall = \frac{\sum_{c=1}^{NOC} TP_c}{\sum_{c=1}^{NOC} TP_c + \sum_{c=1}^{NOC} FN_c} \tag{11}$$

$$micro-averaged\ F1-score = 2 \times \frac{micro-averaged\ precision \times micro-averaged\ precision}{micro-averaged\ precision + micro-averaged\ precision} \tag{12}$$

$$macro-averaged\ precision = \frac{1}{NOC} \times \sum_{c=1}^{NOC} \frac{TP_c}{TP_c + FP_c} \tag{13}$$

$$macro-averaged\ recall = \frac{1}{NOC} \times \sum_{c=1}^{NOC} \frac{TP_c}{TP_c + FN_c} \tag{14}$$

$$macro-averaged\ F1-score = \frac{1}{NOC} \times \sum_{c=1}^{NOC} 2 \times \frac{\frac{TP_c}{TP_c+FP_c} \times \frac{TP_c}{TP_c+FN_c}}{\frac{TP_c}{TP_c+FP_c} + \frac{TP_c}{TP_c+FN_c}} \tag{15}$$

In the above equations, NOC is the number of different classes.

2.3 The second scenario (CNN-Sc) for classifying GI tract lesions from endoscopic video frames

In the second scenario, an end-to-end CNN model is designed, and KVASIR V1 dataset is fed into it as its inputs. The hyper-parameter tuning is determined with the Grid search method. The architecture model with the best performance compared to the examined ones is listed in Table 3.

The second scenario is trained for 100 epochs with Adam optimizer [18] with a learning rate of 0.001, and a batch size of 512.

2.4 The third scenario (Incept_TL) for classifying GI tract lesions from endoscopic video frames

For the last scenario, we assess different pre-trained CNNs models such as MobileNet [29], Inception-V3 [15], VGG16 and VGG19 [34] to compare their results. Figure 2 demonstrates the performance measures of different models are evaluated in the last scenario.

Table 3 s scenario (CNN-Sc) architecture for classifying GI tract lesions

Layer	Output shape	#Parameters
Input Layer	(None, 32, 32, 3)	0
Conv2D	(None, 32, 32, 32)	896
MaxPooling	(None, 16, 16, 32)	0
Batch- Normalization	(None, 16, 16, 32)	128
Dropout	(None, 16, 16, 32)	0
Conv2D	(None, 16, 16, 64)	18,496
MaxPooling2D	(None, 8, 8, 64)	0
LeakyReLU	(None, 8, 8, 64)	0
Batch- Normalization	(None, 8, 8, 64)	256
Dropout	(None, 8, 8, 64)	0
Conv2D	(None, 8, 8, 64)	36,928
MaxPooling2D	(None, 4, 4, 64)	0
LeakyReLU	(None, 4, 4, 64)	0
Batch- Normalization	(None, 4, 4, 64)	256
Dropout	(None, 4, 4, 64)	0
Flatten	(None, 1024)	0
Dense	(None, 128)	131,200
Batch- Normalization	(None, 128)	512
Dropout	(None, 128)	0
Dense	(None, 8)	1032

As is shown in Fig. 2, Inception-V3 demonstrates the best performance among the compared and examined pre-trained CNNs for transfer learning in the proposed method.

2.4.1 Transfer learning

We use the pre-trained CNNs trained previously on Image-net dataset. The convolutional layers of the pre-trained CNNs are locked to prevent from changing their connection weights. To avoid the overfitting, dropout layers are added. The last layer of CNN is replaced with a dense layer with eight neurons, and the SOFTMAX activation function. Finally, the root-mean-square propagation (RMSprop) optimizer is used with the learning rate of 0.01 to tune last dense layer's weight. The model is trained for 100 epochs with the batch size of 512. The images are resized to 75×75 pixels and are fed into the model as inputs.

Table 4 indicates the best value hyper-parameters of models in different scenarios.

3 Results and discussion

In this section, different scenarios are compared based on the classification performance measures to find the scenario which has the best efficiency.

In Table 5, the performance measures for each proposed scenario for classifying GI tract lesions are reported based on five-fold cross validation strategy.

As illustrated by Table 5, the accuracies of 99.87 ± 0.001, 92.07 ± 0.086, and 90.55 ± 0.111 are achieved for CAE-2DV-CNN, CNN-Sc, and Incept_TL, respectively.

Table 4 The hyper-parameters of different scenarios

Scenarios	Models	Hyper-parameters						
		Loss function	Activation function of layers	Activation function of the last layer	Optimizer	Learning rate	Batch size	No. of epochs
First scenario (CAE-2DV-CNN)	CAE	MSE	ReLU	ReLU	Adam	0.01	1000	8
	CNN	Categorical cross-entropy	ReLU	Softmax	Adam	0.001	512	100
Second scenario (CNN-Sc)	CNN	Categorical cross-entropy	ReLU	Softmax	Adam	0.001	512	100
Third scenario (Incept_TL)	Inception-V3	Categorical cross-entropy	ReLU	Softmax	RMSprop	0.01	512	100

Table 5 The performance metrics of the scenarios for classifying GI tract lesions from endoscopic video frames

Scenario	Metrics	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± std
First scenario (CAE-2DV-CNN)	Accuracy	99.75	99.87	99.87	99.87	100.00	99.87 ± 0.001
	AUC	100.00	100.00	100.00	100.00	100.00	100.00 ± 0.000
	Macro-Precision	100.00	100.00	100.00	100.00	100.00	100.00 ± 0.000
	Macro- Recall	100.00	100.00	100.00	100.00	100.00	100.00 ± 0.000
	Macro-F1-Score	100.00	100.00	100.00	100.00	100.00	100.00 ± 0.000
	Micro-Precision	99.75	99.87	99.87	99.87	100.00	99.87 ± 0.001
	Micro- Recall	99.75	99.87	99.87	99.87	100.00	99.87 ± 0.001
	Micro-F1-Score	99.75	99.87	99.87	99.87	100.00	99.87 ± 0.001
Second scenario (CNN-Sc)	Accuracy	76.99	88.12	96.74	98.75	99.75	92.07 ± 0.086
	AUC	95.94	99.07	99.96	99.99	99.99	98.99 ± 0.016
	Macro-Precision	78.00	88.00	97.00	99.00	100.00	92.40 ± 0.084
	Macro- Recall	77.00	88.00	97.00	99.00	100.00	92.20 ± 0.087
	Macro-F1-Score	76.00	88.00	97.00	99.00	100.00	92.00 ± 0.091
	Micro-Precision	77.00	88.00	96.75	98.75	99.75	92.05 ± 0.086
	Micro- Recall	77.00	88.00	96.75	98.75	99.75	92.05 ± 0.086
	Micro-F1-Score	77.00	88.00	96.75	98.75	99.75	92.05 ± 0.086
Third scenario (Incept_TL)	Accuracy	70.87	85.62	96.87	99.50	99.87	90.55 ± 0.111
	AUC	95.04	98.84	99.94	100.00	100.00	98.76 ± 0.019
	Macro-Precision	71.00	86.00	97.00	100.00	100.00	90.80 ± 0.112
	Macro- Recall	72.00	86.00	97.00	100.00	100.00	91.00 ± 0.108
	Macro-F1-Score	71.00	86.00	97.00	100.00	100.00	90.80 ± 0.112
	Micro-Precision	70.88	85.63	96.88	99.50	99.88	90.55 ± 0.111
	Micro- Recall	70.88	85.63	96.88	99.50	99.88	90.55 ± 0.111
	Micro-F1-Score	70.88	85.63	96.88	99.50	99.88	90.55 ± 0.111

So, the first scenario (CAE-2DV-CNN) has the best efficiency compared to the other scenarios. It demonstrates our innovative approach to using 2-D maps instead of initial images to train with CNN is thoughtful.

Some researchers have trained and evaluated their proposed methods on the KVASIR V1 dataset. Their techniques, and the results are listed and are compared with the proposed method in Table 6.

By assessing the performances in Table 6, it is apprehended that our first scenario leads to superior performance compared to the former studies that used the KVASIR V1 dataset.

Figure 3 Illustrates the confusion matrix for each scenario per each data fold used for five-fold C.V. As realized by Fig. 3, the first scenario is classified as 2-D maps correctly. In the last scenario, the misclassification problem alarmed in the previous studies is reduced.

Figure 4 displays Roc curve ofeach class per fold for different scenarios. As shown in Fig. 4, the AUC of various scenarios is highly desirable..

Figure 5 indicates the accuracy and the loss function per epoch in each fold for the various scenarios. As illustrated by Fig. 5, the accuracy and loss function of the first scenario for each fold except, the first fold, are smooth while the second and third scenarios have fluctuated (Fig. 6).

Table 6 Comparing the performance of the scenarios in the proposed method with the previous researches which have worked on KVASIR V1 dataset

Author	Year	Technique	Accuracy	Precision	Recall	F1-Score
The proposed method	2021	CAE-2DV-CNN	99.87 ± 0.001	100.00 ± 0.000	100.00 ± 0.000	100.00 ± 0.000
The proposed method	2021	CNN-Sc	92.07 ± 0.086	92.40 ± 0.084	92.20 ± 0.087	92.00 ± 0.091
The proposed method	2021	Incept_TL	90.55 ± 0.111	90.80 ± 0.112	91.00 ± 0.108	90.80 ± 0.112
Khan et al. [17]	2022	Extreme learning machine (ELM)	98.20	-	-	-
Ahmad et al. [1]	2017	96 Features + SVM	95.00 (for pylorus)	75.40	75.50	75.30
Pogorelov et al. [28]	2017	6 Global features Logistic Model Tree	93.70	74.80	74.80	74.70
Asperti et al. [2]	2017	Ensemble of Inception + fine tuning + data augmentation	91.50	91.50	91.50	91.50

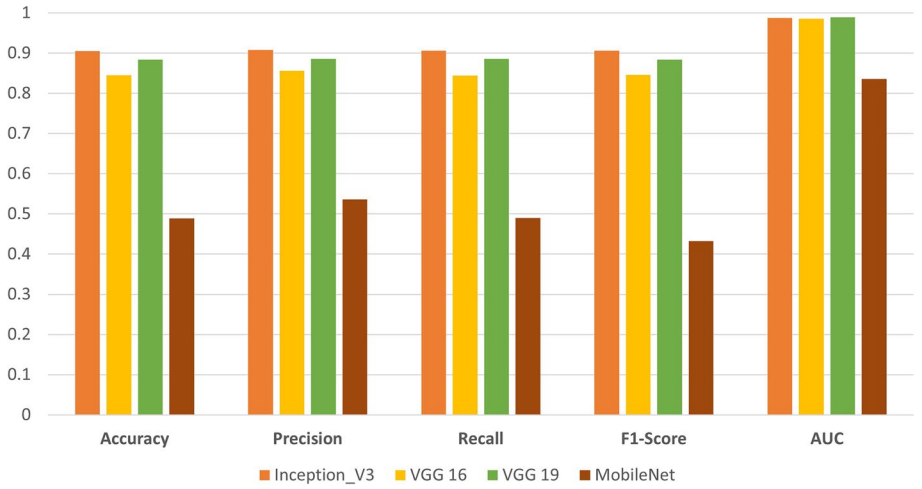

Fig. 3 Comparing the performance measures of different pre-trained CNNs examined in the third scenario

Table 7 represents the processing time details for steps of each scenario in the proposed method, which is calculated by "Google Colab". The maximum amount of RAM is 12.76 GB and the maximum amount of disk is 68.40 GB, which is assigned to each user. The GPU models in "Google Colab" are NVIDIA K80, P100, P4, T4, and V100 GPUs. All deep learning models have used the python libraries like, Scikit-learn, TensorFlow, and Keras.

The principal goal of the proposed method is to suggest novel scenarios that use the advantages of DNNs for classifying GI tract lesions from endoscopic video frames.

Our proposed method in the first scenario can solve some issues which are reported by the previous studies, such as the misclassification of the classes or overfitting of the models.

In addition, in the first scenario, visualizing a 2-D data map represents an innovative approach that leads to the used of the power of CAE in extracting the high-level features and outstanding performance of the model.

4 Conclusions

Misdiagnosis of GI tract abnormalities is exceedingly common during the screening of the GI tract. The accurate diagnosis of the mentioned abnormalities is highly related to the expertise level of the physician and the quality of the images captured by the endoscope and shown on the monitor. Several previous studies have proposed methods with the power of automatic recognition of GI tract abnormalities.

However, low performance in diagnosing some abnormality types are some instances in which suffered from some limitations and drawbacks.

During the advent of computer vision and artificial intelligence, automatic detection has become a common area of research. DNNs are the subset of machine learning methods based on artificial intelligence with the ability the learning. Extracting high-level features from images leads to learning and accessing the superior performance of these models.

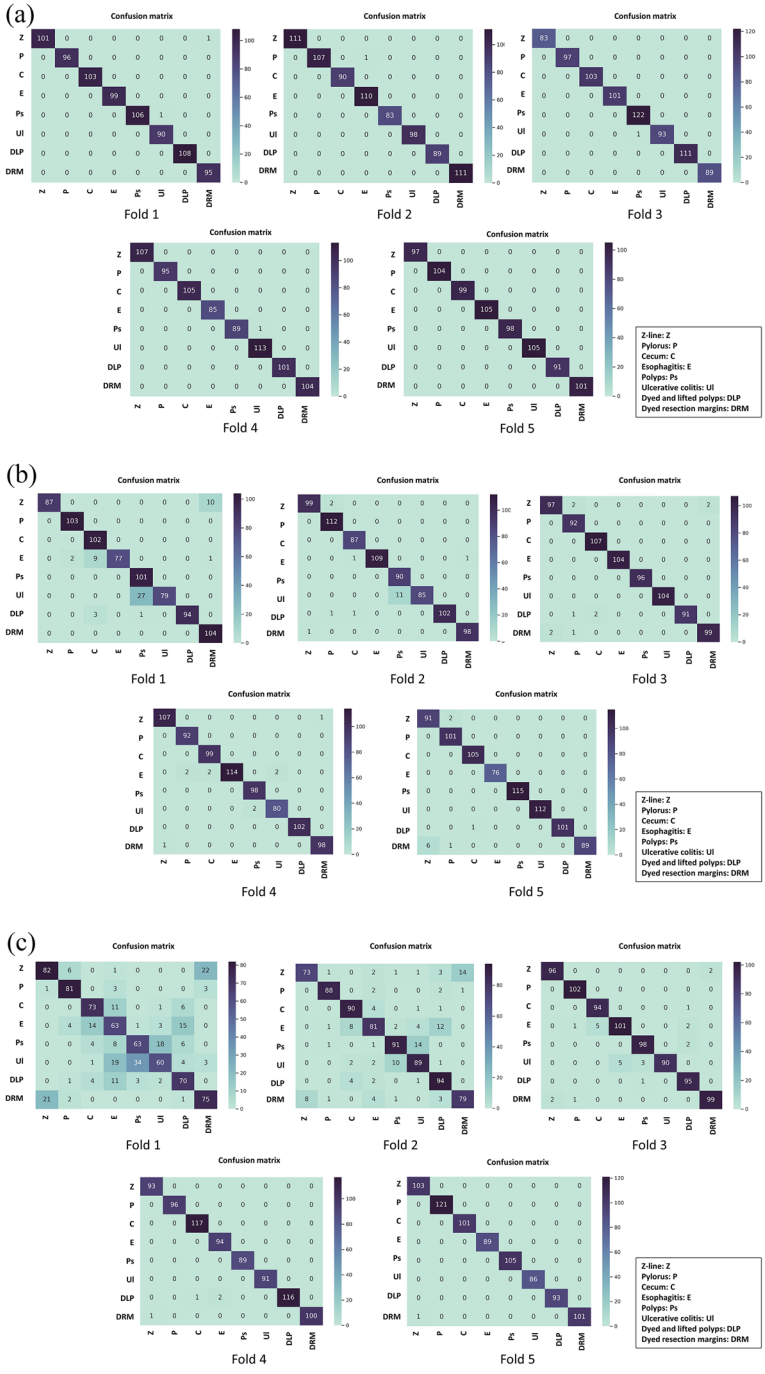

Fig. 4 The confusion matrix for different scenarios per fold. **a** The first scenario for classifying GI tract lesions from endoscopic video frames (AE-2DV-CNN). **b** The second scenario for classifying GI tract lesions from endoscopic video frames (CNN-Sc). **c** The third scenario for classifying GI tract lesions from endoscopic video frames (Incept_TL)

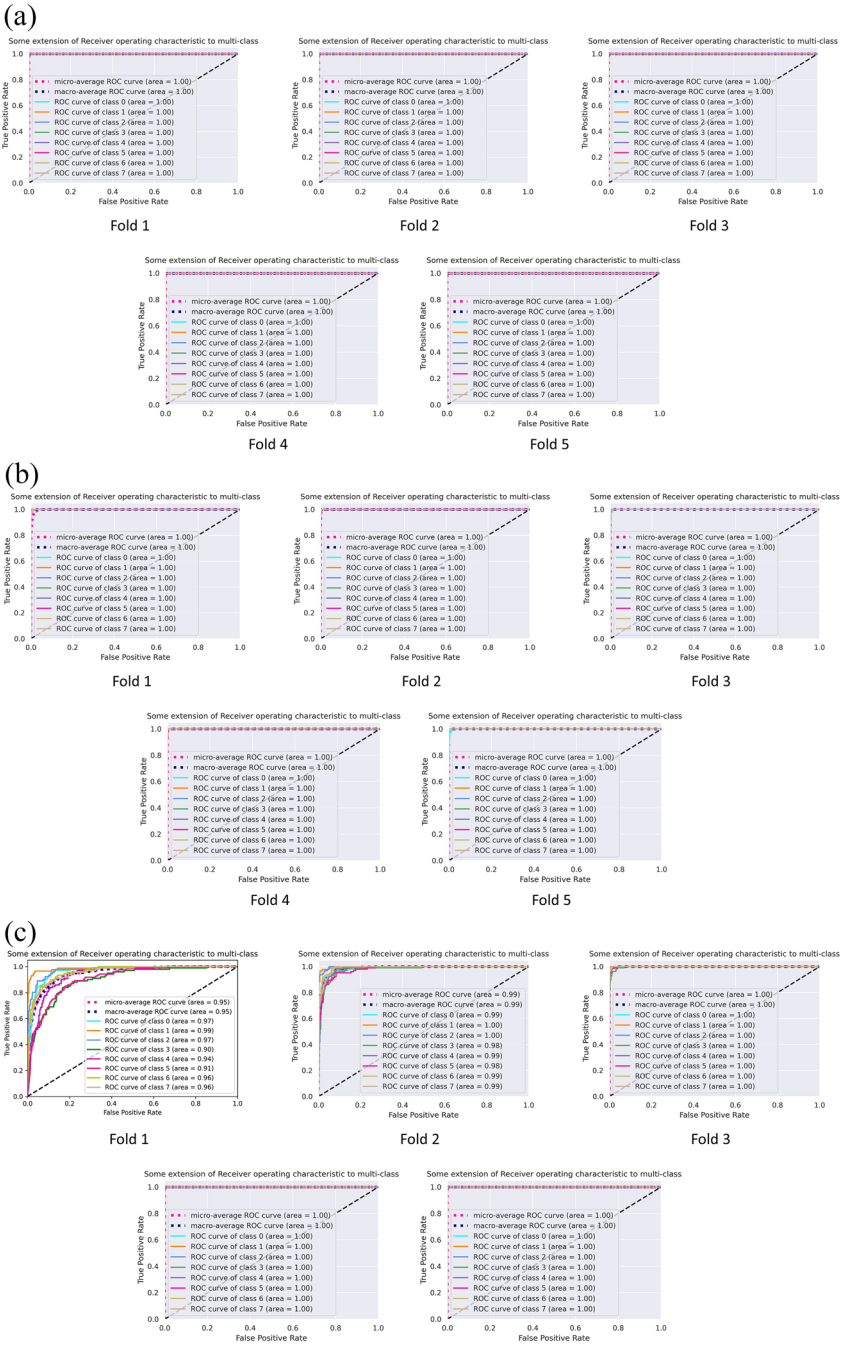

Fig. 5 The ROC curve of each class per fold for different scenarios. **a** The first scenario for classifying GI tract lesions from endoscopic video frames (AE-2DV-CNN). **b** The second scenario for classifying GI tract lesions from endoscopic video frames (CNN-Sc). **c** The third scenario for classifying GI tract lesions from endoscopic video frames (Incept_TL)

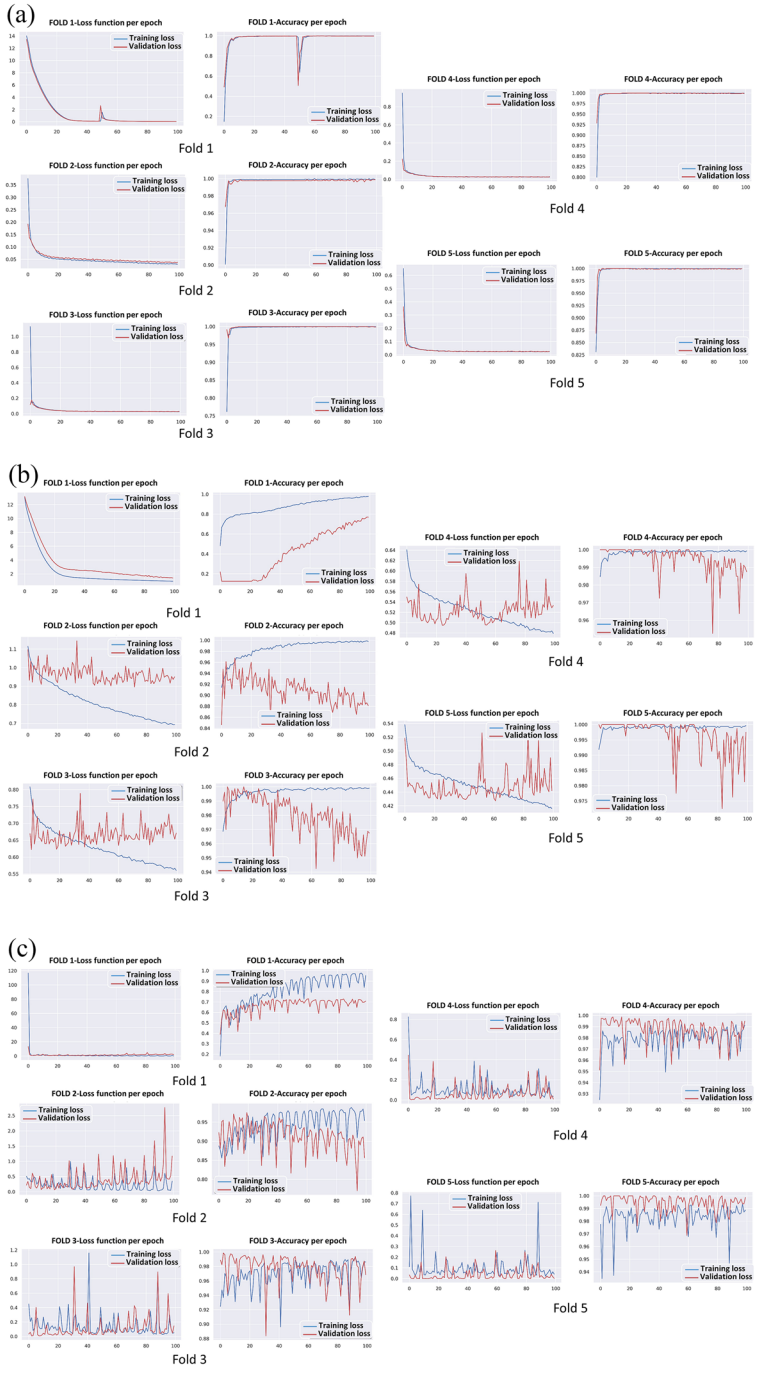

Fig. 6 Training and validation curves of each fold (accuracy and loss function per epoch) for different scenarios. **a** The first scenario for classifying GI tract lesions from endoscopic video frames (AE-2DV-CNN). **b** The second scenario for classifying GI tract lesions from endoscopic video frames (CNN-Sc). **c** The third scenario for classifying GI tract lesions from endoscopic video frames (Incept_TL)

Table 7 The processing time for each step of each scenarios

Scenarios	Processing steps	Processing time (Sec.)
First scenario (CAE-2DV-CNN)	Image preprocessing and patch extraction per image	0.01
	Encoding image by CAE per image	15.60
	Visualizing data into two dimensions and generating a 2-D data map per image	0.32
	Training end-to-end CNN on the training data per fold	924
	Applying the first scenario to testing data per fold	1
Second scenario (CNN-Sc)	Training end-to-end CNN on the training data per fold	575
	Applying the second scenario to testing data per fold	0.71
Third scenario (Incept_TL)	Training end-to-end pre-trained CNN on the training data per fold	2410
	Applying the third scenario to testing data per fold	9.52

So, we aim to use the DNNs model to visualize the high-level features of each image into a 2-D data map and feed them into CNN.

Therefore, in the proposed method, a novel approach is designed and introduced to classify the GI tract lesions from an endoscopic video frame with superior accuracy. The dataset analyzed in the proposed method is KVASIR V1.

In the proposed method, three various scenarios are designed, and their results are compared. The average accuracy of the first, second, and third scenarios is 99.87 ± 0.001, 92.07 ± 0.086 and 90.55 ± 0.111, respectively. The experimental, results demonstrate the superiority of the first proposed scenario in the proposed method over others compared with other ones, and the previous related works focused on the KVASIR V1 dataset. The main novelty of the first scenario is visualizing a 2-D data map from the features extracted by CAE and feeding them into the CNN as inputs.

As a future research direction, it proposes to use and analyze the extended dataset consisting of more abnormalities like Hyperkvasir which is gathered in recent years [10].

Another research opportunity is using other data analysis techniques to extract features with significant correlation from original images and visualize them into 2-D data maps as the inputs of CNN and evaluate their results.

Data availability We use KVASIR V1 dataset in this study for designing and examining our proposed which is publicly available at https://datasets.simula.no/kvasir/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest in this study.

References

1. Ahmad J, Muhammad K, Lee MY, Baik SW (2017) Endoscopic image classification and retrieval using clustered convolutional features, (in Eng). J Med Syst 41(12):196. https://doi.org/10.1007/s10916-017-0836-y
2. Asperti A, Mastronardo C (2017) The effectiveness of data augmentation for detection of gastrointestinal diseases from endoscopical images. arXiv preprint arXiv:1712 03689. https://doi.org/10.1016/j.compmedimag.2020.101852

3. Borgli H et al (2020) HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci Data 7(1):1–14. https://doi.org/10.1038/s41597-020-00622-y
4. Caroppo A, Leone A, Siciliano P (2021) Deep transfer learning approaches for bleeding detection in endoscopy images. Comput Med Imaging Graph 88:101852. https://doi.org/10.1016/j.compmedimag.2020.101852
5. Chauhan NK, Singh K (2018) A review on conventional machine learning vs deep learning. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON), pp 347–352, https://doi.org/10.1109/GUCON.2018.8675097
6. Elhami G, Weber RM (2019) Audio feature extraction with convolutional neural autoencoders with application to voice conversion. Conference: infoscience
7. Ghosh T, Chakareski J (2021) Deep transfer learning for automated intestinal bleeding detection in Capsule Endoscopy Imaging. J Digit Imaging. https://doi.org/10.1007/s10278-021-00428-3
8. Guo X, Yuan Y (2020) Semi-supervised WCE image classification with adaptive aggregated attention. Med Image Anal 64:101733. https://doi.org/10.1016/j.media.2020.101733
9. Han J, Kamber M, Pei J (2011) Data mining concepts and techniques, 3rd edn. The Morgan Kaufmann Series in Data Management Systems 5(4):83–124. https://doi.org/10.1016/C2009-0-61819-5
10. Hasan MM, Hossain MM, Mia S, Ahammad MS, Rahman MM (2022) A combined approach of non-subsampled contourlet transform and convolutional neural network to detect gastrointestinal polyp. Multimedia Tools Appl 81(7):9949–9968. https://doi.org/10.1007/s11042-022-12250-2
11. Heidari M, Mirniaharikandehei S, Khuzani AZ, Danala G, Qiu Y, Zheng B (2020) Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. Int J Med Inform 144:104284. https://doi.org/10.1016/j.ijmedinf.2020.104284
12. Hu H et al (2021) Content-based gastric image retrieval using convolutional neural networks. Int J Imaging Syst Technol 31(1):439–449. https://doi.org/10.1002/ima.22470
13. Hwang M et al (2020) An automated detection system for colonoscopy images using a dual encoder-decoder model, (in Eng). Comput Med Imaging Graph 84:101763. https://doi.org/10.1016/j.compmedimag.2020.101763
14. Jain S et al (2021) A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images. Comput Biol Med 137:104789. https://doi.org/10.1016/j.compbiomed.2021.104789
15. Jha D et al (2020) Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling, 2020. Springer, Berlin, pp 451–462. https://doi.org/10.1007/978-3-030-37734-2_37
16. Jia X, Meng MQ (2017) 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), 11–15 July 2017, pp 3154–3157, https://doi.org/10.1109/EMBC.2017.8037526
17. Khan MA et al (2022) GestroNet: a framework of saliency estimation and optimal deep learning features based gastrointestinal diseases detection and classification. Diagnostics 12(11):2718. [Online]. Available: https://www.mdpi.com/2075-4418/12/11/2718
18. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980. https://doi.org/10.48550/arXiv.1412.6980
19. Leung WK, Cheung KS, Li B, Law SYK, Lui TKL (2021) Applications of machine learning models in the prediction of gastric cancer risk in patients after Helicobacter pylori eradication. Aliment Pharmacol Ther 53(8):864–872. https://doi.org/10.1111/apt.16272
20. Li L et al (2020) Multi-task deep learning for fine-grained classification and grading in breast cancer histopathological images. Multimedia Tools Appl 79:14509–14528. https://doi.org/10.1007/s11042-018-6970-9
21. Maggipinto M, Masiero C, Beghi A, Susto GA (2018) A convolutional autoencoder approach for feature extraction in virtual metrology. Procedia Manuf 17:126–133. https://doi.org/10.1016/j.promfg.2018.10.023
22. McClelland JL, Rumelhart DE, Group PR (1986) Parallel distributed processing. MIT Press, Cambridge. https://doi.org/10.7551/mitpress/5236.001.0001
23. Mohapatra S, Nayak J, Mishra M, Pati GK, Naik B, Swarnkar T (2021) Wavelet transform and deep convolutional neural network-based smart healthcare system for gastrointestinal disease detection, (in Eng). Interdiscip Sci. https://doi.org/10.1007/s12539-021-00417-8
24. Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: ICML. https://dl.acm.org/doi/10.5555/3104322.3104425
25. Owais M, Arsalan M, Choi J, Mahmood T, Park KR (2019) Artificial intelligence-based classification of multiple gastrointestinal diseases using endoscopy videos for clinical diagnosis. J Clin Med 8(7):986. https://doi.org/10.3390/jcm8070986
26. Öztürk Ş, Özkaya U (2020) Gastrointestinal tract classification using improved LSTM based CNN. Multimedia Tools Appl 79(39):28825–28840. https://doi.org/10.1007/s11042-020-09468-3

27. Pannu HS, Ahuja S, Dang N, Soni S, Malhi AK (2020) Deep learning based image classification for intestinal hemorrhage. Multimedia Tools Appl 79:21941–21966. https://doi.org/10.1007/s11042-020-08905-7
28. Pogorelov K et al (2017) KVASIR: a Multi-Class Image dataset for computer aided gastrointestinal disease detection. https://doi.org/10.1145/3193289
29. Ponnusamy R, Sathiamoorthy S (2019) Prediction of esophagitis and Z-line from wireless capsule endoscopy images using fusion of low-level features. Int J Recent Technol Eng (IJRTE) 8(3):6024–6028. https://doi.org/10.35940/ijrte.C5568.098319
30. Raksasat R et al (2021) Accurate surface ultraviolet radiation forecasting for clinical applications with deep neural network. Sci Rep 11(1):5031. https://doi.org/10.1038/s41598-021-84396-2
31. Rau A et al (2019) Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int J Comput Assist Radiol Surg 14(7):1167–1176. https://doi.org/10.1007/s11548-019-01962-w
32. Safarov S, Whangbo TK (2021) A-denseunet: Adaptive densely connected unet for polyp segmentation in colonoscopy images with atrous convolution. Sensors 21(4):1–15, Art no. 1441. https://doi.org/10.3390/s21041441
33. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45(4):427–437. https://doi.org/10.1016/j.ipm.2009.03.002
34. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J R Stat Soc: Ser B (Methodological) 36(2):111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x
35. Sung H et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Cancer J Clin. https://doi.org/10.3322/caac.21660
36. Vieira PM, Freitas NR, Valente J, Vaz IF, Rolanda C, Lima CS (2020) Automatic detection of small bowel tumors in wireless capsule endoscopy images using ensemble learning. Med Phys 47(1):52–63. https://doi.org/10.1002/mp.13709
37. Xing X, Yuan Y, Meng MQH (2020) Zoom in lesions for better diagnosis: attention guided deformation network for WCE image classification. IEEE Trans Med Imaging 39(12):4047–4059. https://doi.org/10.1109/TMI.2020.3010102
38. Yuan Y, Meng MQH (2017) Deep learning for polyp recognition in wireless capsule endoscopy images. Med Phys 44(4):1379–1389. https://doi.org/10.1002/mp.12147
39. Zhang C, Zhang N, Wang D, Cao Y, Liu B (2020) Artifact detection in endoscopic video with deep convolutional neural networks. In: 2020 Second International Conference on Transdisciplinary AI (TransAI), pp 1–8. https://doi.org/10.1109/TransAI49837.2020.00007