

# Performance analysis of convolutional neural network architectures over wireless capsule endoscopy dataset

Parminder Kaur<sup>1</sup>, Rakesh Kumar<sup>2</sup>

<sup>1</sup>Department of Computer Science, Dr. B.R. Ambedkar Government College, Kaithal, Haryana, India

<sup>2</sup>Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India

## Article Info

### Article history:

Received Jan 25, 2023

Revised Jun 4, 2023

Accepted Sep 11, 2023

### Keywords:

Capsule endoscopy

Convolution

Inception

MobileNet

Visual geometry group

## ABSTRACT

Wireless capsule endoscopy is one of the diagnostic methods used to record the video of the gastrointestinal tract. The endoscopy capsule stays in the digestive system for at least eight hours. It is difficult for gastroenterologists to examine such a lengthy video and identify the ailment. Convolutional neural networks (CNN) are a powerful solution to several computer vision problems. CNN can speed up the reviewing time of the recorded video by classifying video frames into various categories. The primary emphasis of this research paper is to examine and evaluate the performance of three different CNN architectures-VGG, inception, and MobileNet-in classifying the disease. Experimental results demonstrate that MobileNetV2's accuracy is 91%, whereas InceptionV3 and VGG16 have an accuracy of 94% which is better than the accuracy of MobileNetV3. However, MobileNetV2 performed relatively better than the other CNN models in terms of computational time and cost. The model's F-score, precision, and recall values are computed and compared also.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Parminder Kaur

Department of Computer Science, Dr. B. R. Ambedkar Government College

Kaithal, Haryana, India

Email: kaur.parminder490@gmail.com

## 1. INTRODUCTION

Wireless capsule endoscopy (WCE) [1]–[3] is one of the most commonly used diagnostic techniques for detecting abdominal and gastrointestinal disorders. The patient in this technique swallows a pill-sized capsule camera. It stays inside the patient's body for approximately eight to nine hours and records video of the digestive system. One of the biggest challenges that a gastroenterologist faces is extrapolating out the disease after observing an 8-9 hour endoscopic video. A potential solution to this problem is an automated disease classification process. An automated classification technique cuts the gastroenterologist's video reviewing time. Many approaches to classification or video summarization of wireless capsule endoscopy utilize color as a prime feature for detecting higher-importance frames. Color is a prominent feature of the human visual attention system. All electronic devices use red, green, blue (RGB) color models to capture and process the images. However, the red, green, and blue channels of the RGB color model do not separate the color and intensity information. So, most video summarization approaches based on color models convert the color space from RGB to HSV or HSI. Using a similar approach, converted the endoscopy images from RGB to hue, saturation, intensity (HSI) colorspace to detect the frame's bleeding and non-bleeding regions [4]. Different organs have different colors and textures, leveraging the fact used hue, saturation, value (HSV) colorspace and texture to extract the representative frames [5]. The texture is another feature to which the human eye is sensitive. A video summarization technique that primarily groups similar frames, to form a clip is proposed in [6]. The similarity between the adjacent frames was calculated based on

color and texture data. While gray level co-occurrence matrix (GLCM) was used to calculate texture, an HSV color histogram represented the color. The author used an adaptive threshold to group similar frames into one clip using an adaptive threshold. And finally, to generate the video summary, an adaptive K-means clustering technique was applied that selects the representative frame from each cluster.

Numerous video summarization approaches are based on the Keyframe extraction technique. A video's representative frames containing important information are called Keyframes. The video is first separated into shots; then, Keyframes are identified in those shots. A shot is a segment of a video with similar content. The shot boundaries can be identified in several ways. Employed a Siamese neural network for extracting the high-level features of the frames and finding the similarity and dissimilarities between those frames based on the features [7]. A support vector machine (SVM) classifier is used to identify the threshold of similarity detection, and the similar frames form a shot. Later, the K-means algorithm is applied to choose a representative frame from that shot. How the capsule moves through the gastrointestinal (GI) tract varies, depending on where it is in the tract. Sushma and Aparna [8] used motion score, direction, and energy for shot boundary identification. To make an automated analysis system that quickly diagnoses a disease, a multi-layer perceptron-based computationally light model that could be deployed over a microcontroller inside a capsule camera [9]. This proposed methodology has been restricted to finding bleeding areas only. A modified multiclass SVM classifier [4] was used to categorize the frames into bleeding, non-bleeding, and background frames. However, adopting a multiclass SVM classifier has the drawback of making the model more complex by merging multiple SVMs. The proposed framework of [9], [10] can only identify gastrointestinal tract bleeding. Since blood is red, most of these bleeding detection techniques exploit the color feature. Numerous studies aim to identify a particular ailment like abnormal bleeding, tumor, polyp, or ulcer, but if a problem-specific method is adopted, that proposed framework can be applied to identifying a specific type of disease only, and it cannot be considered a generic model. A method for polyp detection with a lower false-positive rate [11]. The author used the shape feature because polyps are rounded or curved growths in the colon. A method for polyp detection was developed in [12], using the textural characteristics of polyps and an improved bag of features. Uniform local binary pattern (LBP) for detecting the edges of the polyps [13]. A method to detect Crohn's disease using a deep convolutional network is proposed in [14]. A method for tumor detection that uses color and textural features is employed in [15]. The proposed technique feeds the feature maps produced by the LBP operator to the SVM to identify tumors. Was very effective in finding the ulcers [16]. AlexNet convolutional neural network (CNN) detected small intestine lesions in [17]. Several researchers used various unsupervised learning methods to generate video summaries. Lan and Ye [18] proposed an unsupervised learning-based WCE video summarization technique that primarily has a deep summarizer network. The summarizer network uses long short-term memory (LSTM), a variational autoencoder, generative adversarial networks, and other components to generate the summaries. Similarly, unsupervised learning and a feature weight-assigning algorithm to generate the video summaries [19]. Employing a weight assignment algorithm to assign weight to features helped eliminate the redundant frames and prioritize the informative frames.

A method that can identify just one type of disease is impossible as a broad solution for disease identification. Therefore, a generic method capable of detecting all gastrointestinal diseases is required. Additionally, the method should use semantic or deep characteristics to help capture the precise details. Deep learning techniques might be the answer to this. Both feature extraction and deep learning can be done with neural networks. Neural networks learn the deep features from a labeled dataset. A CNN is built using the convolution operator. There are several CNN architectures. This paper discusses the performance of various CNN on the WCE Endoscopy dataset. Section 2 discusses the methodology. Section 3 discusses the three results of the experiments. Finally, in section 4, conclusions are formed.

## 2. METHOD

An algorithm for deep learning is a CNN. A neural network qualifies to be considered as a CNN even if it has only one convolution layer because it is essentially a neural network with some extra convolution layers. They accomplish the deep semantic feature extraction of images at their finest. CNN are the most commonly used technique in computer vision for classifying and identifying objects. The convolution operation makes CNN unbiased to any local variations and image modifications. A CNN can efficiently recognize an object in a frame regardless of its position, size, or orientation. The role of CNN in image analysis is significant, especially in fields like medical image processing, where the data is highly susceptible. Many researchers have employed CNN for feature extraction, transfer learning, or abnormality detection. Figure 1 illustrates a CNN's basic structure-convolution layer, pooling layer, and dense layer. Every CNN has these essential components or layers with some transformations or adaptations.

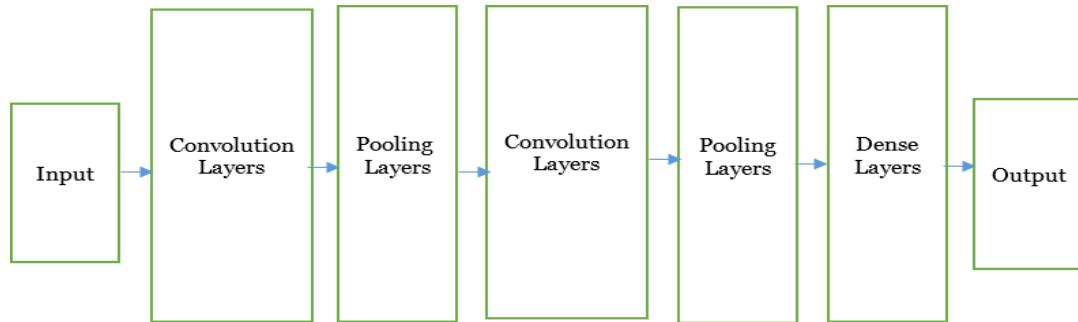


Figure 1. The basic architecture of CNN

A convolution layer applies a kernel (filter) to an input image. The filter size may vary depending on the architecture. A kernel can be applied both pointwise and depthwise. After applying a kernel to an image, a feature map of the image is obtained that can be further processed. The feature map size is more ample than the input image, so another pooling layer is added to the CNN, reducing the feature map's size. A diminished feature map contributes to reducing the overall computational time and cost. The most popular pooling technique is called max pooling. As the name implies, a max pooling procedure chooses a maximum value from a patch of the feature map. A dense layer has densely linked neurons. It is also known as a fully connected layer. Its function is to accept the input (learned features) from the preceding layers and produce the output. Specifically, it completes the data categorization task in the last step.

A CNN can solve a problem by either building a network from scratch or retraining an already-developed model for a new type of problem. Since building a new CNN from scratch is time-consuming, this paper used the three well-known CNN models visual geometry group (VGG), inception, and MobileNet. ImageNet [20] database, which contains millions of annotated images for object classification, was used to train VGG, Inception, and MobileNet. This ImageNet database contains 1000 classes. In the 2014 imagenet large scale visual recognition challenge (ILSVRC), InceptionV1 and VGG16 were introduced, winning first and second position, respectively. This work focuses on solving gastrointestinal abnormality identification and classifying that abnormality into a particular type of disease class. The procedure adopted for this purpose uses the already-developed and well-trained models over the capsule endoscopy dataset for disease classification. The methodology involves: i) generating the frames from a video dataset; ii) normalizing and preprocessing the frames; iii) using the CNN-VGG16, InceptionV3, MobileNetV2, and MobileNetV3 for classifying the disease; and iv) evaluating the results from all the models.

### 2.1. Visual geometry group neural network

The VGG is an Oxford University research team. The team of researchers developed the VGG architectures. VGG16 [21] and VGG19 are intense neural networks but they do not have complex architecture. The 16 layers of the VGG16 network are stacked on top of one another. It accepts an image with a size of  $224 \times 224 \times 3$ . When compared to other architectures, VGG16 attained a significantly greater accuracy. As shown in Figure 2, the architecture comprises thirteen convolution layers, three completely connected layers, and five pooling layers. In the convolution layer, a  $3 \times 3$  filter with a stride of size one is employed, and a  $2 \times 2$  filter with a stride of size two is used in the pooling layer. It uses the max pooling technique. The final, fully connected layer provides the output. The VGG19 variant of VGG16 contains three additional convolution layers above VGG16, however, the rest of the architecture is the same. It is observed that increasing the number of convolutional layers in an architecture improves the system's performance and effortlessly drives challenging problem-solving tasks.

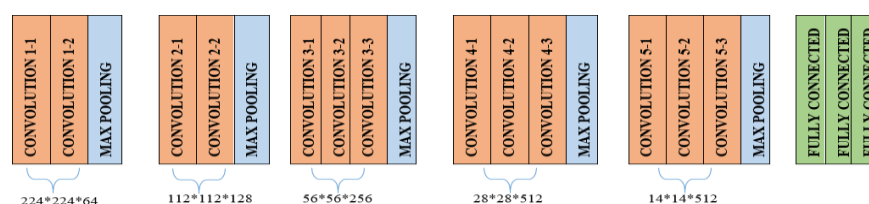


Figure 2. The architecture of the VGG16 network

## 2.2. Inception

The architecture of inception [22] networks is very intricate and sophisticated. An inception network focuses more on the broader network than a deep neural network. Inception network is available in many different versions, including InceptionV1, InceptionV2, InceptionV3, InceptionV4, and Inception-Resnet. An inception network consists of a naive module (Figure 3) and a dimensionality reduction module (Figure 4). As shown in Figure 3, the naive module contains convolutions of various sizes. The goal of this module is to locate the information that has been dispersed around the frame in any area by applying filters of different sizes. Dimensionality reduction-the second module of the inception model is similar to the first module except that an additional convolution of size  $1 \times 1$  is added to the convolution layers of size  $3 \times 3$  and  $5 \times 5$ , as depicted in Figure 4. Adding a convolution of size  $1 \times 1$  reduces the number of parameters and as the number of parameters gets reduced, the computational cost also gets reduced. All versions of inception models share these two modules. Only a few minor modifications have been made to filter sizes. The large-sized filters in InceptionV3 are further factorized into smaller, asymmetrical filters. For instance, two asymmetric filters of sizes  $3 \times 1$  and  $1 \times 3$  are used in place of one filter of size  $3 \times 3$ . The computational cost is reduced when the filter size is factored.

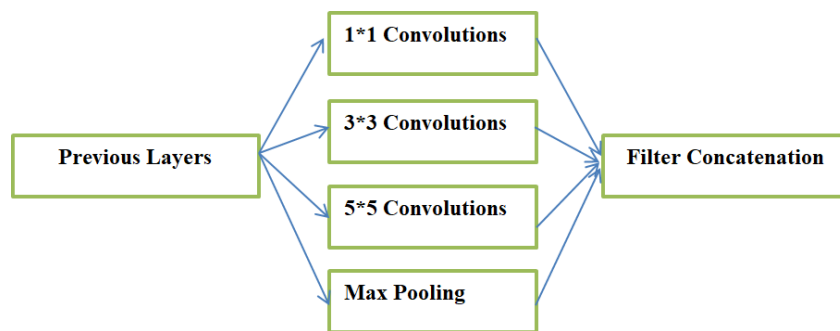


Figure 3. Naive module of inception network

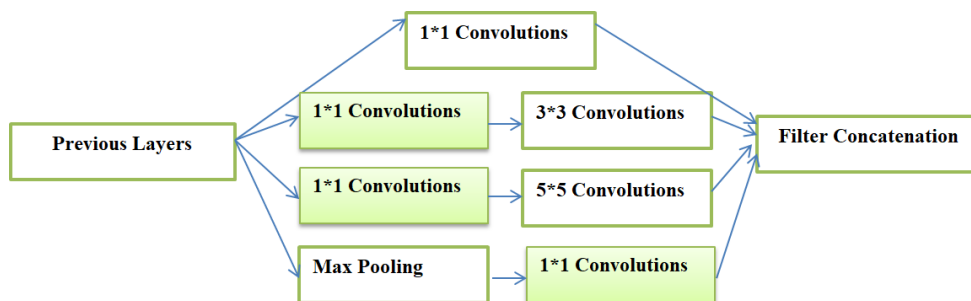


Figure 4. Inception module with dimensionality reduction

## 2.3. MobileNet

MobileNet is developed by Howard *et al.* [23]. MobileNets, as the name suggests, were created to offer a computationally light and valuable model that can be used on mobile devices. This model improves accuracy while reducing the computational latency. MobileNet employs depth-wise separable convolution, where each channel of the input image is first subjected to a depth-wise filter before being combined by a pointwise filter. It also uses the width and resolution multiplier, the two hyperparameters that help the MobileNet to be smaller than conventional convolution networks. MobileNet is available in various versions, including MobilenetV1, MobileNetV2 [24], and the most recent Mobilenet V3 [25]. The recent versions of MobileNet models are computationally more affordable and speedy than the earlier ones.

A comparison between the discussed models is listed in Table 1. The comparison is drawn based on the number of kernels used by each model, the pooling technique, the size of the input image, and the number of parameters. VGG16 has the highest number of parameters but the least number of layers; on the contrary MobileNet has the lowest number of parameters but the highest number of layers.

Table 1. Comparison of VGG16, inception, and MobileNet

CNN models	Proposed by	Input image size	Type of convolution	Number and size of kernels	Pooling technique	Pooling number of layers	Number of parameters
VGG16	Simonyan and Zisserman [21]	224×224	Standard convolution	3×3 for convolution and 2×2 for max pooling	Max pooling	16	138 million
Inception	Szegedy <i>et al.</i> [22]	229×229	2 standard convolution increasing the width of the network	3 kernels 1×1, 3×3, 5×5 some versions have asymmetric kernel of 1×3, 3×1, 7×1, 1×7	Average pooling, max pooling	48	24 million
MobileNet	Howard <i>et al.</i> [23]	Any input size greater than 32×32	Depthwise separable convolution	3×3 depth wise convolution followed by 1×1 pointwise convolution	Average pooling	53	13 million

Object detection, feature extraction, and image classification are some of the areas where CNN provides very efficient solutions. Pretrained models save time and cut down on computational costs. InceptionV3 was used by [26] to create the feature maps of the WCE images, and the feature maps were then provided as input to the K-means algorithm to construct the keyframes of the endoscopic video. ResNet, often referred to as residual networks [27] with an SVM classifier to categorize the images into similar and dissimilar groups.

### 3. RESULTS AND DISCUSSION

The models InceptionV3, VGG16, MobileNetV2, and MobileNetV3 were implemented to determine the best-performing model for a wireless capsule endoscopy dataset.

#### 3.1. Hardware specifications

The experiments were conducted on a desktop computer running the Windows 10 operating system. The computer was equipped with an Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz 2.50 GHz processor, which provided sufficient processing power for the experiments. Additionally, the computer had 8 GB of RAM, which was ample for the experimental tasks. The use of this hardware ensured that the experiments were conducted efficiently and without any performance issues.

#### 3.2. Experimental results

The three models discussed and compared in the above sections were implemented in Python. The dataset used for training and validation was obtained from Kaggle [28] (WCE curated colon disease dataset deep learning). This data set consists of images captured by a wireless capsule during endoscopy to diagnose any abnormal condition. The dataset has three sets-training set, test set, and validation set. The WCE dataset considered for the experiments has labeled images. There are four labels-normal, ulcerative colitis, polyps, and esophagitis as shown in Figure 5. The normal (Figure 5(a)) labeled images represent the condition of a healthy, disease-free GI tract. Figure 5(b) shows the condition of Ulcerative colitis, which is an inflammatory disease of the colon, in this situation the patient's colon has ulcers. In some worse cases, a patient with ulcerative colitis may develop colon cancer. Figure 5(c) shows a gastrointestinal tract condition with Polyps, while Figure 5(d) represents esophagitis. The performance of the VGG16, InceptionV3, MobileNetV2, and MobileNetV3 networks over the endoscopy dataset was evaluated.

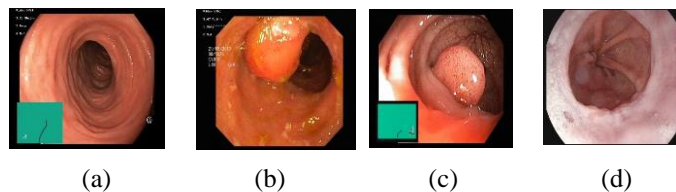


Figure 5. Images from the endoscopy dataset representing different conditions of the GI tract (a) normal, (b) ulcerative colitis, (c) polyps and, (d) esophagitis

### 3.2.1. Evaluation parameters

The performance of all the models discussed is evaluated using F-score. F-score is computed using (1). It is computed based on recall (2) and precision (3):

$$F_{Score} = \frac{2 * Recall * Precision}{Recall + Precision} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

True positives (TP) are the positives that are correctly predicted as positives. False positives (FP) are the negatives that are incorrectly predicted as positives. True negatives (TN) are the negatives that are correctly predicted as negatives. False negatives (FN) are the positives that are incorrectly predicted as negatives. The accuracy of all the models is computed as well.

### 3.2.2. Performance comparison

The models are trained over 30 epochs. F-score (Table 2) indicates that VGG16 and InceptionV3 performed equally well in classifying the diseases. MobileNetV3 could have performed better. F-score value of MobileNetV2 is close to the performance of VGG16 and InceptionV3. In totality, F-score performance of VGG16, Inception3, and MobileNetv2 are the same. The accuracy of the three models is computed over both training data and test data. From Table 3, it can be easily concluded that InceptionV3 and VGG 16 have an accuracy of 0.94, which is better than MobileNet. Precision and recall values also indicate that VGG16 and InceptionV3 achieved the same values of 0.94. If only MobileNetV2 and MobileNetV3's performance are considered, MobileNetV2 performs better than MobileNetV3. It is also reflected in the bar graph of Figure 6 that InceptionV3 and MobileNetV2 have higher accuracies not only on the training dataset but also on the test dataset.

Table 2. F-score of different classes

Class	InceptionV3	VGG16	MobileNetV2	MobileNetV3
Normal	0.97	0.97	0.96	0.84
Ulcerative colitis	0.88	0.88	0.82	0.30
Polyps	0.91	0.91	0.86	0.43
Esophagitis	0.98	0.98	0.97	0.87

Table 3. Precision, accuracy, and recall of different models

	VGG16	InceptionV3	MobileNetV2	MobileNetV3
Precision	0.94	0.94	0.91	0.61
Recall	0.94	0.94	0.91	0.62
Accuracy	0.94	0.94	0.91	0.62

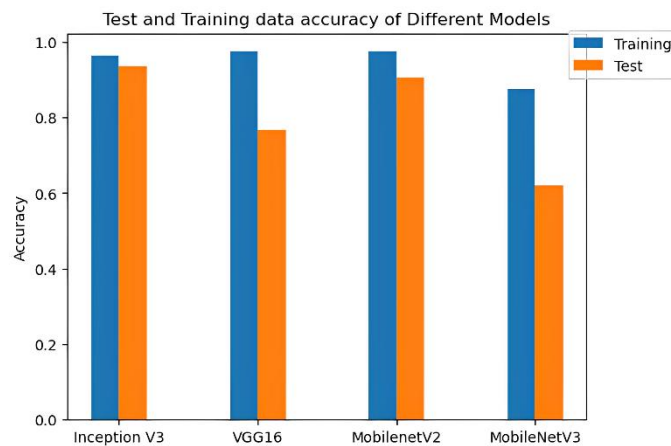


Figure 6. The accuracy of different models over the training and test dataset

From Tables 2 and 3, it can be concluded that VGG16 and inception performed better than that of MobileNet. However, along with accuracy, many other aspects must be considered before judging a model as the best performer. The number of parameters is one of the contributing factors in approximating the complexity of the model. The computational complexity of any model is directly proportional to the number of parameters. Table 4 lists the parameters of the models. It is observed that the highest number of parameters is in InceptionV3, whereas the lowest number of parameters is in MobileNetV2. MobileNetV2 produces results in a very short time if compared with other models.

Table 4. Parameters of different models

	VGG16	InceptionV3	MobileNeV2	MobileNetV3
Trainable parameters	529412	21,802,784	1,315,844	988,164
Total parameters	15,244,100	23,905,060	3,573,828	3,984,516

#### 4. CONCLUSION

In this study, the wireless capsule endoscopy images were classified into various intestinal diseases with the help of InceptionV3, VGG16, MobileNetV2, and MobileNetV3. According to the experimental findings, InceptionV3 and VGG16 outperformed MobileNetV2 in accuracy by a small margin. However, MobileNetV2 and MobileNetV3 performed well when the computational time was considered. In the context of a WCE, the diagnosis punctuality and the findings' accuracy are both crucial. Therefore, the optimum endoscopic video summarization or classification model must possess accuracy and timeliness. MobileNetV2 is the safest alternative in terms of accuracy as well as time. The results from the automated diagnosis system should be followed up by expert advice for the best results. This work can be extended by developing a transfer learning system that only utilizes the trained weights of MobileNetV2 for extracting high-level semantic features. These features can be provided as input for any classification or clustering algorithm.

#### ACKNOWLEDGEMENTS

We thank Kaggle for providing the WCE Curated Colon Disease Dataset.

#### REFERENCES




- [1] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, pp. 417–417, May 2000, doi: 10.1038/35013140.
- [2] P. Swain, "Wireless capsule endoscopy," *Gut*, vol. 52, no. 90004, pp. 48–50, Jun. 2003, doi: 10.1136/gut.52.suppl\_4.iv48.
- [3] A. Wang *et al.*, "Wireless capsule endoscopy," *Gastrointestinal Endoscopy*, vol. 78, no. 6, pp. 805–815, Dec. 2013, doi: 10.1016/j.gie.2013.06.026.
- [4] E. Tuba, M. Tuba, and R. Jovanovic, "An algorithm for automated segmentation for bleeding detection in endoscopic images," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 4579–4586, doi: 10.1109/IJCNN.2017.7966437.
- [5] Jia Sen Huo, Yue Xian Zou, and Lei Li, "An advanced WCE video summary using relation matrix rank," in *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*, Jan. 2012, pp. 675–678, doi: 10.1109/BHI.2012.6211673.
- [6] J. Chen, Y. Wang, and Y. X. Zou, "An adaptive redundant image elimination for Wireless Capsule Endoscopy review based on temporal correlation and color-texture feature similarity," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, Jul. 2015, pp. 735–739, doi: 10.1109/ICDSP.2015.7251973.
- [7] J. Chen, Y. Zou, and Y. Wang, "Wireless capsule endoscopy video summarization: A learning approach based on Siamese neural network and support vector machine," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec. 2016, pp. 1303–1308, doi: 10.1109/ICPR.2016.7899817.
- [8] B. Sushma and P. Aparna, "Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion Analysis," *IEEE Access*, vol. 9, pp. 13691–13703, 2021, doi: 10.1109/ACCESS.2020.3044759.
- [9] M. Hajabdollahi, R. Esfandiarpour, S. M. R. Soroushmehr, N. Karimi, S. Samavi, and K. Najarian, "Segmentation of Bleeding Regions in Wireless Capsule Endoscopy Images an Approach for inside Capsule Video Summarization," *arxiv Computer Science*, 2018, doi: 10.48550/arXiv.1802.07788.
- [10] Y. Yuan, B. Li, and M. Q.-H. Meng, "Bleeding Frame and Region Detection in the Wireless Capsule Endoscopy Video," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 2, pp. 624–630, Mar. 2016, doi: 10.1109/JBHI.2015.2399502.
- [11] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, Feb. 2016, doi: 10.1109/TMI.2015.2487997.
- [12] Y. Yuan, B. Li, and M. Q.-H. Meng, "Improved Bag of Feature for Automatic Polyp Detection in Wireless Capsule Endoscopy Images," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 529–535, Apr. 2016, doi: 10.1109/TASE.2015.2395429.
- [13] T. Ojala, M. Pietikainen, and T. Maenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.
- [14] D. M.- Santos, J. A. C.- Fernandez, I. P.- Borrero, H. P.- Manrique, and M. E. G.- Arias, "Automatic detection of crohn disease in






- wireless capsule endoscopic images using a deep convolutional neural network,” *Applied Intelligence*, vol. 53, no. 10, pp. 12632–12646, May 2023, doi: 10.1007/s10489-022-04146-3.
- [15] B. Li and M. Q.-H. Meng, “Tumor Recognition in Wireless Capsule Endoscopy Images Using Textural Features and SVM-Based Feature Selection,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 3, pp. 323–329, May 2012, doi: 10.1109/TITB.2012.2185807.
  - [16] Y. Yuan, J. Wang, B. Li, and M. Q.-H. Meng, “Saliency Based Ulcer Detection for Wireless Capsule Endoscopy Diagnosis,” *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 2046–2057, Oct. 2015, doi: 10.1109/TMI.2015.2418534.
  - [17] S. Fan, L. Xu, Y. Fan, K. Wei, and L. Li, “Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images,” *Physics in Medicine & Biology*, vol. 63, no. 16, pp. 1–25, Aug. 2018, doi: 10.1088/1361-6560/aad51c.
  - [18] L. Lan and C. Ye, “Recurrent generative adversarial networks for unsupervised WCE video summarization,” *Knowledge-Based Systems*, vol. 222, Jun. 2021, doi: 10.1016/j.knsys.2021.106971.
  - [19] M. M. Ben Ismail, O. Bchir, and A. Z. Emam, “Endoscopy video summarization based on unsupervised learning and feature discrimination,” in *2013 Visual Communications and Image Processing (VCIP)*, Nov. 2013, pp. 1–6, doi: 10.1109/VCIP.2013.6706410.
  - [20] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
  - [21] K. Simonyan and A. Zisserman, “Very deep convolutional networks for Large-Scale image recognition,” *Computer Vision and Pattern Recognition*, Sep. 2014, [Online]. Available: <http://export.arxiv.org/pdf/1409.1556>
  - [22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
  - [23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” *Arxiv Computer Vision and Pattern Recognition*, 2017, doi: 10.48550/arxiv.1704.04861.
  - [24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.
  - [25] A. Howard *et al.*, “Searching for MobileNetV3,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.
  - [26] V. Raut and R. Gunjan, “Transfer learning based video summarization in wireless capsule endoscopy,” *International Journal of Information Technology*, vol. 14, no. 4, pp. 2183–2190, Jun. 2022, doi: 10.1007/s41870-022-00894-0.
  - [27] A. Biniiaz, R. A. Zoroofi, and M. R. Sohrabi, “Automatic reduction of wireless capsule endoscopy reviewing time based on factorization analysis,” *Biomedical Signal Processing and Control*, vol. 59, May 2020, doi: 10.1016/j.bspc.2020.101897.
  - [28] J. F. Montalbo, “WCE Curated Colon Disease Dataset Deep Learning,” 2022. [Online]. Available: <https://www.kaggle.com/Datasets/Francismon/Curated-Colon-Dataset-for-Deep-Learning>. Accessed 30 September 2022.

## BIOGRAPHIES OF AUTHORS



**Parminder Kaur**    is an assistant professor at the Department of Computer Science, Dr. Bhim Rao Ambedkar Government College Kaithal, Haryana, India. She is pursuing Ph.D. from the Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India. Her research interests are video processing, computer vision, and machine learning. She can be contacted at email: [kaur.parminder490@gmail.com](mailto:kaur.parminder490@gmail.com).



**Rakesh Kumar**    is a professor and chairman at the Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana India. He obtained his B.Sc. Degree, Master's degree – Gold Medalist (Master of Computer Applications) and Ph.D. (Computer Science & Applications) from Kurukshetra University, Kurukshetra. His research interests are in genetic algorithms, software testing, artificial intelligence, and networking. He has published research papers in many esteemed international journals and participated in national and international conferences. He has been a member of technical program committees and organizing committees of various International and National Conferences. He can be contacted at email: [rakeshkumar@kuk.ac.in](mailto:rakeshkumar@kuk.ac.in).