Wireless Capsule Endoscopy Video Summarization using Transfer Learning and Random Forests

Parminder Kaur[1], Dr. Rakesh Kumar[2]
Dept. of Computer Science, Dr. B.R. Ambedkar Government College, Kaithal, Haryana, India[1]
Dept. of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India[2]

***Abstract*—Wireless Capsule Endoscopy (WCE) is a diagnostic technique for identifying gastrointestinal diseases and abnormalities. Gastroenterologists face a considerable challenge when reviewing a lengthy video to identify a disease. The solution to this problem is generating an automated video summarization technique that generates the WCE Video summaries. This paper presents a Video Summarization technique that summarizes the WCE video. The proposed method uses transfer learning and a Random Forest classifier. Using a computationally light and pre-trained MobileNetV2 for feature extraction helped deliver results quickly. Managing small datasets and mitigating the overfitting risk was effectively addressed using Random Forest. The Random Forest's hyperparameters are optimized through the use of Bayesian optimization. The approach proposed has achieved an accuracy of 98.75% in disease prediction while significantly reducing the viewing time for the video summary. Furthermore, it has attained an average F-Score of 0.98, demonstrating its efficacy and reliability.**

Keywords—Bayesian optimization; capsule endoscopy; MobileNetV2; random forest classifier; transfer learning

I. Introduction

Wireless Capsule Endoscopy [1-3] is a technology used for performing the endoscopy of a patient to diagnose an illness. A pill-sized capsule camera captures the video of the gastrointestinal tract and assists the doctor in diagnosing any gastrointestinal abnormality. Capsule Endoscopy is effortless and does not interfere with a person's routine; due to this, it is getting more popular. However, specific challenges are associated with it - the battery may get low, and the capsule may get stuck in the gastrointestinal tract, and analyzing the long endoscopy video to identify the disease. An automated process for analyzing and generating the WCE video summary may save the doctor's time spent analyzing the video. Many researchers have utilized machine learning and deep learning methods to summarize WCE videos. Numerous studies aim to identify a particular ailment from a WCE video, like abnormal bleeding, tumor, polyp, or ulcer. However, if a problem-specific method is adopted, that proposed framework can only be applied to identifying a specific type of disease. One such approach was adopted in [4] for polyp detection having a lower false-positive rate. Since polyps are rounded or curved growths in the colon, the author also utilized the shape and texture features for polyp identification and localization. Similarly, [5] developed a classification method for polyp detection by using the textural characteristics of polyps and an improved bag of features. [6] Used Uniform Local Binary Pattern (LBP) to detect the polyps' texture. A method to detect Crohn's disease using a deep convolutional network is proposed in [7]. According to a study [8], two critical factors can aid in detecting tumors: color and textural features. To achieve this, the SVM utilizes the LBP operator's feature maps. This approach has proven to be quite effective in accurately identifying tumors. It achieved an accuracy of 92.4% in detecting tumors. A saliency map-based ulcer detection method was proposed in [9]. A multilevel approach was used for detecting saliency and identifying ulcers. In [10], a CNN-based approach for bleeding detection is proposed. The CNN developed has a low complexity because the input to the CNN is a single patch, and it outputs a segmented patch of the same size. An approach for detecting multiple bleeding detection was proposed in [11]. For detecting the small intestine lesions, [12] used AlexNet. The model achieved an accuracy of over 95.16%. [13] Proposed a technique for lesion detection using the high-level features extracted by ResNet50 [14] and InceptionV4. The lesion and non-lesion frames are classified by using the SVM classifier.

Specific video summarization approaches focus on keyframe extraction. A key frame is the most informative and relevant frame. [15], used a keyframe extraction approach for video summarization. The irrelevant frames are discarded in the first phase of image quality assessment. From the remaining frames, keyframes are extracted using low-level and deep features. Another keyframe extraction-based video summarization technique is proposed in [16]. Convolutional autoencoder is used to extract high-level features and shot boundary detection. A shot is part of a video having similar content. Motion profiles are used to extract keyframes from each shot. In [17], temporal segmentation of the video into shots was accomplished with the Prune Exact Linear Time (PELT) algorithm, and the high-level features were extracted by using pre-trained VGG19. A temporal segmentation of endoscopy video for detecting abnormal video segments was proposed by [18]. The abnormal shot covers any gastrointestinal abnormality. To identify the abnormality a Graph Convolutional Network (GCNN) was used. A keyframe selection strategy for polyp identification by utilizing the depth information of polyps is proposed in [19]. One approach for constructing keyframes of endoscopic videos uses pre-trained InceptionV3 to create feature maps of WCE images, which are then fed into a K-means algorithm implemented in [20].

A machine learning model's efficiency largely depends upon the amount of data used to train the model. If the training dataset is prominent, the model may learn adequately; otherwise, it may overfit, leading to inaccurate results with the test data. Nevertheless, in certain situations and domains, immense amounts of training data are unavailable; training a deep learning model for that problem becomes tedious. Transfer learning is the solution in such cases. Transfer learning is transferring knowledge from a learned model to a new one. However, both models should perform similar kinds of tasks. Applying transfer learning has several advantages over making a model learn from scratch. Training a model from scratch is time-consuming and requires tremendous training data. Moreover, there is no point in training a model from scratch if it performs a task similar to that performed by some other model.

The main objective of this research paper is to create a computationally efficient model that delivers precise results quickly. The proposed solution aims to overcome the following challenges:

1) One of the challenges in developing a machine-learning model for WCE is the need for adequate training data. With sufficient data, it is possible to train the model effectively. However, transfer learning is a solution to this problem. One can overcome the data shortage by leveraging pre-trained models with their weights and still successfully train a model.

2) Timely delivery of final results is crucial for the diagnostic procedure of endoscopy. Any delays are deemed unacceptable and can have serious consequences. MobileNetV2 [21] is a lightweight model in terms of computational requirements. Despite this, it can provide accurate solutions on time.

3) Machine learning models need a lot of computational resources. However, a model that requires significantly less computational resources and is so computationally light that it can be used in a mobile device is a favored solution.

This paper proposes a WCE video summarization technique that extracts the frames' deep features using the MobileNetV2 model. Further, the extracted features are provided as input to a Bayesian hyperparameter-optimized Random Forest Classifier. The Random Forest Classifier categorizes frames into classes based on features and then sorts frames by entropy values within each class. However, outliers may occasionally contain valuable information. The proposed approach examines frames from each predicted class to avoid discarding crucial information owing to outliers.

The rest of the paper is structured as follows: Section II details the methodology employed for WCE video summarization. Section III presents the experimental results and provides a thorough analysis. The final section summarizes the main conclusions drawn from this study.

II. Related Work

There have been several studies that aim for video summarization in WCE. The approaches for video summarization can be categorized into two categories: The first is a Generic approach that primarily focuses on Keyframe extraction or Shot Segmentation, and the second is Disease identification specific. Several approaches uses keyframe extraction techniques for video summarization. Keyframe extraction is a technique for video summarization that involves extracting the most informative frames from a long video that has many redundant frames. Researchers have explored various criteria and algorithms for selecting keyframes representing essential video information.

A. Generic Approaches

A general approach to video summarization in WCE is keyframe identification or Shot Segmentation. Keyframes are the frames that contain the most informative part of the video and can be considered as a representative frame of the entire video. On the other hand, Shot segmentation is dividing a long video into small shots. A shot is a part of the video that contains similar frames. A technique for keyframe extraction that uses depth maps not only for keyframe identification but also for localization of the polyps was proposed in [19]. The proposed approach detects the keyframes that contain the polyps. In addition to the depth information, the proposed technique utilizes the image moments and edge magnitudes to select the keyframes.

The author in [20] proposed a method for generating video summaries by utilizing the deep features extracted by using InceptionV3 and then using the K-Means clustering algorithm to group similar frames in one cluster. Frames from the clusters are selected to generate a final summary. A technique for keyframe extraction that first extracts the deep features of the frames using a Convolutional Autoencoder Neural Network (CANN) was used in [16]. The frames are then grouped into similar and dissimilar frames, representing shots of the WCE video. From each shot, keyframes are then selected using motion analysis.

B. Disease-Specific Approaches

Specific approaches focus on disease identification along with video summarization. One of the most common signs of gastrointestinal abnormality is bleeding. Several researchers worked towards identifying bleeding regions, polyps, ulcers, erosions, and tumors in the WCE Videos.

1) Bleeding detection: The author in [10] proposed a method for automatically detecting and segmenting bleeding regions by leveraging a CNN structure. The CNN utilized by the author is of low complexity.

2) Polyps: Most polyp detection approaches utilize the shape or texture features for polyp identification. However, [4] used an approach that considers both the context and shape of the polyp. Using a context-based reduces the chances of misclassifying some other polyp-like structure as a polyp. Whereas shape features help to capture the geometric information of polyps.

In [5] the author used a synthetically designed high-dimensional feature using Local Binary Patterns – Local, Uniform, and Complete, combined with the Histogram of oriented gradients (HOG). The high-dimensional descriptors provide the visual words as output, after they are fed as an input of the K-means clustering method. Finally, SVM and Fisher's linear discriminated analysis (FLDA) are used for polyp classification. The author in [6] combined the textural features and the Local Fractal Dimensions. The proposed method first detects the keypoints of the frames using SIFT followed by the textural feature extraction of the neighborhood of the keypoints. In the end, the classification is carried out by an SVM classifier.

3) Ulcers: A two-staged automated ulcer detection system was proposed in [9]. In the first stage, superpixels are identified. A superpixel is a group of pixels under some restriction of local image features such as color, intensity, or texture. Then, the saliency regions are identified based on texture and color. The color and texture-based saliency maps are fused together to create a better salient representation of the ulcers. In the second stage, ulcer classification is done using a Bag of Words Model. A CNN-based approach to detect small intestinal ulcers and erosion in WCE images was proposed in [12].

4) Crohn's disease: The authors in [7] developed a CNN (Convolutional Neural Network) to classify WCE images into two categories- normal images and the images that are likely to have evidence of Crohn's lesions.

5) Tumor: A tumor is a mass of abnormal cells. They can be cancerous in rare situations. Researchers are exploring this field to develop techniques for automated detection of tumors in the WCE images. The author in [8] exploit the image's color and textural features; later, a support vector machine (SVM) is utilized for feature selections for tumor detection.

Considering the different approaches for WCE video summarization it can be concluded that a generic approach that is able to identify the abnormal frames of a WCE video is a better approach for video summarization. An approach that identifies only a specific disease cannot find other diseases, there may be a case where a patient has multiple abnormalities. Therefore a generic method for WCE video summarization is a better approach for generating WCE video summaries. The video summaries generated by a generic method can further be evaluated by the gastroenterologist for pinpointing the particular ailment.

III. Methodology

The proposed method leverages the benefits of classification for video summarization. It generates a video summary of a long video. A video is an ordered set of frames. For processing a video, the first step is to generate the frames from the videos. CEV is the Capsule Endoscopy Video, which can be represented as

CEV= {f1, f2, f3…... fn}

Where, {f1, f2, f3……… fn } represents the ordered set of frames. Video summarization is the process of extracting the informative frames of the video. Let VS be the Video summary then, VS $\subset$CEV, and VS= {f1, f2, f3, f4….. fk} where k<n. We may assume this by renumbering the frames.

Fig. 1 depicts a block diagram of the proposed video summarization approach.

```
Algorithm 1
Step 1: Generating frames from the video.
CEV= {$f_{1,}$ $f_2$, $f_3$ . . . . . . . . . ,$f_n$}, where $f_1$,$f_2$,$f_3$…..$f_n$ are the video frames.
Step 2: Extract features from the frames by using pre-trained MobileNetV2.
For i=1 to n do
$y_i \leftarrow M(f_i)$ , $M(f_i)$ represents the MobileNetV2 used for feature extraction
Projecting each feature vector $y_i$ to embedding $\gamma$
$\gamma_i \leftarrow y_i$
$\gamma$= {$\gamma_{1,}$ $\gamma_{2\ldots}$ $\gamma_n$}
Step 3: The Random Forest algorithm processes the input as feature vectors extracted from frames in Step 2.
Step 4: Identifying the best parameters for the Random Forest using Bayesian Optimization.
Step 5: Testing the Model and generating Video Summaries.
```

C. MobileNetV 2

MobileNet architecture was introduced by Google in 2018. As the name indicates, MobileNet is a lightweight convolutional neural network developed for mobile or embedded devices. It works efficiently with devices that have limited computational resources. MobileNetV2's first layer is a depth-wise convolution that performs lightweight filtering by applying a single convolutional filter per input channel. The second layer is point-wise convolution, which computes linear combinations of the input channels to generate new features. The fewer parameters and matrix multiplications significantly contribute to reducing the complexity of MobileNetV2. Some key features of MobileNetV2 are:

- The depth-wise convolution layer applies a separate filter to each input channel, producing intermediate feature maps. A point-wise convolution followed them to combine these features linearly. This two-step convolution approach significantly reduces the number of parameters and computations compared to traditional convolutional layers.

- Inverted residual blocks of MobileNetv2 are capable of capturing more complex patterns. An inverted residual block consists of a bottleneck layer, which downsizes the number of input channels followed by a depth-wise separable convolution and a linear projection layer to upsize the number of channels back. Skip connections are also used to retain low-level features.

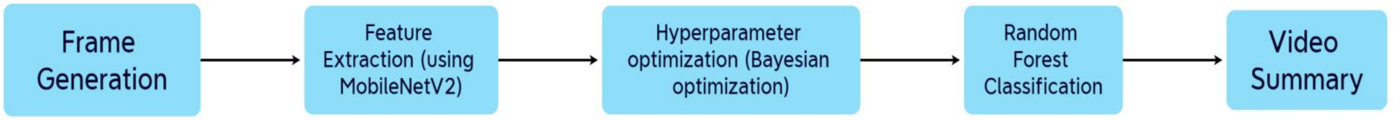
Frame Generation
Feature Extraction (using MobileNetV2)
Hyperparameter optimization (Bayesian optimization)
Random Forest Classification
Video Summary

Fig. 1. Block diagram of video summarization.

- There are two hyperparameters in MobileNetV2- width multiplier and resolution multiplier. The width multiplier reduces the number of channels in each layer. The resolution multiplier scales down the input image size. Due to the resolution multiplier, some spatial information is sacrificed. The hyperparameters reduce the model's complexity and the computational requirements.

D. Random Forest

Random Forest [22] is a supervised machine-learning algorithm developed by Leo Breiman. It uses an ensemble of multiple decision trees for generating predictions. Ensemble means combining multiple models. Thus, a Random Forest uses a collection of decision trees to make predictions rather than an individual decision tree. The random forest algorithm delivers a precise and cohesive output by aggregating these tree's outputs. Fig. 2 shows the working of a random forest classifier. The Random Forest classifier classifies an image in one out of the different output classes. Every decision tree casts a vote to which the input vector belongs. However, in the WCE video, multiple frames need to be classified; there may be a part of a video that contains redundant information, and a small part of the video contains abnormality. In other words, the outlier may contain the most informative information. To avoid missing any informative part of the endoscopic video, a technique for video summarization is adopted that incorporates a change in the voting module of the random forest. The proposed voting module calculates the entropy of each image. The total number of votes an output class received is sorted based on the entropy values, and the top 10% frames from each class are combined to generate the video summary.

Algorithm 2

Step 1: In the Random forest model a subset of features are randomly selected and decision trees are constructed from each sample.

```
for t=1 to T do
        D= {D1, D2, D3….. DT}, where
        D is the set of decision trees
        T is the total number of decision trees
        Dt ⊂ γ, where 1<t<T
```

Each decision tree is constructed from a subset of feature embedding γ.

Step 2: For each input image, each decision tree will generate an output and cast a vote to one of the output classes.

If L={L_1, L_2, L_3 ….L_k} and C= {C_1,C_2, C_3, ….. C_k} represents the set of labels of the Output class and C represents the count for each class label then,

```
for i=1 to n do
        for t=1 to T do
        Pit ←Dt(yi), P is the predicted Label for yi and Pit it
        takes value from L, and update the corresponding
        Label’s count
```

Step 3: Final output is considered based on Majority Voting for Classification. However, for every L_i, where1< i< k
The proposed voting module also calculates the entropy-based ranks for each vote cast for each class.

Step 4: The top 10% of the total votes (frames) that a class got are selected to generate a final video summary. And the final summary VS is generated for the video VCE.

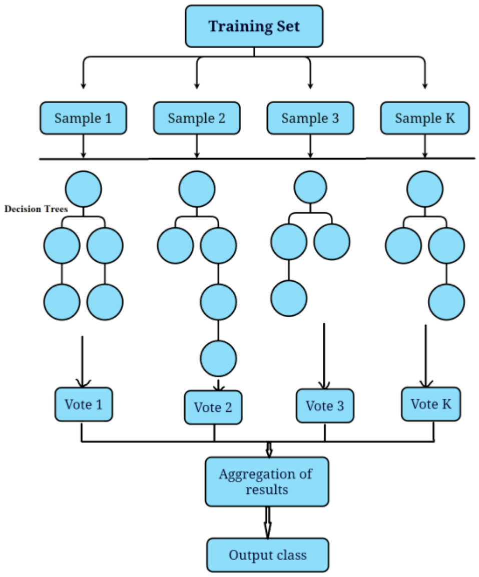

Fig. 2. Working of random forest classifier.

Each decision tree of the Random forest is constructed by using a different random sample from the training data every time, reducing the chances of overfitting. This tree construction task can be run on multiple CPU cores, reducing the training time. Random forests' voting and aggregation feature helps them effectively deal with missing and noisy data. A Random forest's most significant advantage is dealing with small sample sizes, high-dimensional feature space, and complex data structures.

E. Hyperparameter Tuning

Hyperparameter tuning is finding a hyperparameter setting for a machine-learning model to increase its accuracy. There are several techniques of hyperparameter optimization. Grid Search and Random Search are the most reliable techniques for hyperparameter tuning. Grid Search is an exhaustive technique; it considers each possible combination of hyperparameters to determine the optimum value. However, Random Search explores random values of the hyperparameters in a given search space. Both techniques are not adaptable; the results generated every time are independent of the previous outcomes. Random and Grid searches are significantly slow because of their exhaustive nature for search space exploration. Bayesian optimization [23] uses a probabilistic model to guide the search, which helps explore the hyperparameter space more intelligently and reduces the number of evaluations needed, making it one of the computationally efficient techniques. The optimized values of hyperparameters after Bayesian optimization are depicted in Table I.

IV. Experiments and Results

The experiments were implemented in Python, and the training and validation dataset was obtained from Kaggle

(WCE Curated Colon Disease Dataset Deep Learning) [24]. This data set consists of images a wireless capsule captures during endoscopy to diagnose abnormal conditions. It has three sets: training set, test set, and validation set. The WCE dataset has labeled images. The dataset has four labels: Normal, Ulcerative Colitis, Polyps, and Esophagitis.

TABLE I. HYPERPARAMETER OPTIMIZATION RESULTS

Hyperparameter	Optimized Value	Significance
n_estimators	174	Number of trees in the forest
min_samples_split	2	Minimum number of samples required to split an internal node.
max_depth	15	Maximum depth of the tree
random_state	42	Controlling the randomness

The performance of the Random Forest is presented in the confusion matrix of Fig. 3. The predicted labels are close to the true labels of the images. It represents a model with high accuracy. The model obtained an accuracy of 98.75% over 50 iterations.

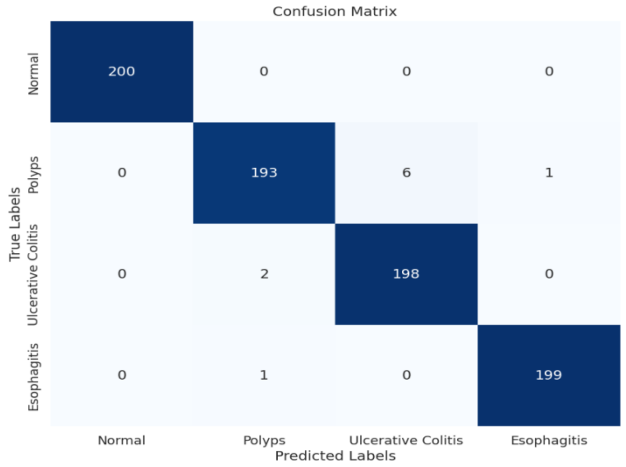

Fig. 3. Confusion matrix.

The model's performance is also compared over Precision, Recall, and F-score (Table II). F-Score is computed using (1). It is computed based on recall (2) and precision (3). Recall measures the proportion of actual positive instances that were correctly predicted as positive by the model.

$$\text{F\_Score} = \frac{(2*\text{Recall}*\text{Precision})}{(\text{Recall}+\text{Precision})} \quad (1)$$

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

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

TABLE II. CLASSIFICATION REPORT

	Precision	Recall	F-Score
Normal	1.00	1.00	1.00
Polyps	0.97	0.97	0.97
Ulcerative Colitis	0.97	0.97	0.97
Esophagitis	1.00	1.00	1.00

True Positives (TP) are the positives that are correctly predicted as positives. False Positives (FP) are the negatives 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 average value of the F-score is obtained as 0.985. The classification report indicates that the model is 100% precise in predicting the Normal and Esophagitis labels.

The proposed voting module of the Random Forest Classifier not only votes for a particular class but also calculates the entropy of each image. Fig. 4 shows the predicted frames of the output classes: Polyps, Ulcerative Colitis, and Esophagitis according to the decreasing entropy values. The Final summary is generated by combining the top 10% of the frames from every predicted class, as shown in Fig. 5. The selection of 10% of the total images of a particular class was experimentally determined. The contribution of each output class to generate the final summary ensures that outliers don’t get missed.

Fig. 5 shows a video summary of a patient suffering from Esophagitis. Although the disease identified is esophagitis the video summary generated has an abnormal bleeding part of the esophagus, which may otherwise have been missed.

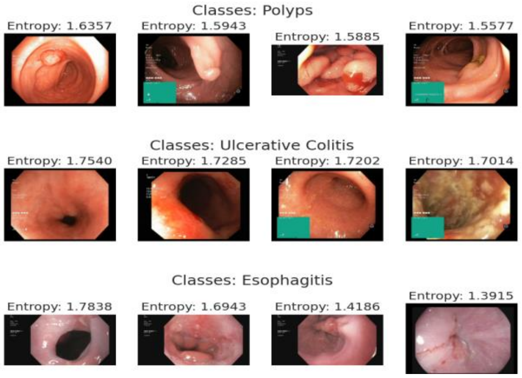

Fig. 4. Images from different output classes in decreasing order of entropies.

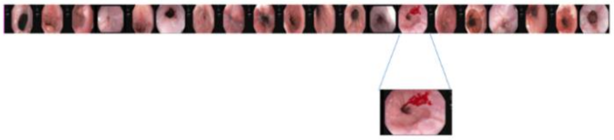

Fig. 5. Video Summary generated for an esophagitis patient.

V. CONCLUSION

This paper introduces a WCE Video Summarization technique that uses transfer learning, random forest, and an entropy-based ranking mechanism to select informative frames and generate the video summary. Using MobileNetv2 for feature extraction allowed for prompt and efficient results to be obtained with excellent computational efficiency. Moreover, employing a Random Forest reduces the chances of overfitting. Selecting the most informative frames from each predicted class prevents the exclusion of outliers with valuable content. The proposed approach obtained an accuracy of 98.75% in classifying the disease, and the Video summary generated by the model has a significantly reduced viewing time. In the future, a scalable WCE video summarization technique can be proposed that predicts the disease and maintains the temporal relation of the frames.

REFERENCES

[1] G. J. Iddan, G. Meron, A. Glukhovsky, and P. Swain, “Wireless capsule endoscopy,” *Nature*, vol. 405, no. 6785, p. 417, May 2000, doi: 10.1038/35013140.

[2] P. Swain, “Wireless capsule endoscopy,” Gut, vol. 52, no. 90004, pp. 48iv–4850, 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] N. Tajbakhsh, S. R. Gurudu and J. Liang, "Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information," in *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630-644, Feb. 2016, doi: 10.1109/TMI.2015.2487997.

[5] Y. Yuan, B. Li and M. Q. . -H. Meng, "Improved Bag of Feature for Automatic Polyp Detection in Wireless Capsule Endoscopy Images," in *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 529-535, April 2016, doi: 10.1109/TASE.2015.2395429.

[6] M. E. Ansari and S. Charfi, "Computer-aided system for Polyp detection in wireless capsule endoscopy images," 2017 International Conference on Wireless Networks and Mobile Communications (WINCOM), Rabat, Morocco, 2017, pp. 1-6, doi: 10.1109/WINCOM.2017.8238211.

[7] D. Marin-Santos, J. A. Contreras-Fernandez, I. Perez-Borrero, H. Pallares-Manrique, and M. E. Gegundez-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, Sep. 2022, doi: 10.1007/s10489-022-04146-3.

[8] 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.

[9] 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.

[10] M. Hajabdollahi, R. Esfandiarpoor, K. Najarian, N. Karimi, S. Samavi and S. M. Reza Soroushmehr, "Low Complexity CNN Structure for Automatic Bleeding Zone Detection in Wireless Capsule Endoscopy Imaging," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 7227-7230, doi: 10.1109/EMBC.2019.8857751.

[11] O. Bchir, M. M. B. Ismail, and N. Alzahrani, “Multiple bleeding detection in wireless capsule endoscopy,” Signal, Image and Video Processing, vol. 13, no. 1, pp. 121–126, Jul. 2018, doi: 10.1007/s11760-018-1336-3.M. L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, Jan. 2001, doi: 10.1023/a:1010933404324.

[12] 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, p. 165001, Aug. 2018, doi: 10.1088/1361-6560/aad51c

[13] A. Caroppo, P. Siciliano, and A. Leone, “An expert system for lesion detection in wireless capsule endoscopy using transfer learning,” *Procedia Computer Science*, vol. 219, pp. 1136–1144, Jan. 2023, doi: 10.1016/j.procs.2023.01.394.

[14] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[15] A. Biniaz, 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, p.101897, May 2020, doi:10.1016/j.bspc.2020.101897.

[16] B. Sushma and P. Aparna, "Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion Analysis," in *IEEE Access*, vol. 9, pp. 13691-13703, 2021, doi: 10.1109/ACCESS.2020.3044759.

[17] S. Adewole et al., “Unsupervised shot boundary detection for temporal segmentation of long capsule endoscopy videos.,” arXiv (Cornell University), Oct. 2021, [Online]. Available: http://export.arxiv.org/abs/2110.09067

[18] S. Adewole et al., “Graph Convolutional Neural Network For Weakly Supervised Abnormality Localization In Long Capsule Endoscopy Videos,” 2021 IEEE International Conference on Big Data (Big Data), Dec. 2021, doi: 10.1109/bigdata52589.2021.9671281.

[19] P. Sasmal, A. Paul, M. K. Bhuyan, Y. Iwahori, and K. Kasugai, "Extraction of Keyframes From Endoscopic Videos by using Depth Information," IEEE Access, vol. 9, pp. 153004–153011, Jan. 2021, doi: 10.1109/access.2021.3126835.

[20] V. G. 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, Feb. 2022, doi: 10.1007/s41870-022-00894-0.

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Salt Lake City, UT, USA, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474

[22] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, Jan. 2001, doi: 10.1023/a:1010933404324.

[23] F. Hutter, H. H. Hoos, and K. Leyton-Brown, “Sequential Model-Based optimization for general algorithm configuration,” in Springer eBooks, 2011, pp. 507–523. doi: 10.1007/978-3-642-25566-3_40.

[24] “WCE Curated Colon Disease Dataset Deep Learning,” Kaggle, Apr. 15, 2022. https://www.kaggle.com/datasets/francismon/curated-colon-dataset-for-deep-learning.