Colon Cancer Detection Based on Deep Learning Using Image Processing

C Krishna Kumari

Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *It is very important to make an objective evaluation of colorectal cancer histological images. Current approaches are generally based on the use of different combinations of textual features and classifiers to assess the classification performance. For virtually every patient with colorectal cancer (CRC), hematoxylineosin (HE) stained tissue slides are available. These images contain quantitative information, which is not routinely used to objectively extract prognostic biomarkers. In the present study, we investigated whether deep convolutional neural networks (CNNs) can extract prognosticators directly from these widely available images. Classification is still challenging. In this proposed method, we proposed the best classification methodology based on the selected CNN networks used as Densenet 201. Then, we used deep learning technology to distinguish between healthy and diseased large intestine tissues. The results showed that the accuracy of the recognition of histopathological images was significantly better than the conventional methods.*

I. INTRODUCTION

The colon is the part of the digestive system, and its function is to absorb fluid and waste products of the body's processes. Cancer is a disease in which abnormal cells divide without control and destroy body tissue. Colon cancer usually begins with small, non-cancerous clumps of cells called polyps that form in the large intestine. Over time, these polyps can develop into colon cancer. According to the American Cancer Society, more than 50,000 Americans die from colon cancer each year. This amount makes colorectal cancer the third leading cause of cancer-related deaths in the US. Meanwhile, in Indonesia, Global Cancer Observatory estimates 30.017 cases of colon cancer in 2018, making 8.6% of Indonesia's total cancer cases. Even though Indonesia's number is relatively small compared to the United States, a preventive measure must be considered. The massive leap in technology for the past years inevitably affects the medical field. The growth in medical data collection opens a new possibility for researchers to develop their methods to help medical diagnosis more sophisticated. Machine learning is one kind of approach that is widely used for medical diagnosis.

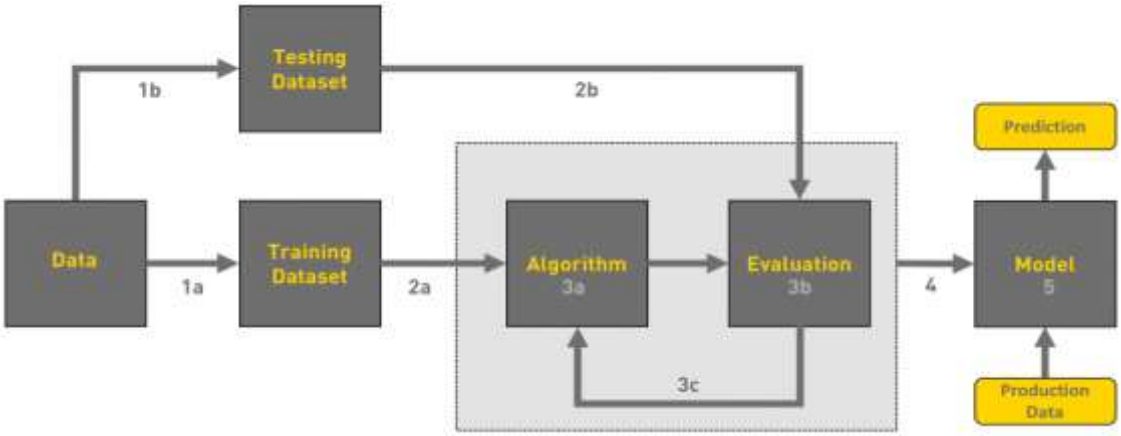
Artificial intelligence (AI) is the ability of a computer program or a machine to think and learn. It is also a field of study which tries to make computers "smart". As machines become increasingly capable, mental facilities once thought to require intelligence are removed from the definition. AI is an area of computer sciences that emphasizes the creation of intelligent machines that work and reacts like humans. Some of the activities computers with artificial intelligence are designed for include: Face recognition, Learning, Planning, Decision making etc., Artificial intelligence is the use of computer science programming to imitate human thought and action by analysing data and surroundings, solving or anticipating problems and learning or self-teaching to adapt to a variety of tasks.

1.1. Deep Learning

Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised. Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance.

In deep learning, each level learns to transform its input data into a slightly more abstract and composite representation. In an image recognition application, the raw input may be a matrix of pixels; the first representational layer may abstract the pixels and encode edges; the second layer may compose and encode arrangements of edges; the third layer may encode a nose and eyes; and the fourth layer may recognize that the image contains a face. Importantly, a deep learning process can learn which features to optimally place in which level on its own. This does not eliminate the need for hand-tuning; for example, varying numbers of layers and layer sizes can provide different degrees of abstraction.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Deep learning has changed our way of thinking about the problem. The below block diagram explains the working of Deep Learning algorithm:



1.1.1. Features of Deep Learning:

- Deep learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Deep learning is much similar to data mining as it also deals with the huge amount of the data.

1.1.2. Classification of Machine Learning:

At a broad level, deep learning can be classified into two types:

1. Supervised learning
2. Unsupervised learning

Supervised Learning

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.

Supervised learning can be grouped further in two categories of algorithms:

- **Classification**
- **Regression**

Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data.

It can be further classifieds into two categories of algorithms:

- **Clustering**
- **Association**

1.2. CNN:

CNN is an example of a Deep Learning algorithm that takes an input image and assigns priority to different aspects of the image, allowing it to distinguish one image from another based on its features. In this system, two convolutional layers in the CNN model are used where each convolutional layer used convolutional 2D. In both convolutional 2D layers, 'Relu activation' is utilized. For complete connectivity, two Dense Layers are used. 'Relu activation' for the first dense layer and 'Sigmoid activation' for the second dense layer is used. Aside from these layers, there are several hidden layers, as well as an input layer. In this study, two pooling layers: Max Pooling 2D and Average Pooling 2D, are implemented. Finally, for the classification of image data Densenet 201 classifier is used.

1.3. Dense Convolutional Network (DenseNet)

Dense Convolutional Network (DenseNet) is connecting each layer to every other layer in a feed-forward fashion. They alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters.

DenseNet works on the idea that convolutional networks can be substantially deeper, more accurate, and efficient to train if they have shorter connections between layers close to the input and those close to the output. The figure below is from the original paper which gives a nice visualization of scaling.

Here, we will implement the Deep Learning model to create the network based on the Densenet 201 Network. DenseNet-201 is a convolutional neural network that is 201 layers deep. You can load a pretrained version of the network trained on more than a million images from the ImageNet database of the system. DenseNet was developed specifically to improve the declined accuracy caused by the vanishing gradient in high-level neural networks.

II. LITERATURE REVIEW

[1] TITLE: Topographical reconstructions from monocular optical colonoscopy images via deep learning

AUTHOR NAME: Faisal Mahmood; Nicholas J. Durr

DESCRIPTION: Colorectal cancer is a leading cause of cancer deaths worldwide, but mortality can be mitigated by the detection and removal of premalignant lesions. Unfortunately, conventional 2D optical colonoscopy does not capture topographical information of the surface of the mucosa and thus has a high lesion miss rate. In this short paper, we use a joint deep convolutional neural network-conditional random field (CRF) framework for depth estimation from monocular colonoscopy images. Unlike previous approaches, this method does not make any geometric assumptions. The estimated depth is used to reconstruct the topography of the surface of the colon. Using digitally generated synthetic endoscopy data and CT-Phantom data, with corresponding ground truth depths, we train the unary and pairwise potential functions of a conditional random filed in a joint network. Results show that this approach can estimate depths for test data with a84% accuracy. We show that estimated depth maps can be used for reconstructing the topography of the mucosa from real colonoscopy images. This topographical information can be used for improving learning-based algorithms for detection, segmentation and identification of lesions.

[2] TITLE: Deep Learning Based Colorectal Cancer (CRC) Tumors Prediction

AUTHOR NAME: Rahul Deb Mohalder; Kamrul Hasan Talukder

DESCRIPTION: Cancer detection and prediction using computer assisted systems has become the most leading research area in recent times. It has a big demand in the medical sector for identifying not only cancer but also any diseases detected and predicted from pathological data or images. Colorectal Cancer or Colon Cancer (CRC) detection is also one of them. Because CRC has become a global health issue day by day. In this paper we used a dataset of 10,000 histopathological images with the same dimension of colonic tissue. We used ensemble methods and classifiers for classifying images. We

obtained the best accuracy 99% from XGBoost classifier and from others were 98%, 97%, 96%, 92%, 92% and 89% which exactly classifying 523 colon adenocarcinoma images and 477 benign colonic tissue images from 1,000 histopathological images.

[3] TITLE: https://ieeexplore.ieee.org/abstract/document/1017309/An improved automatic system for aiding the detection of colon polyps using deep learning

AUTHOR NAME: LishanCai;ReginaBeets-Tan;Sean Benson

DESCRIPTION: Colorectal cancer is responsible for the most cancer deaths after lung cancer. It has been well-established that early detection and removal of polyps can prevent colorectal cancer. It is therefore essential that automated polyp detection has the highest sensitivity and precision possible in order to detect the most cases and prevent unnecessary treatment. We present a deep learning model based on YOLOv3 that was trained to detect polyps. Training made use of the 39308 images of 78 polyps and 393 completely healthy images from the SUN database. The model was subsequently validated using both the public CVC-clinic and ETIS-Larib datasets containing both standard defintion (SD) and high definition (HD) images. The per-image polyp detection sensitivity(precision) was calculated as 91.5(96.6)% and 86.5(94.2)% for the CVC-clinic and Etis-Larib datasets, respectively. These results represent the best-known performance in the validation datasets in comparison with the results of a recent review.

[**4] TITLE:** Analysis of Deep Feature Extraction for Colorectal Cancer Detection

AUTHOR NAME: Devvi Sarwinda;

DESCRIPTION: Colorectal cancer is the most common cancer worldwide in the third. Detection of colon cancer is an essential task for the histopathologist as they have to analyze the morphology of the images at different magnifications. In this study, we classified benign and adenocarcinoma using 10000 images of benign colon tissue. We proposed a feature extraction method by the deep convolutional neural network. First, we learn the features of data from ResNet-50 and DenseNet-121. Then, we conduct colon cancer classification by popular classifiers such as SVM, Random Forest, K-Nearest Neighbor, and XGBoost. We evaluated our models on two kinds of testing data (25% and 15% of the whole dataset). In this research, the data was conducted on the Kaggle colon tissue dataset. The experimental results indicate that the extraction of features in DenseNet-121 based architecture leads to higher accuracy, sensitivity, and specificity of ResNet-50 architecture for all classifiers. DenseNet-121 gets about 98.53% and 98.63% with KNN classifier for accuracy and sensitivity, respectively.

[5] TITLE: Gland Segmentation in Histopathological Images by Deep Neural Network

AUTHOR NAME: SafiyehRezaei;AliEmami;NaderKarimi;ShadrokhSamavi

DESCRIPTION: Histology method is vital in the diagnosis and prognosis of cancers and many other diseases. For the analysis of histopathological images, we need to detect and segment all gland structures. These images are very challenging, and the task of segmentation is even challenging for specialists. Segmentation of glands determines the grade of cancer such as colon, breast, and prostate. Given that deep neural networks have achieved high performance in medical images, we propose a method based on the LinkNet network for gland segmentation. We found the effects of using different loss functions. By using Warwick-Qu dataset, which contains two test sets and one train set, we show that our approach is comparable to state-of-the-art methods. Finally, it is shown that enhancing the gland edges and the use of hematoxylin components can improve the performance of the proposed model.

III. PROBLEM STATEMENT

- For colorectal cancer, one of the most prevalent tumor types, there are in fact no published results on multiclass texture separation.
- Although histological images typically contain more than two tissue types, only few studies have addressed the multi-class problem.
- Machine learning classifiers are used for classify the input data given to the system.
- Classifier such as nearest neighbor and Ensemble of decision trees are used for classification.

Disadvantage

- It will take too much resource for the processing.

- Take too much processing time.
- Accuracy is less for the prediction in testing images

IV. DEVELOPMENT PROCESS

4.1 Requirement Analysis and Specifications

The requirement engineering process consists of feasibility study, requirements elicitation and analysis, requirements specification, requirements validation and requirements management. Requirements elicitation and analysis is an iterative process that can be represented as a spiral of activities, namely requirements discovery, requirements classification and organisation, requirement negotiation and requirements documentation.

4.1.1 Input Requirement and Output Requirements

Input Design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

Objectives

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2.It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus, the objective of input design is to create an input layout that is easy to follow

Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2.Select methods for presenting information.

3.Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the

- ❖ Future.
- ❖ Signal important events, opportunities, problems, or warnings.
- ❖ Trigger an action.
- ❖ Confirm an action.

V. PROPOSED METHOD

- Deep learning has gained enormous popularity in scientific computing due to CNN, and its algorithms are widely used by industries to solve complex problems.
- In this proposed method, the project demonstrated the Deep Learning model based on Densenet 201 model of the system.
- The block diagram illustrates the recognition process, which will train the data and find the accuracy of the training.
- We divide the dataset into 70% of data for training and 30% of data for testing.
- The last stage of our architectural parameters involved classifying the histological images through each CNN model's neural network architecture by training the model to identify the types from different disease classes.

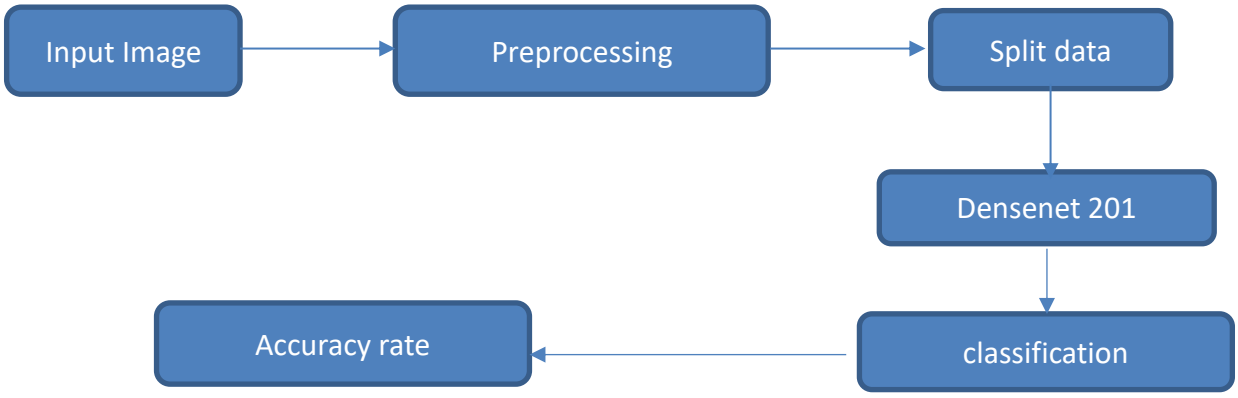
5.1 Advantages

- An improved version of deep learning parameters was proposed in this proposed method to improve the accuracy of classification.
- Based on the experimental results, our method was superior to the techniques described in the conventional methods.

5.2 Algorithm

CNN:

CNN is an example of a Deep Learning algorithm that takes an input image and assigns priority to different aspects of the image, allowing it to distinguish one image from another based on its features. In this system, two convolutional layers in the CNN model are used where each convolutional layer used convolutional 2D. In both convolutional 2D layers, 'Relu activation' is utilized. For complete connectivity, two Dense Layers are used. 'Relu activation' for the first dense layer and 'Sigmoid activation' for the second dense layer is used. Aside from these layers, there are several hidden layers, as well as an input layer. In this study, two pooling layers: Max Pooling 2D and Average Pooling 2D, are implemented. Finally, for the classification of image data Densenet 201 classifier is used.



5.3 Modules Used

1. Datasets Collection:

A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question.

A Collection of data is called datasets. Here, the datasets collection is based on the two types of Colon Cancer classification. Here, an input data can be obtained in the image format to classify the cancer based on the user chosen image of the system.

2. Preprocessing:

Pre-processing routines prepare the data for analysis. Before we start the actual processing, the data has to be pre-processed to remove the detector effects. Preprocessing is the most important aspect of data processing. Hence, data filtering, data ordering, data editing and noise modeling play an important role in any data preprocessing.

A Pre-Processing is one of the techniques is used to reduce the noise in the image format of the system. Smoothing and de-trending are processes for removing noise and linear trends from data, while scaling changes the bounds of the data. Grouping and binning methods are techniques that identify relationships among the data variables. Here, the preprocessing techniques will be used in the Image Enhancement to enhance the image to show the results based on the High-Resolution pattern of the system. Image Enhancement pattern will be obtained to resize the image based on the user convenient of the system.

3. Network Creation:

Here, we will implement the Deep Learning model to create the network based on the Densenet 201 Network. DenseNet-201 is a convolutional neural network that is 201 layers deep. You can load a pretrained version of the network trained on more than a million images from the ImageNet database of the system. DenseNet was developed specifically to improve the declined accuracy caused by the vanishing gradient in high-level neural networks.

4. Classification:

Classification neural networks used for feature categorization are very similar to fault-diagnosis networks, except that they only allow one output response for any input pattern, instead of allowing multiple faults to occur for a given set of operating conditions. Classification is a process related to categorization, the process in which ideas and objects are recognized, differentiated and understood the system. Classification is a term used both about the process to classify the condition of the patient. Among the most important contributing disciplines are philosophy, biology, knowledge organization, psychology, statistics and mathematics. A classification process can be obtained to training the network of the system. Here, classification techniques can be obtained by using Deep Learning model to classify the result of the model.

VI. FUTURE ENHANCEMENT

In our future work, we can implement the concept of colon cancer detection disease using image processing in deep learning techniques. In this project, we are processing in software. In future, we will implement in hardware side.

VII. CONCLUSION

In our Project, Interpretation of complex images by deep CNNs is presently transforming many domains in medical imaging, but clinical translation of this technology is still in its infancy. One reason for this delay is that CNNs per se need huge annotated training data sets that are not readily available in the context of histopathology. Another reason is that neural network–based risk assessment needs to be validated in clinically characterized validation cohorts. In the present study, we addressed both of these difficulties: we assembled a large data set of This dataset contains 25,000 histopathological images with 5 classes. All images are 768 x 768 pixels in size and are in jpeg file format. The images were generated from an original sample of HIPAA compliant and validated sources, consisting of 750 total images of lung tissue (250 benign lung tissue, 250 lung adenocarcinomas, and 250 lung squamous cell carcinomas) and 500 total images of colon tissue (250 benign colon tissue and 250 colon adenocarcinomas) and augmented to 25,000 using the Augmentor package. publicly available data sets. Finally, to classify the data's based on the Colon cancer databases.

REFERENCES

[1] Sánchez-Peralta, L.F.; Bote-Curiel, L.; Picón, A.; Sánchez-Margallo, F.M.; Pagador, J.B. Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. Artif. Intell. Med. 2020.

[2] F. M. Javed Mehedi Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi and S. Shultana, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020.

[3] P. Ghosh, F. M. Javed Mehedi Shamrat, S. Shultana, S. Afrin, A. A. Anjum and A. A. Khan, "Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm," 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Bangkok, Thailand, 2020.

[4] Pronab Ghosh, Sami Azam, Mirjam Jonkman, Asif Karim, F.M. Javed Mehedi Shamrat, Eva Ignatious, Shahana Shultana, Abhijit Reddy Beeravolu, Friso De Boer, "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques," in IEEE Access, doi: 10.1109/ACCESS.2021.3053759.

[5] Toraman, S.; Girgin, M.; Üstündağ, B.; Türkoğlu, İ. Classification of the likelihood of colon cancer with machine learning techniques using FTIR signals obtained from plasma. Turk. J. Electr. Eng. Comput. Sci. 2019.

[6] Jiao, Liping & Chen, Qi & Li, Shuyu & Xu, Yan. (2013). Colon Cancer Detection Using Whole Slide Histopathological Images. IFMBE Proceedings.

[7] S. Rathore, M. Hussain, and A. Khan, "Automated colon cancer detection using hybrid of novel geometric features and some traditional features," Comput. Biol. Med., 2015.

[8] Yuan, Z.; IzadyYazdanabadi, M.; Mokkapati, D.; Panvalkar, R.; Shin, J.Y.; Tajbakhsh, N.; Gurudu, S.; Liang, J. Automatic polyp detection in colonoscopy videos. Med. Imaging 2017 Image Process. 2017.

[9] Babu, T.; Gupta, D.; Singh, T.; Hameed, S. Colon Cancer Prediction on Different Magnified Colon Biopsy Images. In Proceedings of the 10th International Conference on Advanced Computing (ICoAC), Chennai, India, 13–15 December 2018.

[10] Mo, X.; Tao, K.; Wang, Q.; Wang, G. An Efficient Approach for Polyps Detection in Endoscopic Videos Based on Faster R-CNN. In Proceedings of the International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018.