COLORECTAL CANCER DETECTION BASED ON CONVOLUTIONAL NEURAL NETWORKS (CNN) AND

RANKING ALGORITHM

**<Boggada Mounika> <Y. eesha sai sri> <Pancheti divija reddy> <K. Sangeetha>**

**Student, Dept of Computer Science Engineering, PEC College, Tamil Nadu, India**

**Student, Dept of Computer Science Engineering, PEC College, Tamil Nadu, India**

**Student, Dept of Computer Science Engineering, PEC College, Tamil Nadu, India**

**Associate Professor, Dept of Computer Science Engineering, PEC College, Tamil Nadu**

**ABSTRACT :**

Now-a-days, with the development of targeted therapies, many treatments are based on molecular studies, which require sampling tumor tissue from paraffin blocks for sequencing. An automated solution could potentially reduce the workload of the pathologists by acting as a screening device and may reduce the subjectivity in diagnosis. In tissue-based diagnostics, most of the work still needs to be done manually by a pathologist using a microscope to examine stained slides. The foundation of such tasks is to accurately distinguish cancer/malignant cells from normal/benign cells. However, the determination of tumor content is poorly reproducible with significant variation. As the size of tumor regions can be very small, pathologists are often required to use high magnification for detecting tumor cells. This requirement significantly increases the workload for pathologists. As digital pathology datasets have become publicly available and have opened up the possibility of evaluating the feasibility of applying deep learning techniques to improving the efficiency and quality of histologic diagnosis. In this project we introduce an application to detect Colorectal cancer based on the Convolutional Neural Network and Ranking algorithm.

**INTRODUCTION :**

In the modern era, cancer is the most spreading complex disease. Identifying cancer without biopsy at an early stage is further imperative. Also, taking a biopsy is not good for health also. In general, cancer has been caused by hereditary instability and accumulation of multiple molecular alterations. It is also caused by cellular genes abnormal activation that controls cell growth or cell mitosis. Colorectal cancer is cancer from uncontrolled cell growth in the colon or rectum. This was the third most commonly diagnosed cancer in the world. Colorectal cancer is also known as colon cancer, bowel cancer or colorectal adenocarcinoma. The main negative aspect of cancer is its diagnosis and treatment too late. Due to this problem, cancer has overtaken heart disease as the leading cause of death for any age on. Therefore, early detection of cancer is important.

**PROBLEM DEFINITION**

Colorectal cancer is also known as colon cancer, bowel cancer or colorectal adenocarcinoma. The main negative aspect of cancer is its diagnosis and treatment too late. Due to this problem, cancer has overtaken heart disease as the leading causeof death for any age on. Therefore, earlydetection of cancer is important. With the development of targeted therapies, many treatments are based on molecular studies, which require sampling tumor tissue from paraffin blocks for sequencing. An automated solution could potentially reduce the workload of pathologists by acting as a screening device and may reduce the subjectivity in diagnosis. As datasets have become publicly available and have opened up the possibility of evaluating the feasibility of applying deep learning techniques to improving the efficiency and quality of histologic diagnosis. In this project we introduce an application to detect Rectal cancer based on Convolutional Neural Network and Ranking algorithm

**LITERATURE SURVEY:**

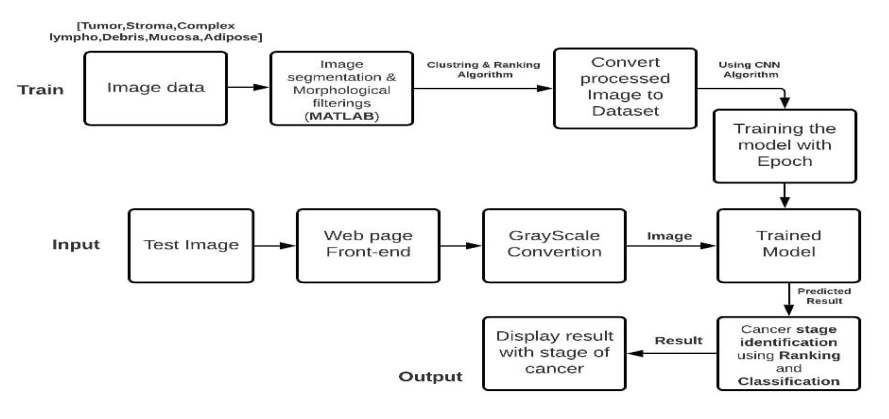
In [1] theColorectal Cancer Detection was done in MRI Images Using Image Processing Techniques. The key problems in thetreatment of cancer is the early detection of the disease. Often, cancer is detected in its later stages, when it hascompromised the function of one or more vital organ systems and is widespread throughout the body. Methodsfor the early detection of cancer are of utmost importance and are an active area of current research. Magneticresonance imaging (MRI) established itself as the primary method for detection and staging in patients withcolorectal cancer. In this paper, MRI images of colorectal cancer are used to detect the area and mean values oftumor area and distance from tumor area to other parts for staging cancer. This research paper describesalgorithms for pre-processing, clustering and segmentation of MRI images. The implementation of thisclustering, segmentation and pre-processing is done with MATLAB 2015 (a). By using this proposed methodology,the cancer is detected in its early beginning stage.

In paper[2], itaddressesthe current problem in medical image processing, the detection of colorectal cancer from colonoscopy videos. According to worldwide cancer statistics, colorectal cancer is one of the most common cancers. The process of screening and the removal of pre-cancerous cells from the large intestine is a crucial task to date. The traditional manual process is dependent on the expertise of the medical practitioner. In this paper, a two-stage classification is 4 proposed to detect colorectal cancer. In the first stage, frames of colonoscopy video are extracted and are rated as significant if it contains a polyp, and these results are then aggregated in a second stage to come to an overall decision concerning the final classification of that frame to be neoplastic and nonneoplastic. In doing so, a comparative study is being made by considering the applicability of deep learning to perform this two-stage classification. The CNN models namely VGG16, VGG19, Inception V3, Xception, GoogLeNet, ResNet50, ResNet100, DenseNet, NASNetMobile, MobilenetV2, InceptionResNetV2 and fine-tuned version of each model is evaluated. It is observed that the VGG19 model is the best deep learning method for colonoscopy image diagnosis.

In [3] an automated system for grading of colorectal cancer using image processing method. Almost, half a million people die every year due to colon cancer. Histopathological tissue analysis is a common method for its detection, which needs an expert pathologist. Screening for this cancer is effective for prevention as well as early detection. The method proposed segment the glands automatically by using intensity-based thresholding and organizational properties for classification. In existing literature, the majority of studies based on gland segmentation in healthy or benign samples, but rarely on intermediate or highgrade cancer. Unlike most of the existing methods this system is fully automated and grades the images as benign healthy, benign adenomatous, moderately differentiated malignant and poorly differentiated malignant. The proposed method achieves overall accuracy of 81% when tested on 165 histology images.

# PROPOSED SYSTEM :

In our proposed system we use CNN and LSTM simulator models using python to identify colorectal cancer affected tumor tissue. Tissues are classified into 7 types Tumor, Stroma, Lympho, Adipose, Complex, Debris, Mucosa. These seven types are used to identify the stage. According to a datasheet from cancer.net, the stage of cancer is classified as three types Tumor(T), Node(N), Metastasis(M). Stages are classified based on similarity in multiple tissue types. Combination of these 7 types [Tumor, Stroma, Lympho, Adipose, Complex, Debris, Mucosa] gives us the stage. This is achieved using CNN and LSTM where CNN trains the model to predict the tumor similarity and LSTM helps in identifying.



**Fig 3: Flow diagram**

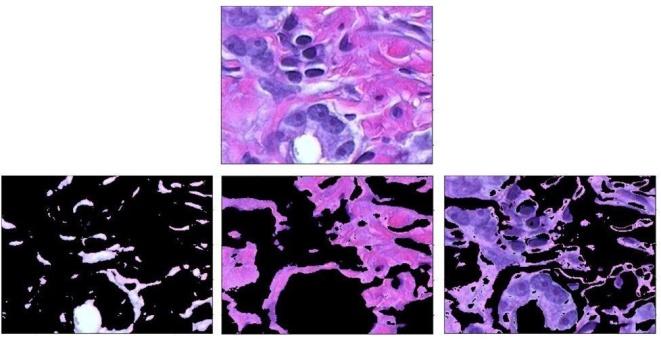
**EXISTING SYSTEM :**

The key problem in the treatment of cancer is the early detection of the disease. Often, cancer is detected in its later stages, when it has compromised the function of one or more vital organ systems and is widespread throughout the body. Methods For the early detection of cancer are of utmost importance and are an active area of current research. Magnetic Resonance imaging (MRI) established itself as the primary method for detection and staging in patients with colorectal cancer. In this paper, MRI images of colorectal cancer are used to detect the area and mean values of tumor area and distance from tumor area to other parts for staging cancer. This research paper describes algorithms for pre-processing, clustering and segmentation of MRI images. The implementation of this clustering, segmentation and pre-processing is done with MATLAB 2015 (a). By using this proposed methodology, the cancer is detected in its early beginning stage.

**CLUSTERING :**

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects

in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of [exploratory data analysis](https://en.wikipedia.org/wiki/Exploratory_data_analysis), and a common technique for [statistical](https://en.wikipedia.org/wiki/Statistics) [data analysis](https://en.wikipedia.org/wiki/Data_analysis), used in many fields, including [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), [image analysis](https://en.wikipedia.org/wiki/Image_analysis), [information](https://en.wikipedia.org/wiki/Information_retrieval) [retrieval](https://en.wikipedia.org/wiki/Information_retrieval), [bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics), [data compression, computer graphics an](https://en.wikipedia.org/wiki/Data_compression)d [machine](https://en.wikipedia.org/wiki/Machine_learning) [learning](https://en.wikipedia.org/wiki/Machine_learning). Cluster analysis itself is not one specific [algorithm](https://en.wikipedia.org/wiki/Algorithm), but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small [distances](https://en.wikipedia.org/wiki/Distance_function) between cluster members, dense areas of the data space, intervals or particular [statistical](https://en.wikipedia.org/wiki/Statistical_distribution) [distributions](https://en.wikipedia.org/wiki/Statistical_distribution).



**Fig 1: Image Segmentation Using Clustering**

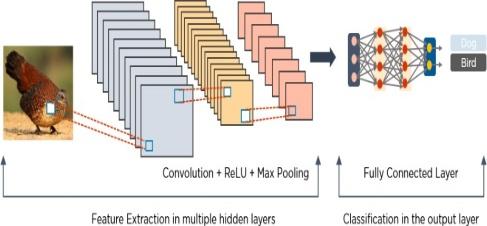
# RANKING ALGORITHM :

Randomized On-Line Matching, a representative of a class of algorithms, is a sequential algorithm that exploits a randomized efficient on-line [matching](https://www.sciencedirect.com/topics/computer-science/matching-algorithm) [algorithm](https://www.sciencedirect.com/topics/computer-science/matching-algorithm) that calculates maximal matchings in bipartite graphs, named the Ranking algorithm , as its basis. The Ranking algorithm considers that the

nodes of one part of the [bipartite graph](https://www.sciencedirect.com/topics/computer-science/bipartite-graphs) arrive on-line, that is, one after the other, and calculate a matching in an on-line fashion. Specifically, the algorithm calculates a [random permutation](https://www.sciencedirect.com/topics/computer-science/random-permutation) of the nodes in one part of the graph and then considers on-line arrival of the nodes in the other part; each incoming node of the second graph part is matched with the first appropriate node in the permutation of the first graph part. Ranking calculates a maximal matching, as has been proved.

**CNN ALGORITHM :**

In [deep learning](https://en.wikipedia.org/wiki/Deep_learning), a convolutional neural network (CNN, or ConvNet) is a class of [deep neural networks](https://en.wikipedia.org/wiki/Deep_neural_network), most commonly applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on the shared-weight architecture of the convolution kernels that shift over input features and provide translation [equivariant](https://en.wikipedia.org/wiki/Equivariant_map) responses. Counter-intuitively, most convolutional neural networks are only [equivariant](https://en.wikipedia.org/wiki/Equivariant_map), as opposed to [invariant](https://en.wikipedia.org/wiki/Translation_invariant), to translation. They have applications in [image and video](https://en.wikipedia.org/wiki/Computer_vision) [recognition](https://en.wikipedia.org/wiki/Computer_vision), [recommender systems](https://en.wikipedia.org/wiki/Recommender_system), [image classification](https://en.wikipedia.org/wiki/Image_classification), [Image segmentation](https://en.wikipedia.org/wiki/Image_segmentation), [medical image analysis](https://en.wikipedia.org/wiki/Medical_image_computing), [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing), [brain-computer interfaces](https://en.wikipedia.org/wiki/Brain%E2%80%93computer_interface), and financial [time series](https://en.wikipedia.org/wiki/Time_series).



**Fig 2: Images processed via CNN**

CNNs are [regularized](https://en.wikipedia.org/wiki/Regularization_(mathematics)) versions of [multilayer perceptron](https://en.wikipedia.org/wiki/Multilayer_perceptron). Multilayer perceptrons usually mean fully connected networks, that is, each neuron in one [layer](https://en.wikipedia.org/wiki/Layer_(deep_learning)) is connected to all neurons in the next [layer](https://en.wikipedia.org/wiki/Layer_(deep_learning)). The "full connectivity" of these networks makes them prone to [overfitting](https://en.wikipedia.org/wiki/Overfitting) data. Typical ways of regularization, or preventing overfitting, include: penalizing parameters during training (such as weight decay) or trimming connectivity. CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler patterns embossed in their filters. Therefore, on a scale of connectivity and complexity, CNNs are on the lower extreme.

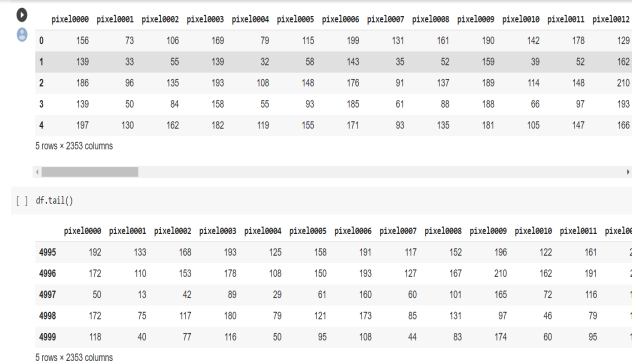
**EXISTING SYSTEM :**

The key problem in the treatment of cancer is the early detection of the disease. Often, cancer is detected in its later stages, when it has compromised the function of one or more vital organ systems and is widespread throughout the body. Methods For the early detection of cancer are of utmost importance and are an active area of current research. Magnetic Resonance imaging (MRI) established itself as the primary method for detection and staging in patients with colorectal cancer. In this paper, MRI images of colorectal cancer are used to detect the area and mean values of tumor area and distance from tumor area to other parts for staging cancer. This research paper describes algorithms for pre-processing, clustering and segmentation of MRI images. The implementation of this clustering, segmentation and pre-processing is done with MATLAB 2015 (a). By using this proposed methodology, the cancer is detected in its early beginning stage.

**Fig 3: Flow diagram**

**DATASET :**

A dataset is a collection of [data](https://en.wikipedia.org/wiki/Data). In the case of tabular data, a data set corresponds to one or more [database tables](https://en.wikipedia.org/wiki/Table_(database)), where every [column](https://en.wikipedia.org/wiki/Column_(database)) of a table represents a particular variable, and each [row](https://en.wikipedia.org/wiki/Row_(database)) corresponds to a given record of the data set. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum.



**Fig 4: Dataset**

**RESULT ANALYSIS:**

**ACCURACY**

**Accuracy :** It gives you the overall accuracy of the model, meaning the fraction of the total samples that were correctly classified by the classifier. To calculate accuracy, use the following formula: (TP+TN)/(TP+TN+FP+FN).

**Misclassification Rate :** It tells you what fraction of predictions were incorrect. It is also known as Classification Error. You can calculate it using (FP+FN)/(TP+TN+FP+FN) or (1-Accuracy).

**Precision:** It tells you what fraction of predictions as a positive class were actually positive. To calculate precision, use the following formula: TP/(TP+FP).

**Recall:** It tells you what fraction of all positive samples were correctly predicted as positive by the classifier. It is also known as True Positive Rate (TPR), Sensitivity, Probability of Detection. To calculate Recall, use the following formula: TP/(TP+FN).

**Specificity:**It tells you what fraction of all negative samples are correctly predicted as negative by the classifier. It is also known as True Negative Rate (TNR). To calculate specificity, use the following formula: TN/(TN+FP).

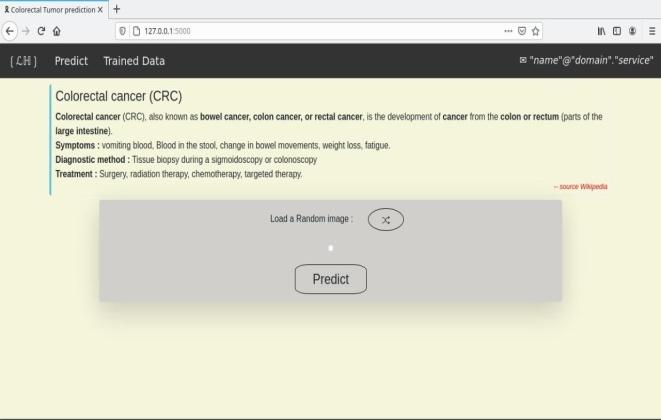
**F1-score:** It combines precision and recall into a single measure. Mathematically it’s the harmonic mean of precision and recall. It can be calculated as follows:

### F1-score = 2\* precision \* Recall/ precision + Recall = 2TP/ 2TP + FP + FN

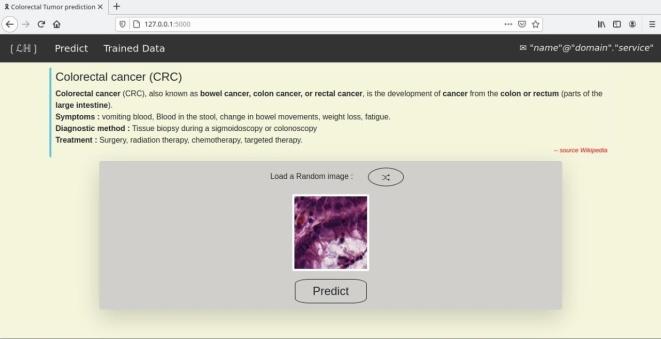
Now, in a perfect world, we would want a model that has a precision of 1 and a recall of 1. That means a F1-score of 1, i.e. a 100% accuracy which is often not the case for a machine learning model. So what we should try, is to get a higher precision with a higher recall value.

**SCREEN SHOTS :**

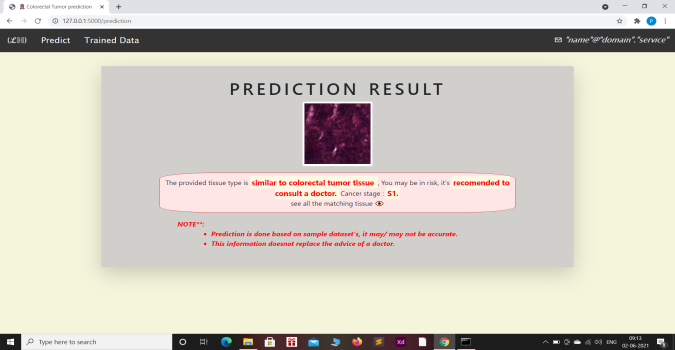
In this project we build a web application for displaying the predicted output. We used Amazon EC2 as our front end for webpage development which is provided by AWS. Amazon EC2 is a web service that provides secure, resizable compute capacity in the cloud. It is designed to make web-scale cloud computing easier. It allows us to rent virtual computers on which to run their own computer application. Here in the backend we use flask command in the command prompt to link the data to the webpage. Flask command gets the permission AWS for using their web services.



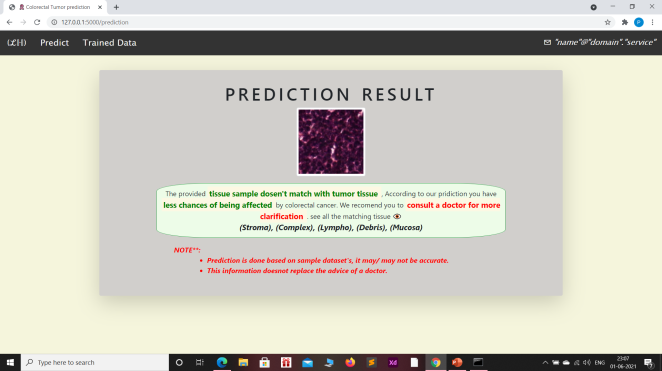
**Fig 4: Web Page**

****

**Fig 5: Uploading the image**

****

**Fig 6: Result more chance of colorectal cancer**

****

**Fig 7: Result with less chance of colorectal cancer**

# CONCLUSION AND FUTURE WORK:

Early detection of cancer is very important in the medical field. In this work, we present an image-based feature extraction, segmentation and training approaches for classification and screening the cancer tissues. The previous work focused on k-mean clustering which is less efficient and accuracy was 73%. In our work we have increased the accuracy to 94% by training the model with EPOCH. LSTM is used for fast processing and stores the best result for comparison in the future. Future works will be directed towards analysis of additional data sets acquired under controlled imaging conditions. Since the datasets under analysis in this work represent a huge variety of imaging condition variabilities, the observations from the experimental analysis are more generalizable zableyet limited in classification capabilities.

**REFERENCE :**

[1].M.Jayakandan, T.Manivannan(February 2018) “Colorectal Cancer Detection in MRI Images Using Image Processing Techniques”International Journal of Engineering Science & Research Technology

[2]. Bunil Kumar Balabantaray,Kangkana Bora, Kunio Kasugai and Pallabi Sharma(2020) **“Two Stage Classification With CNN For Colorectal Cancer Detection”**Oncologie (SCI)

[3]. [Jiri Prinosil](https://ieeexplore.ieee.org/author/37393794700),[Malay Kishore Dutta](https://ieeexplore.ieee.org/author/38256943900), [Namita Sengar](https://ieeexplore.ieee.org/author/37085439483), [Neeraj Mishra](https://ieeexplore.ieee.org/author/37086201371), [Radim](https://ieeexplore.ieee.org/author/37666304800) [Burget](https://ieeexplore.ieee.org/author/37666304800)(2016)**“Grading Of Colorectal Cancer Using Histology Images”** 39th International Conference on Telecommunications and Signal Processing,[29 June](https://ieeexplore.ieee.org/xpl/conhome/7750937/proceeding) [2016](https://ieeexplore.ieee.org/xpl/conhome/7750937/proceeding)

[4]. KorsukSirinukunwattana et al.,(2016)[**“Locality Sensitive Deep Learning for**](http://ieeexplore.ieee.org/document/) **Detection and Classification of**[**Nuclei in Routine Colon Cancer Histology**](http://ieeexplore.ieee.org/document/)[**Images”**](http://ieeexplore.ieee.org/document/),[IEEE Transactions on Medical Imaging](http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=42)(Volume:35, [Issue: 5](http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=7463083)

**[5] .**P. P. Mainenti, A. Stanzione, S. Guarino, V. Romeo, L. Ugga, F. Romano, et al., "**Colorectal cancer: Parametric evaluation of morphological functional and molecular tomographic imaging**", World J. Gastroenterol., vol. 25, no. 35, pp. 5233-5256, Sep. 2019.

[6]. Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei and W. Liu, "**CCNet: Criss-cross attention for semantic segmentation**", Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 603-612, Oct. 2019.

[7]. H. Dong, G. Yang, Y. Mo, Y. Guo and F. Liu, "**Automatic brain tumor detection and segmentation using u-net based fully convolutional networks**", Proc. Conf. Med. Image Understand. Anal., pp. 506-517, Jul. 2017.

[8]. J. Wang, J. Lu, G. Qin, L. Shen, Y. Sun, H. Ying, et al., **"Technical note: A deep learning-based autosegmentation of rectal tumors in MR images"**, Med. Phys., vol. 45, no. 6, pp. 2560-2564, Jun. 2018

[9]. N. Hoshino, T. Sakamoto, K. Hida and Y. Sakai, "**Diagnostic accuracy of computed tomography colonography for tumor depth in colorectal cancer: A systematic review and meta-analysis**", Surgical Oncol., vol. 30, pp. 126-130, Sep. 2019.

[10]. R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. S. Meester, A. Barzi, et al., "**Colorectal cancer statistics 2017"**, CA Cancer J. Clinicians, vol. 67, no. 3, pp. 177-193, May 2017.

[11]. Y. Yuan, "**Prognostic and survival analysis of 837 chinese colorectal cancer patients**", World J. Gastroenterol., vol. 19, no. 17, pp. 2650-2659, 2013.

[12]. L. Cheng, C. Eng, L. Z. Nieman, A. S. Kapadia and X. L. Du, " **v**", Amer. J. Clin. Oncol., vol. 34, pp. 573-580, Dec. 2011.