May - June 2020
ISSN: 0193-4120 Page No. 4130-4134

Colon Cancer Detection using Data Mining

[1]Kalpana G, [2]Kallish Kumar N, [3]Sagayasree Z, [4]Sushanth Arunachalam

[1,2,3,4]Department of Computer Science and Engineering, Rajalakshmi Institute of Technology, Chennai, India
[1]g.gkalpana@gmail.com, [2]kallishkumar.n.2017.cse@ritchennai.edu.in,
[3]sagayasree.z.2017.cse@ritchennai.edu.in, [4]sushantharunachalam.2017.cse@ritchennai.edu.in

Article Info
Volume 83
Page Number: 4130-4134
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract

In this ever-changing world there are many health issues, that humans have not been able to win over. Cancer is one such disease that has been for decades challenging the medical industry throughout the world. There are many stages and types of Cancer classified by the part of the body which is affected by the cancer cells such as Colon, Rectal, Lung, Prostate, Ovarian etc., Diagnosis of cancer is done by physical tests and few laboratory tests such as Computerized Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Test (PET) etc.,. While the mortality rate of Cancer is high all over the world, it is seen that detecting cancer at an early stage dramatically increase the chances of it getting cured. Data Mining and Machine Learning are playing a key role in detecting potential cancer patient's way earlier based on their clinical reports. The various reports from the pathology lab and imaging facilities have been put through scientific models to alert the medical fraternity about a potential cancer suspect at a very early stage.

Keywords: *Cancer, Data Mining, Dataset, Machine Learning*

1. Introduction

India hosts world's second largest population and is no exception to the wrath of Cancer. According to WHO[13],Cancer is a large group of diseases that can start in almost any organ or tissue of the body when abnormal cells grow uncontrollably, go beyond their usual boundaries to invade adjoining parts of the body and/or spread to other organs. Malignant and Benign are two major classifications of tumors of which Benign tumors are not cancerous, however malignant tumors are cancerous and can affect other tissues and organs. There are around 2.25 Million people living with cancer, on an average 11.5 Lakh new patients are registered every year and the cancer related death accounts to 7.48 Lakh. 47.2% of all cancers are related to the top 5 cancers in men (Oral Cavity, Lung, Stomach, Colorectal, Esophagus) and women (Breast, Oral, Cervix, Lung, Gastric). These cancers prevented if they are detected in an early stage[12]. Data Mining, as the space is largely known for, helps the human race to understand complex problems with the power of large datasets and analytics. This ability to wrangle large datasets, learning about the trends and formulating a model comprising of various nuances of decision making, is proving of great assistance in identifying Cancer at early stage by putting the rather limitedly used Medical reports and Imageries, to use.

2. Related Work

[1][2][6]Data mining is a concept that is used in various fields of science such as classification of tumor, clustering of gene expression data, cancer classification etc. Support Vector Machine and Bagging concepts are used largely to predict Colon Cancer. Colon is the 3rd most threatening cancer in India and is ranked the 4th in world. According to Namazi, Radio Therapy Center data, out of 567 patients of colon cancer, in stage 1-4, 338 patience are alive and 268 are deceased, a mortality rate of almost 47%.Weka tool was used in the prediction of survival rate of colon cancer patients. The efficiency and performance of the SVM and Bagging were checked using the confusion matrix. The rate of specificity, sensitivity and accuracy of the support vector machine was 84%, 87% and 84.4% and for the bagging method it was 78%, 88% and 89.3% respectively. It is evident that out of the two methods that were used to predict the survival of the colon cancer patients, bagging method is more efficient compared to the support vector machine method.

May - June 2020
ISSN: 0193-4120 Page No. 4130-4134

[3] Gene level extraction is developing in healthcare sector, it is useful in many research fields. Data mining in gene expression data is an emerging research field which can be used to predict colon cancer from the hereditary traits. Clustering with Machine and Deep learning neural network algorithms are key to the analysis of Gene Expression data obtained. A model is designed based on shallow neural network with the required variates, this uses helps in the prediction of colon cancer.

[4] Predicting colon cancer by using hybrid of novel geometric features and traditional features as explained in the reference publication, provides clear view on the stages of cancer. The proposed solution uses a colon classification technique, Hybrid Feature Space Based Colon Classification (HFS-CC), which comprises of classifiers with different types of discriminative features such as image texture, geometric structure etc., and conventional features such as SIFT and EFDs.

[5] Another colon cancer prediction method called, the SMOTE i.e.(synthetic minority over sampling technique) was used to predict the survival and non-survival rate of cancer patients in a particular area. The efficiency of the method was found to be high in terms of survival prediction. The main aim is to reduce the mortality rate from colon cancer. Using this method, the given samples were tested year after year to see the percentage of decrease in non - survival rate. After 5 years, 13 attributes were selected using the SMOTE method from the total 64 attribute set to predict the results.

[7] Data mining techniques are used in predicting diseases that cause death in the long run. Colon cancer can be detected collecting related data from the patients like symptoms, side effects etc., grouping the same, analyzing the same.

[8] Colon cancer is the second dangerous cancer type in the world. The primary treatment for colon cancer is surgery for the patients. Cancer Recurrence Support System (CARES) is used to resolve the colon cancer problem. It also uses Case-Based Reasoning (CBR). It makes comparison between the positive cancer patients and negative cancer patients in order to detect further cancer cases.

3. System Methodology

3.1 Symptoms of Colon Cancer[9]

Symptoms of colon cancer include:

1. A change in bowel habits, diarrhea, constipation or a change in the consistency of stool that lasts longer than four weeks.
2. Rectal bleeding or blood in stool
3. Persistent abdominal discomfort, such as cramps, gas or pain
4. A feeling that bowel doesn't empty completely.
5. Weakness or fatigue

3.2 Screening[10]

Colon is a type of cancer that is visible only at an advanced stage and hence the screening test are done to check the stage of the polyp that has grown. The screening tests are as follows:

1. Colonoscopy: A flexible tube with light called the colonoscopy is send into the rectum to check for grown polyp or cancer.
2. Computer tomography or colonography: This is done if the polyp disturbs the examining process or if the patient is prone to Anesthesia.
3. Sigmoidoscopy: This is similar to colonoscopy, but sigmoidoscopy can be used to remove polyps if not in the advanced stage.
4. Fecal Occult Blood Test (FOBT) and Fecal Immunochemical Test(FIT):
5. Double Contrast Barium Enema (DCBE): Stool DNA tests.

3.3 System Design

The proposed system uses a Machine Learning algorithm called Logistic Regression. This model works based on the input criteria and returns the output as either probability of one result or another .Its prediction range lies between 0 and 1.This algorithm is majorly of three types, such as

a) Binomial Logistic Regression Model
b) Multinomial Logistic Regression Model
c) Ordinal Logistic Regression Model

The Binomial Logistic Regression model is a type of logistic regression model which has only two categories of outputs. The output will be based on the true or false output or to classify other things etc. The prediction values are based on the probability ranging from 0 and1.

The Multinomial Logistic Regression Model is a type belonging to logistic regression where this model involves three or more than three categories of outcomes without ordering the outcomes. The outcomes predicted by this model are not being ordered.

The Ordinal Logistic Regression Model is a type of logistic regression which involves three or more than three categories, but the outcome will be in an ordered manner.

Each type of logistic regression model has its respective features. The proposed system makes use of the multinomial logistic regression model. The regression model consists of various functions involved in it, such as cost function and sigmoid function. There are other parameters involved in the logistic regression such as the setting threshold value, predicting cost function and model evaluation which is used to evaluate the model by generating a confusion matrix. A confusion matrix is in the form of a table which evaluates the performance of the model based on the test dataset being provided to the

TEST Engineering & Management

May - June 2020
ISSN: 0193-4120 Page No. 4130-4134

model. The accuracy of the model lies between 0 and 1.

$$p = \frac{1}{1 + e^{-y}}$$

Figure 1: Sigmoid Function

The Sigmoid Function is also said to be called as logistic function which would always return a value between 0 and 1 and it doesn't matter whatever value provided, but the result

are between 0 and 1.In order to predict it based on the features of algorithm, a threshold value is set.

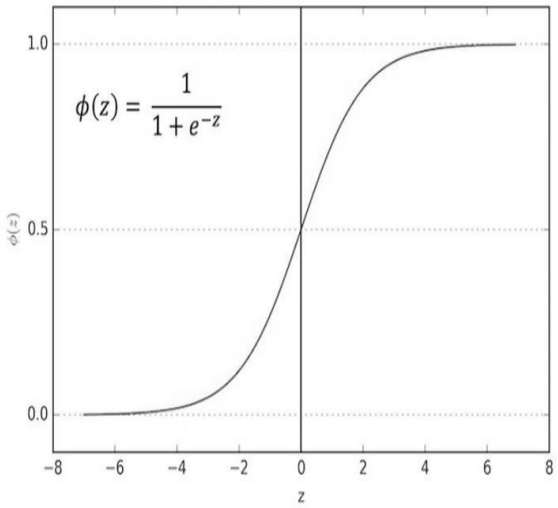

Figure 2: Sigmoid Function Curve

The pictorial representation of the sigmoid function's curve which looks like S shaped. A threshold is being set at the center of 0 and 1 which is the value 0.5.Anything which is above the threshold value is predicted as one class and anything below the threshold value is predicted as other class.

The purpose of using cost function is to reduce the cost for training the model and used to increase the probability value and helps in minimizing the loss which occurs during the training of the model.

$\text{Cost}(h_\theta(x), y) = -y \log(h_\theta(x)) - (1-y) \log(1 - h_\theta(x))$

If y = 1, (1-y) term will become zero, therefore – log ($h_\theta(x)$) alone will be present

If y = 0, (y) term will become zero, therefore – log (1- $h_\theta(x)$) alone will be present

Figure 3: Simplified Cost Function

The proposed model consists of three modules in which the system is being designed.

a) Building dataset and Train Logistic Regression Model
b) Testing the model
c) Predicting the stage of colon cancer

3.3.1 Building dataset and Train Logistic Regression Model

The proposed system first collects the required dataset, i.e., the images with colon cancer as well as images with non-cancer, in the form of an.csv file. The collected images undergo preprocessing steps such as data cleaning, data reduction and then data transformation. After this step the model selects features from the images and then it constructs two kinds of variables such as dependent variable and independent variable. Then they are separated into two columns inside the dataset. The dataset is split into two types of data.

a) Training Data
b) Testing Data

The Training data is used for training the model with the greatest number of images present in it. The Training dataset consists of 75% of the images in the dataset. The Testing dataset is used to test the model for its prediction accuracy and results are applied on the test images. The testing dataset consists of 25% of the images in the dataset.

The model is trained with the training dataset and then the testing dataset is being applied to the model for predicting the accuracy obtained on the test images. Then the required regression model is built with the custom dataset. The model is now subjected to evaluation. The model is evaluated by using a confusion matrix. The confusion matrix is used to evaluate the performance of the model. Then the system is used to visualize the confusion matrix using the heat map technique. The confusion matrix displays its evaluation metrics and based on the metrics being generated, a Receiver Operating Characteristic (ROC) curve is generated. In the Fig. 4, an example for the ROC curve is being examined. It is used to plot a graph which consists of the true positive rate and false positive rate. It also shows the difference between the sensitivity and specificity. The accuracy score is being displayed in the graph and the value always lies between the range of 0 and 1. The accuracy value is 0.86.

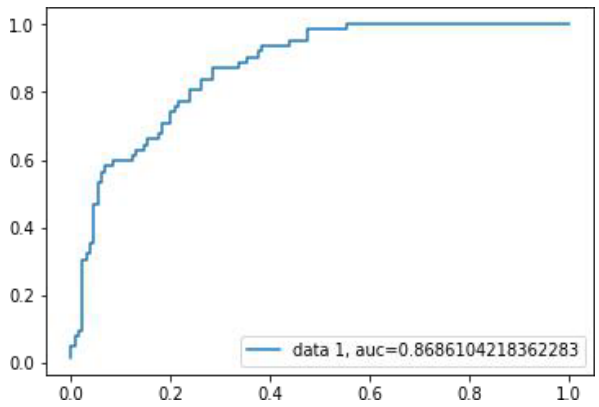

Figure 4: ROC Curve

May - June 2020
ISSN: 0193-4120 Page No. 4130-4134

3.3.2 Testing the Model

Once the model has been designed with the custom dataset, the model is provided with the input dataset which consists of the both cancer and non-cancer images. The dataset undergoes the data preprocessing stage. It consists of three stages such as data cleaning, data reduction and data transformation. After preprocessing the data, the dataset is fed into the trained logistic regression model. The model applies its feature learning techniques over the input images and then an evaluation model is constructed where the model's predicted confusion matrix, ROC curve and the prediction results are being displayed. The test results fall in the range of 0 and 1.The test result based on the accuracy value, predicts that the respective image has cancer or not.

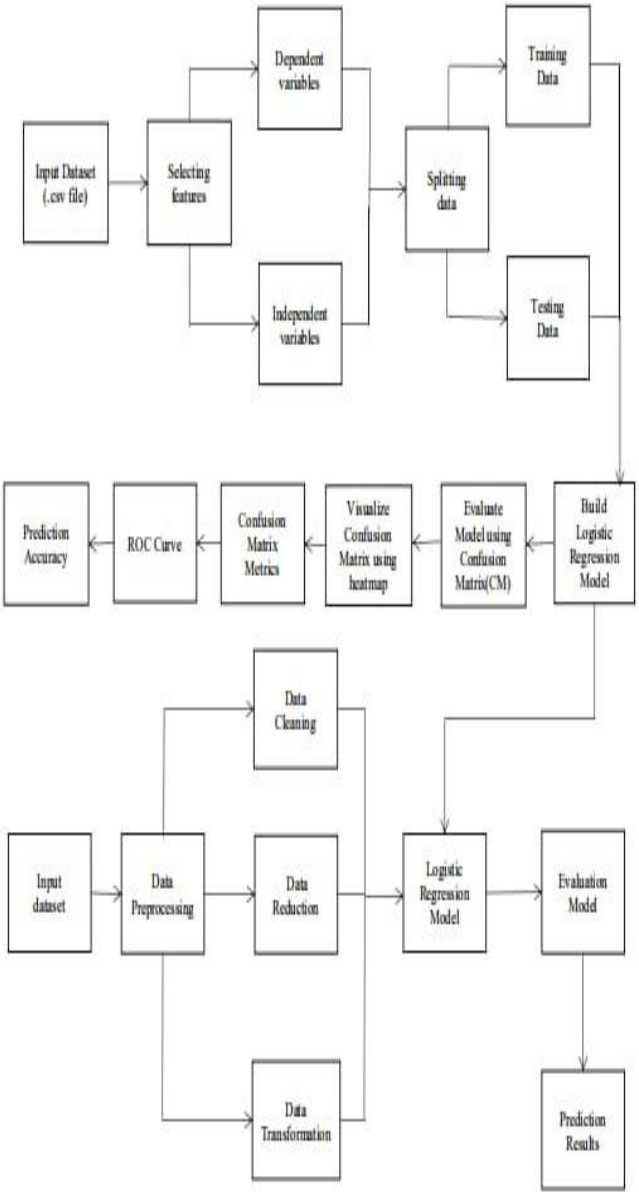

Figure 5: Building dataset, regression model and testing model

3.3.3 Predicting the stage of colon cancer

After testing the model, the prediction results are taken and then it is tested for the stage of cancer whether it is in stage 1,stage 2,stage 3, or stage 4.In order to predict these stages we use ordinal logistic regression model.

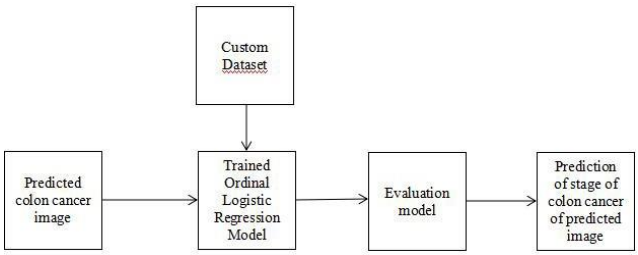

Figure 6: Prediction of colon cancer stage

The ordinal regression model is used to take three or more categories and provides the end results. The results of the predicted image falls under the range of 0 and 1.The cancer positive images are fed to the ordinal logistic regression model where the model is already trained with custom dataset and the algorithm applies its learning features and it evaluates the model and predicts the cancer stage with accuracy value displayed.

4. Results and Conclusion

The proposed logistic regression system is used to collect the dataset and train the model accordingly by providing the training and testing dataset separately .The system trains the model with the images provided and then is evaluated with the confusion matrix plotted in the ROC curve. The curve displays the accuracy value in terms of predicting whether the image has cancer or not.

The proposed system then makes use of the predicted cancer image with ordinal regression algorithm which takes multiple parameters as input and provides an ordered output. This model is already trained with the custom dataset which consists of images such as cancer images, stage ranging from one to fourth stage. Then the model applies its learning feature technique and compares the predicted cancer image with all four stages and constructs an evaluation model where the results are evaluated. The output is displayed as the image with an accuracy value of ranging from 0 and 1 has this stage of cancer respectively.

Thus, the proposed system helps in predicting colon cancer as well as it also predicts the stage of cancer which may help doctors to rectify the problem as early as possible and helps in saving the life of people and increases the probability to recover from the cancer.

References

[1] Lean Suan Ong, Barry Shepherd, Loong Cheong Tong, Francis Seow-Choen, Yik Hong Ho, Choong Leong Tang, Yin Seong Ho, Kelvin Tan, "The colorectal cancer recurrence support (CARES) system", *Artificial Intelligence in*

May - June 2020
ISSN: 0193-4120 Page No. 4130-4134

Medicine 11 (3),175-188,1997.

[2] ario Antonelli, Elena Baralis, Giulia Bruno, Silvia Chiusano, Naeem A Mahoto, Caterina Petrigni, "Analysis of diagnostic pathways for colon cancer", *Flexible Services and Manufacturing Journal* 24 (4), 379-399,2012.

[3] Reda Al-Bahrani, Ankit Agrawal, Alok Choudhary, "Colon cancer survival prediction using ensemble data mining on SEER data", *2013 IEEE international conference on Big Data*, 9-16,2013.

[4] DSVGK Kaladhar, Bharath Kumar Pottumuthu, Padmanabhuni V Nageswara Rao, Varahalarao Vadlamudi, A Krishna Chaitanya, R Harikrishna Reddy, "The elements of statistical learning in colon cancer datasets: data mining, inference and prediction", *Algorithms Research* 2 (1), 8-17,2013.

[5] Huaming Chen, Hong Zhao, Jun Shen, Rui Zhou, Qingguo Zhou,"Supervised machine learning model for high dimensional gene data in colon cancer detection", 2015 IEEE International Congress on Big Data, 134-141,2015.

[6] Saima Rathore, Mutawarra Hussain, Asifullah Khan, "Automated colon cancer detection using hybrid of novel geometric features and some traditional feature", Computers in biology and medicine 65, 279-296,2015.

[7] S Setareh, M Zahiri Esfahani, M Zare Bandamiri, A Raeesi, R Abbasi," Using data mining for survival prediction in patients with colon cancer ", *Iranian Journal of Epidemiology* 14 (1), 19-29, 2018.

[8] Mohammed Othman, Faten F Khatbat, Tarik Al Amsy, "Exploring colorectal cancer genes through data mining techniques", 2018 Advances in Science and Engineering Technology International Conferences (ASET), 1-4,2018.

[9] Colon cancer symptoms are available [Online].https://www.mayoclinic.org/diseases-conditions/colon-cancer/symptoms-causes/syc-20353669.

[10] Screening test for colon cancer are available [Online]. https://www.cancer.net/cancer-types/colorectal-cancer/screening.

[11] Logistic Regression https://en.wikipedia.org/wiki/Logistic_regression.

[12] http://cancerindia.org.in/cancer-statistics/

[13] Cancer Definition -https://www.who.int/health-topics/cancer