A Hybrid Learning Approach for Detection of Cervical Cancer and Prediction of Survival Outcome

by

Name	Roll No.	Registration No.
Rounak Saha	**11700116056**	**161170110057** of 2016-17
Oishik Mukhopadhyay	**11700116066**	**161170110047** of 2016-17
Jaya	**11700116077**	**161170110036** of 2016-17
Rangeet Choudhury	**11700116059**	**161170110054** of 2016-17

Under the guidance of

Asst. Prof. Mr. Anirban Dey

Dept. of Computer Science & Engineering

RCC Institute of Information Technology

PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING
RCC INSTITUTE OF INFORMATION TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
RCC INSTITUTE OF INFORMATION TECHNOLOGY
[Affiliated to West Bengal University of Technology]
CANAL SOUTH ROAD, BELIAGHATA, KOLKATA-700015

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

RCC INSTITUTE OF INFORMATION TECHNOLOGY

TO WHOM IT MAY CONCERN

I hereby recommend that the Project entitled **"A Hybrid Learning Approach for Detection of Cervical Cancer and Prediction of Survival Outcome"** prepared under my supervision by Rounak Saha (11700116056), Oishik Mukhopadhyay (11700116066), Jaya (11700116077) and Rangeet Choudhury (11700116059) of B.Tech (8th Semester), may be accepted in partial fulfillment for the degree of **Bachelor of Technology in Computer Science & Engineering** under West Bengal University of Technology (WBUT).

Anirban Dey
7/7/20

.....

Project Supervisor

Department of Computer Science and Engineering

RCC Institute of Information Technology

Countersigned:

.....

Head

Department of Computer Sc. & Engg,

RCC Institute of Information Technology

Kolkata – 700015

CERTIFICATE OF APPROVAL

The foregoing Project is hereby accepted as a credible study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the project only for the purpose for which it is submitted.

FINAL EXAMINATION FOR

EVALUATION OF PROJECT

1. Anirban Dey 8/7/20

2. Minakshi Banerjee

3. P Ghosh

(Signature of Examiners)

ACKNOWLEDGEMENT

We acknowledge our overwhelming gratitude & immense respect to our revered guide, Mr. Anirban Dey (Asst. Prof., RCC Institute of Information Technology) under whose scholarly guideline, constant encouragement & untiring patience; we have proud privilege to accomplish this entire project work. We feel enriched with the knowledge & sense of responsible approach we inherited from our guide & shall remain a treasure in our life.

Rounak Saha
Roll No: 11700116056

Oishik Mukhopadhyay
Roll No: 11700116066

Jaya
Roll No: 11700116077

Rangeet Choudhury
Roll No: 11700116059

ABSTRACT

"Cancer affects all of us, whether you're a daughter, mother, sister, friend, coworker, doctor, or patient." Cancer is one of the leading causes of death in the World. Cancer is common enough that most people in the world know someone who has had cancer in their lifetime. Cervical cancer is the fourth most frequent cancer in women with an estimated 570,000 new cases in 2018 representing 6.6% of all female cancers. In our paper we are presenting a model which can detect the cancer by using clinical data and medical (endoscopic) images. This model is built using various Statistical Models in which data sufficiency is nearly optimal. The model has greater than 96% Rank-1 accuracy in all the test case scenarios.

CONTENTS

LIST OF SYMBOLS

Symbol	Meaning	Page No.
m_i	*First Order Moment Of* i^{th} *class*	4
n	*Number of Classes*	4
$E(m_i)$	*Expectation of the first order means of the classes*	4
w_i	*Weighted Parameter of* i^{th} *class*	4
W_k	*Weighted first order moment of* k^{th} *patient*	4
$\vec{X}$	*Flattened Vector* $(1\times m)$ *of transformed Image matrix*	5
C_v	*Compressed Statistical Value*	5
$\mu_{\vec{X}}$	*Mean of the* $\vec{X}$	5
$\sigma_{\vec{X}}$	*Standard Deviation of the* $\vec{X}$	5

LIST OF ABBREVIATIONS

Abbreviation	Full Form / Meaning
HPV	Human Papillomavirus
IUD	Intrauterine Device
STD	Sexually Transmitted Disease
SVM	Support Vector Machine
CNN	Convolution Neural Network
RGVF	Radiating Gradient Vector Flow
MRI	Magnetic Resonance Imaging
SEM	Structural Equation Modeling
ReLU	Rectified Linear Unit
KECA	Kernel Entropy Component Analysis

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Cervical cancer:

It is a cancer that starts in a woman's cervix, which is the lower, narrow part of the uterus. The cervix connects the lower part of the uterus to the vagina and, with the vagina, forms the birth canal. Cervical cancer begins when healthy cells on the surface of the cervix change and grow out of control, forming a mass called a tumor. Cervical cancer begins with abnormal changes in the cervical tissue. The risk of developing these abnormal changes is associated with infection with human papillomavirus (HPV). Genetic material that comes from certain forms of HPV (high-risk subtypes) has been found in cervical tissues that show cancerous or precancerous changes.

Cervical cancer is most often diagnosed between the ages of 35 and 44. About 15% of cervical cancers are diagnosed in women over age 65. It is rare for women younger than 20 to develop cervical cancer.

The 5-year survival rate tells you what percent of women live at least 5 years after the cancer is found. Percent means how many out of 100. The 5-year survival rate for all women with cervical cancer is 66%. However, survival rates can vary by factors such as race, ethnicity, and age. It is important to remember that statistics on the survival rates for women with cervical cancer are an estimate. The estimate comes from annual data based on the number of women with this cancer in the United States. Also, experts measure the survival statistics every 5 years. So the estimate may not show the results of better diagnosis or treatment available for less than 5 years.

Cervical cancer is treated in several ways. It depends on the kind of cervical cancer and how far it has spread. Treatments include surgery, chemotherapy, and radiation therapy. It is important to note that this cancer may be nearly untreatable in the late stage but is easily treatable in the early stages; hence correctly diagnosing it in the early stages is of utmost importance.

1.2 Challenges Faced:

A lot of previous work has been done in this field but invariably faced many challenges. A very large data set is usually required. Accuracy often suffers due to data loss in the various models and layers of the neural network. Approaches using SVMs may lead to underfit or overfit in the final result.

1.3 Overcoming the Challenges:

Our model is based on Statistical Data Sufficient methods which are combined with classical Deep learning techniques that are able to capture very complicated patterns with minimum data loss. The previous working models are based mainly using PCA or Classical Image Processing techniques which suffer from data loss.

Literature Review

We have done the following literature review, as discussed in table 1.1:

Table 1.1 : Literature Review

Sl No	Author Name	Paper Name	Contribution	Challenges
1	Zhang et al.	Cervical Cancer Detection Using SVM Based Feature Screening	Pixel-level classification showed improvements to previous methods.	Very large feature vectors for images. High computation time. SVM suffers from data loss and overfit/underfit problems.
2	Hector G et al.	Cervical Cancer Detection Using Colposcopic Images: a Temporal Approach.	Temporal approach and expert system had promising results.	Very large data set as well as expert input required.
3	Rahmadwati et al.	Cervical Cancer Classification Using Gabor Filters	Gabor filters and K-means clustering.	Gabor filters may not give independent features. K-means clustering needs large data set.
4	Sajeena et al.	Automated Cervical Cancer Detection through RGVF segmentation and SVM Classification	Novel approach of Radiating Gradient Vector Flow (RGVF) Snake with SVM to focus on individual cells.	High level of image segmentation required. RGVF has trouble with overlapping/clustered cells and may get trapped in local minima.
5	Roy et al.	Cervical Cancer Detection from MR Images based on multiresolution wavelet analysis	Discrete wavelet transforms have significant advantages and provided good results.	Less "natural" and computationally expensive for fine analysis. Also used SVMs as explained above.
6	Kashyap et al.	Cervical Cancer Detection And Classification Using Independent Level Sets And Multi SVMs	Highly complex image processing methods combined with Multi SVMs provided very high accuracy.	High level of image processing is also computationally expensive. Multi SVMs suffer from similar problems.
7	Iwai et al.	Automatic Diagnosis Supporting System for Cervical Cancer using Image Processing	Novel method focusing only on individual nuclei of the cells.	Somewhat simplistic approach and the classification model is not highly accurate.

CHAPTER 2

Work Flow

Our proposed model has three main phases taking advantage of both the clinical data and the medical image data (endoscopic) before combining the data. The clinical data is fed through a Statistical Data Analyser and then Euclidean Divergence of each patient is calculated. The medical image data is passed through a modified Convolutional Neural Network (CNN) which consists of Convolution, ReLU, Final Vector Generation and Statistical Data Compression layers.

Finally both the processed Image and Clinical data are fed in the Structural Equation Model where the Kernel Entropy Component Analysis is used to classify the Cancerous cell and if Cancer is determined then find which stage it belongs to. The Work Flow diagram is mentioned in Fig 2.1

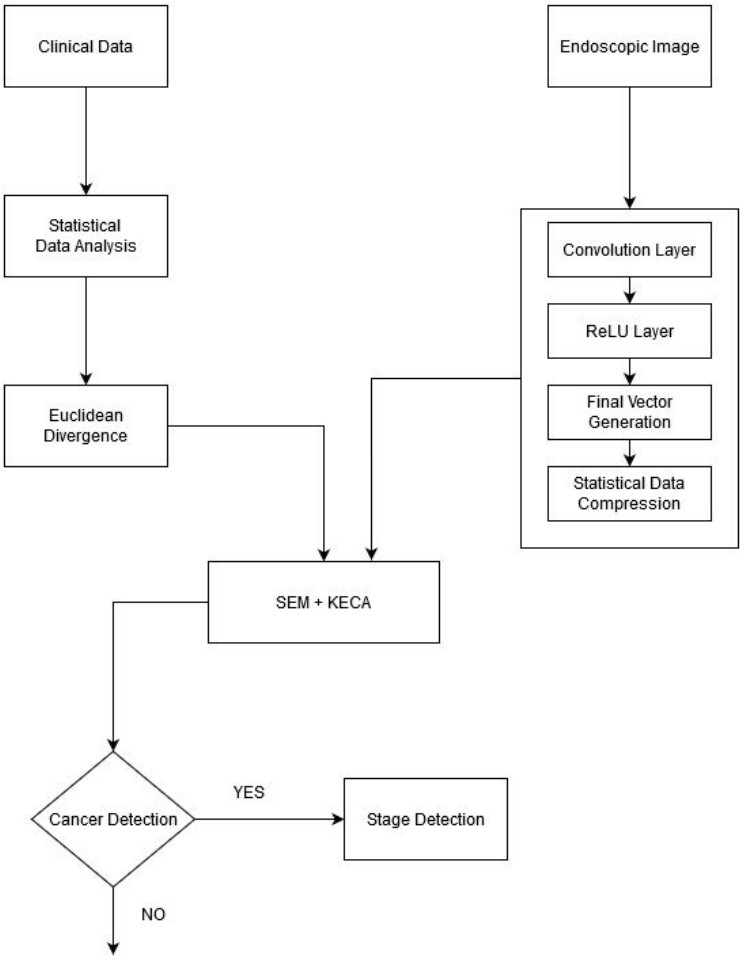

Fig 2.1 This diagram depicts the work flow which we have used in our project. We use both clinical data and endoscopic images and processed them parallely using Statistical Data Analysis and Image Data Analysis respectively.

Methodologies of Implementation

To overcome all the previous challenges this paper proposes a model which not only considers image or clinical data but both. This model mainly works in 3 phrases 1. Statistical Data Analysis 2. Image data Analysis 3. Structural Equation Model And Kernel Entropy Component Analysis

2.1. Statistical Data Analysis:

In the first stage, clinical data such as 1. Age, 2. Number of Sexual Partners, 3. Number of Pregnancy, 4. Smokes(Year), 5. Hormonal Contraceptives(Year), 6. IUDs(Year), 7. STDs(Number) is analysed by our Statistical data Analyser using the following manner.

First, the data is normalised using max divisor approach then the First order moment is calculated. After that, the Expectation/Mean of the first-order moment calculated using the following equation

$$E(m_i) = \frac{\sum_{i=1}^{n} m_i}{n} \tag{2.1.1}$$

Where,
m_i = *First Order Moment Of* i^{th} *class*
n = *Number of Classes (Here n=7)*
$E(m_i)$ = *Expectation of the first order means of the classes*

Then the weighted first-order moment is calculated using the following formula

$$W_k = \frac{\sum_{i=1}^{n} w_i m_i}{n} \tag{2.1.2}$$

Where,

w_i = *Weighted Parameter of* i^{th} *class*
W_k = *Weighted first order moment of* k^{th} *patient*

Finally the Euclidean divergence is calculated using the following formula and sent to the Structural Equation Model.

$$D_E = \sqrt{\left|(W_k)^2 - (E(m_i))^2\right|} \tag{2.1.3}$$

2.2. Image Data Analysis

For image data analysis we have used the CNN but in a modified form where the plain endoscopic image is passed through the Convolution Layer and Relu layer but after that, the transformed image matrix is flattened and converted to a vector. After that Statistical Compression is applied where the mean and standard deviation is calculated from the values of the vector and put in the following equation mentioned below. We used this compression approach to reduce data loss caused by pooling.

$$\vec{X} = \begin{bmatrix} x_1 & \cdots & x_{m-1} & x_m \end{bmatrix}^T \tag{2.2.1}$$

Where,

$\vec{X} = Flattened\ Vector\ (1 \times m)\ of\ transformed\ Image\ matrix$

$$C_v = \frac{\mu_{\vec{X}}}{\frac{\sigma_{\vec{X}}}{\sqrt{m}}} \tag{2.2.2}$$

Where,

$C_v = Compressed\ Statistical\ Value$

$$\mu_{\vec{X}} = \frac{\sum_{i=1}^{m} x_i}{m} \ and \quad \sigma_{\vec{X}} = \frac{\sum_{i=1}^{m} (x_i - \mu_{\vec{X}})^2}{m} \tag{2.2.3}$$

Where,

$\mu_{\vec{X}}, \sigma_{\vec{X}} = Mean\ and\ Standard\ Deviation\ of\ the\ \vec{X}$

After that C_v is sent to the Structural Equation Model.

2.3 Structural Equation Model And Kernel Entropy Component Analysis

Finally after getting the values of D_E and C_v a two-variable equation created with bias B_θ using the following manner.

$$SEM = D_E + C_v + B_\theta \tag{2.3.1}$$

Where,

$SEM = Structural\ Equation\ Model.$

Initially the bias value is randomly selected for each of the data points later on we use Kernel Entropy Component Analysis (KECA) given by *Robert Jenssen* to partition the patients into several classes. After completion of the KECA the data is partitioned into 4 parts **1**. Non-Cancerous **2**. Cancer Type 1 **3**. Cancer Type 2 **4**. Cancer Type 3.

After each iteration done by KECA the B_θ value is adjusted accordingly so that after completion of the partitioning the Stage detection can be determined accordingly.

In this model we use KECA because the Entropy Content of the SEM is greater than 5 so that using normal PCA or KPCA the stage determinism becomes very difficult .

Software & Hardware Requirements

2.2 Software Specifications:

The following software is necessary for this project –

- Python 3 (using PyCharm IDE) for the bulk of the project
- Several python modules including – cv2, numpy, tensorflow, tflearn, matplotlib, tqdm, os
- The required OS and environment to run the project
- The dataset was gathered from : http://biogps.org/dataset/tag/cervical%20cancer/

2.3 Hardware Specifications:

The following hardware is necessary for this project –

- High End PC for image processing and CNN for final output

Our Approach

As has been previously mentioned, our proposed method takes advantage of both the clinical data and the endoscopic images. The clinical data is fed through an efficient statistical data model and the Euclidean Divergence is found. The endoscopic image data is also sent through a modified CNN layer and then compressed in a special way to minimise data loss. The compressed image data along with the Statistically analysed clinical data forms our overall statistical equation model . These SEM values finally partitioned using Kernel Entropy Component Analysis and further Bias Values of the SEM are changed in each iteration until independent partitions are formed. After completion of this stage we get an output as to whether cancer is detected or not. If cancer is detected, then we can even output which stage the cancer is. This approach to combine the data gives us a high level of accuracy with a relatively small dataset.

The dataset used (specified in software specifications) has the important feature that medical (endoscopic) images as well as clinical data are available.

CHAPTER 3

Test Cases & System Validation

We used datasets from various sources like Clinical Reports and Endoscopic Images for creating a robust, error-free system. We noticed that having two parallel data sources reduced classification errors by a big margin. The Clinical Reports and Endoscopic Images example are given in the following images.

	A	B	C	D	E	F	G	
1	Age	Number of sexual partners	Num of pregnancies	Smokes (years)	Hormonal Contraceptives (years)	IUD (years)	STDs (number)	ST
2	18	4	1	0	0	0	0	
3	15	1	1	0	0	0	0	
4	34	1	1	0	0	0	0	
5	52	5	4	37	3	0	0	
6	46	3	4	0	15	0	0	
7	42	3	2	0	0	0	0	
8	51	3	6	34	0	7	0	
9	26	1	3	0	2	7	0	
10	45	1	5	0	0	0	0	
11	44	3	4	0	2	0	0	
12	27	1	3	0	8	0	0	
13	45	4	6	0	10	5	0	
14	44	2	2	0	5	0	0	
15	43	2	5	0	0	8	0	
16	40	3	2	0	15	0	0	
17	41	4	3	0	0.25	0	0	
18	43	3	8	0	3	0	0	
19	42	2	?	0	7	6	2	
20	40	2	?	0	0	1	0	
21	43	2	4	0	15	0	0	
22	41	3	4	0	10	0	1	
23	40	1	1	0	0.25	0	2	
24	40	1	2	0	15	0	0	
25	40	3	3	0	3	0	0	
26	44	3	1	0	0	0	0	
27	39	5	2	0	0	1	0	
28	39	2	4	0	0	0	0	

Sheet1

Fig 3.1 This image is a screenshot from a .csv file containing attributes from Clinical Data of multiple patients

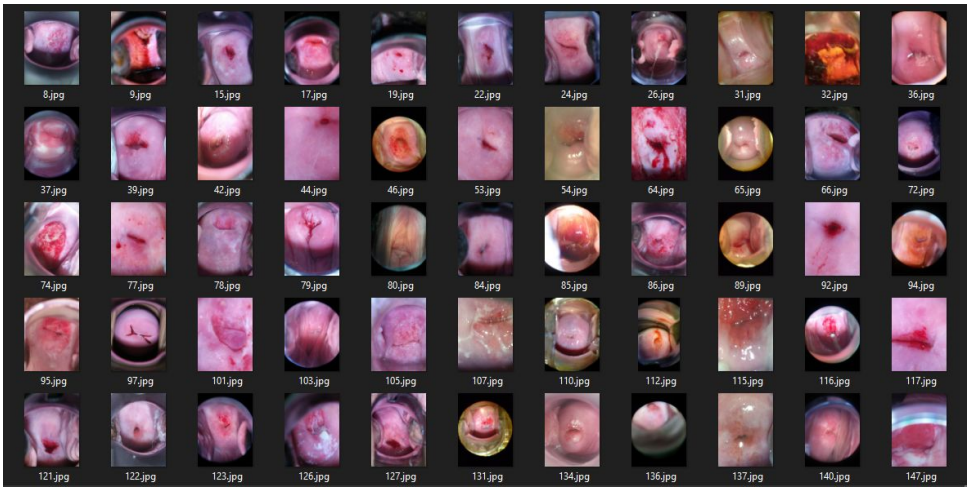

Fig 3.2 This image is a screenshot of a directory containing Endoscopic Images of multiple patients

Observed Output

```
In [13]: clinical_data['euclidean_divergence']

Out[13]: 0      0.207593
         1      0.205399
         2      0.208131
         3      0.208886
         4      0.205845
         5      0.203290
         6      0.203454
         7      0.208248
         8      0.210924
         9      0.202620
         10     0.200371
         11     0.201190
         12     0.202669
         13     0.190569
         14     0.183185
         15     0.183217
         16     0.183891
         17     0.189215
         18     0.197780
```

Fig 3.3 This figure is an example of Euclidean Divergence in Clinical Data

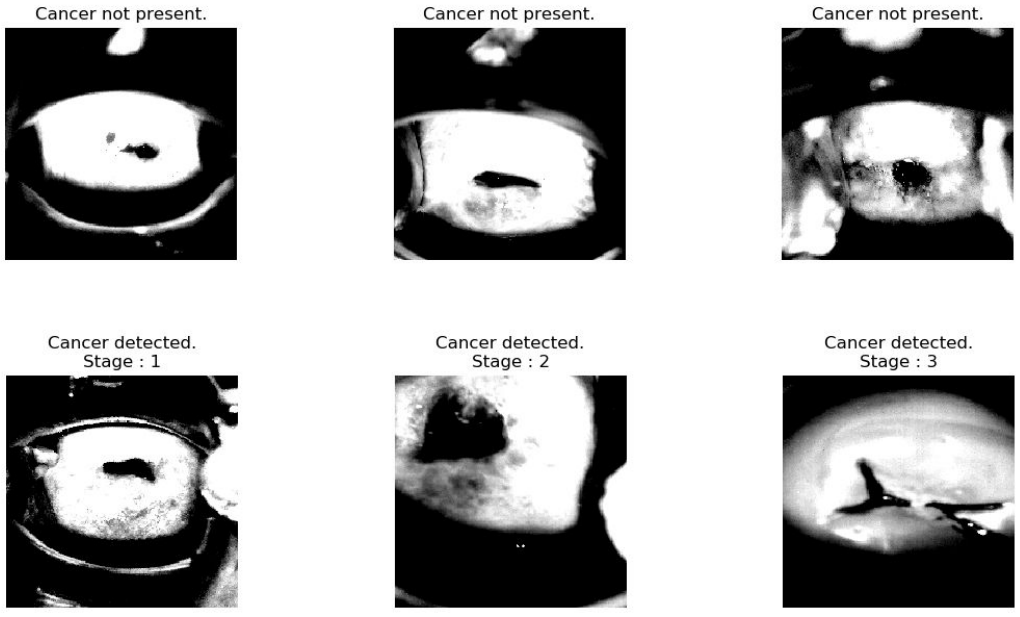

Fig 3.4 This figure is an example of 6 classified random images from the test dataset, 3 belonging to noncancerous class and 3 belonging to cancerous class, with each image in latter class being from each individual cancer stage

Performance Analysis

We have seen that cervical cancer is dangerous as it is difficult to diagnose in early stages and has a high mortality rate in late stages, however if it is caught early it can be treated easily. Due to this it is important to focus on accurate diagnosis early. Previous work in this field has suffered from various problems leading to low accuracy or overfitting/underfitting. However our approach which combines the input of clinical data with endoscopic images along with Data Sufficient methods to ensure minimum data loss and using SEM+KECA results in high accuracy of detection and can also accurately specify which stage the cancer has reached. We achieved an accuracy of 99.80% as compared to best of 98.80% in literature reviewed.

We have used 5000 samples for testing, which contained 3210 cancerous samples and 1790 noncancerous. Using our model we were able to determine 3210 cancer samples and 1780 non-cancer samples correctly.

Hence, we were able to achieve 99.80% accuracy.

Table 3.1 This table shows a comparative analysis between our project and related work as discussed in literature reviewed

Paper Name	Rank 1 Accuracy (%)	Entropy
Cervical Cancer Classification Using Gabor Filters, Rahmadwati et. al.	87	4.2588
Automated Cervical Cancer Detection through RGVF segmentation and SVM Classification, Sajeena et. al.	93.72	4.9856
Cervical Cancer Detection from MR Images based on multiresolution wavelet analysis, Roy et. al.	98.80	5.4439
Cervical Cancer Detection And Classification Using Independent Level Sets And Multi SVMs, Kashyap et. al.	95	5.1286
Automatic Diagnosis Supporting System for Cervical Cancer using Image Processing, Iwai et. al.	60.97	3.1258
Automatic Detection of Anatomical Landmarks in Uterine Cervix Images, Greenspan et. al.	90	4.8753
Prediction of Cervical Cancer using Voting and DNN Classifiers, Rayavarapu et. al.	90	4.7852
Our Approach (A Hybrid Learning Approach for Detection of Cervical Cancer and Prediction of Survival Outcome)	**99.80**	**5.6827**

References

1. Cervical Cancer Classification Using Gabor Filters, Rahmadwati et. al.

2. Automated Cervical Cancer Detection through RGVF segmentation and SVM Classification, Sajeena et. al.

3. Cervical Cancer Detection from MR Images based on multiresolution wavelet analysis, Roy et. al.

4. Cervical Cancer Detection And Classification Using Independent Level Sets And Multi SVMs, Kashyap et. al.

5. Automatic Diagnosis Supporting System for Cervical Cancer using Image Processing, Iwai et. al.

6. Automatic Detection of Anatomical Landmarks in Uterine Cervix Images, Greenspan et. al.

7. Prediction of Cervical Cancer using Voting and DNN Classifiers, Rayavarapu et. al.

8. Kernel Entropy Component Analysis, Robert Jennsen

9. Deep Learning, Ian Goodfellow