A REPORT

ON

DETECTION OF CERVICAL CANCER USING SUPPORT VECTOR MACHINES

BY

Names of the students	ID No.s
Kaustubh Butte	2016A8PS0364G
Ameya Zope	2016A7PS0721G

AT

Central Electronics and Engineering Research Institute, Pilani

A Practice School- I station of

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

(JUNE 2018)

A REPORT

ON

DETECTION OF CERVICAL CANCER USING SUPPORT VECTOR MACHINES

BY

NamesoftheStudents	ID Number	Disciplines
Kaustubh Butte	2016A8PS0364G	Electronics & Instrumentation
Ameya Zope	2016A7PS0721G	Computer Science

Prepared in partial fulfilment of the
Practice School-I Course

AT

CENTRAL ELECTRONICS AND ENGINEERING RESEARCH INSTITUTE, PILANI

A Practice School- I station of

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

(June 2018)

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

BIRLAINSTITUTEOFTECHNOLOGYANDSCIENCE,
PILANI
Practice School Division

Station Name and Centre: Central Electronics Engineering Research Institute, Pilani

Duration: 2months **Date of Start:** 22nd May 2018

Date of Submission: 13th July 2018

Title of the Project: DETECTION OF CERVICAL CANCER USING SUPPORT VECTOR MACHINES

Kaustubh Butte (2016A8PS0364G) B.E. Electronics & Instrumentation
Ameya Zope (2016A7PS0721G) B.E. Computer Science

Experts: Dr. Jagdish Lal Raheja, HOD, Computer Science Department

PS Faculty: Mr. Pawan Sharma

Keywords: GLCM Algorithm, C++, Machine Learning, Support Vector Machines, MATLAB, Feature Extraction, OpenCV

Project Areas: Machine Learning, Image Processing, Computer Vision, C++

Signature(s) of Student(s) Signature of PS Faculty

Date: Date:

ABSTRACT

Cancer is the uncontrolled growth of abnormal cells anywhere in a body. These abnormal cells are termed cancer cells, malignant cells, or tumor cells. These cells can infiltrate normal body tissues.

Humans Suffer from over 100 types of cancer. Furthermore, cancer is not limited to humans only, it can affect animals as well.

The advent of medical image digitalization leads to image processing and computer-aided diagnosis systems in numerous clinical applications. These technologies could be used to automatically diagnose patient or serve as second opinion to pathologists.

The dataset used in this report for cancer detection is of cervical cancer. Since a period of two to three decades is needed for cervical cancer to reach an invasive state, the incidence and mortality related to this disease can be significantly reduced through early detection and proper treatment.

ACKNOWLEDGEMENTS

I would like to extend my heartfelt gratitude to all those responsible for providing me internship with the opportunity of a lifetime to work in one of the most premier Research Institutes as CEERI, Pilani.

I would like to take this opportunity to thank the several entities and personalities responsible for providing me with such a platform.

Firstly, I would like to thank **Prof. Shantanu Chaudhury, Director of CSIR-CEERI, Pilani** to have facilitated the collaboration of BITS Pilani University and CEERI and to have been willing to train interns. I am forever indebted for this learning opportunity.

I would like to extend my gratitude to **Mr. Pawan Sharma, coordinator of the PS Programme** for his extremely insightful role in guiding us and for smoothly facilitating the collaboration. I would like to thank the **PS Division of BITS Pilani University** for taking such a brilliant and necessary initiative. This will forever contribute to my work ethic and learning experience for my career.

I would like to thank **Chief Scientist and Group Head Dr. Jagdish Lal Raheja**, Computer Science Department for taking time out of his busy schedule to guide me and enlighten me with my project.

I would also like to thank **Ms. Shalini Nehra (Project fellow)** who has helped me with my queries on the project. He always helped me understand about the research they were doing and how I could fit into their work.

I would like to thank everyone else who was directly or indirectly involved to providing this opportunity to me.

Table of Contents

1. INTRODUCTION

According to the World Health Organization(WHO), cervical cancer is the fourth most frequent cancer in women with an estimated 530,000 new cases in 2012 representing 7.9% of all female cancers. Approximately 90% of the 270,000 recorded deaths resulting from cervical cancer in 2015 occurred in low and middle-income countries. According to Cancer Research UK [2] on survey conducted in 2012 more than 265,000 women are died from cervical cancer across the world. This signifies the need for cheaper and earlier detection of cervical cancer. Particularly in countries where cervical cancer screening programs are not available, diagnosing cervical cancer at an early stage and providing access to effective treatment can significantly improve the likelihood of survival. Owing to the fact that an increasing number of individuals are gaining access to better and more advanced health care and undergoing Pap smear test, and there is a shortage of skilled and experienced pathologists on the other hand, reviewing Pap smear slides has become quite time consuming which ultimately leads to an increase of workload on the pathologists and can also load to fatigue. This can be a potential cause of inaccuracies in diagnosis. In order to solve this problem, an automated detection system of cervical cancer has been proposed.

Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has led many research teams ,from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance.

A variety of these techniques, including Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Even though it is evident that the use of ML methods can improve our understanding of cancer progression, an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice.

For this project we have used dataset of cervical cancer images of colposcopy test.
Tumor is a strange development of cells. There are more than 100 sorts of tumor, including bosom malignancy, skin disease, lung growth, oral tumor and so on. Cervical malignancy shows up as a development in the cervix. Cervical cancer, can be life undermining if not analyzed and treated early. Side effects change contingent upon the sort of malignancy. Tumor treatment may incorporate chemotherapy, radiation, and/or surgery. Early assessment of cervical precancerous injuries can dramatically affect cervical malignancy death rates.

Cervical cancer is a leading cause of mortality and morbidity, which comprises approximately 12% of all cancers in women worldwide according to World Health Organization (WHO). In fact, the annual global statistics of WHO estimated 470 600 new cases and 233 400 deaths from cervical cancer around the year 2000. Therefore early detection and prediction of cancer should play a vital role in the diagnosis process and increase the survival rate of patients.
In this work, we show the way we tried to implement one of the proposed method to detect cancer using image processing and applying SVM on it

2. RELATED WORK

In a research conducted by B.Ashok and Dr.P.Aruna feature selection methods for diagnosis of cervical cancer using SVM classifier have been specified. Although the accuracy in the above-mentioned research reaches 98.5%, their method requires the use of a microscope with an inbuilt camera attached in order to take pictures and feed to the prediction algorithm. The method presented in this research paper requires only the use of a camera, which in today's world is easily accessible to everyone.

D. Kashyap et. al. achieved an accuracy of 95% using SVM classifier and polynomial kernel on images of cancerous and non-cancerous images of cells. This again necessitates the use of a microscope with an inbuilt camera.

The paper by Soumya MK et. al. proposes to classify cervical cancer images into different stages based on the treatment volume that the particular patient requires but for that it is required for them to know that the patient is affected with cervical cancer which is proposed in our paper.

The paper titled New Features of Cervical Cells for Cervical Cancer Diagnostic System Using Neural Network, by Mustafa et. al. states that though Pap test is the most popular and effective test for cervical cancer, Pap test does not always produce good diagnostic performance.

The paper titled Image Texture Feature Extraction using GLCM Approach by P. Mohaniah et. al. mentions that as the size (dimension) of the image increases, the value of texture features extracted from them also increases. So, this paper proposes an optimal size of the image i.e. 128x128 for feature extraction for better resolution and minimum loss of generality.

S. Kaaviya et. al. proposes a new approach to identify abnormal cervical cells using the area of the nucleus as the feature for classification.

K.Pradeep Chandran ,et al. presents a segmentation method, spatial fuzzy clustering algorithm, for segmenting Magnetic Resonance images to detect the Cancer in its early stages anatomical structures. Here a Probabilistic Neural Network with radial basis function is employed to implement an automated Tumor classification.

Priyanka K Malli and Dr. Suvarna Nandyal proposed an automated, comprehensive machine learning for the detection of cervical cancer. Using the color and shape features of nucleus and cytoplasm of the segmented unit of the cervix cell, they propose to train a k-NN and an Artificial Neural Network(ANN). The above mentioned approach has shown an accuracy of 88.04% for KNN and 54% for ANN. Our method gives an accuracy of 96.67% in classifying images into cancerous and non-cancerous.

Chankong et. al. propose a method that utilizes a set of simple features extracted from the two- dimensional Fourier transform of the cell images in order to avoid the problem of cell and nucleus segmentation. The features used are calculated based on the mean, variance, and entropy obtained from the frequency components along the circle of radius r centered at the center of the spectrum and the frequency components along the radial line at an angle θ. The above-mentioned approach achieves an accuracy of 92% classification rate on a set of 276 cervical single cell images containing 138 normal cells and 138 abnormal cells.

Setu Garg et. al. has proposed the use of edge detection and hybrid segmentation to extract the tumor infected area.

Sajeena T A et. al. proposes a method of cervical cancer detection using Radiation Gradient Vector Flow segmentation and SVM and artificial neural networks.

3. OVERVIEW OF THE PROJECT

a. Problem statement

Develop a machine learning algorithm in C++ that detects Cancer. Given any image the algorithm should classify the image as Cancerous and non-Cancerous.

b. Proposed Solution Statement

Following are the solutions proposed in the review paper published by Santhosh B.[1]

Author	Technique	Advantages
Lalit Gupta	A new feature selection and classification scheme for screening of oral cancer using laser induced fluorescence	Increase discrimination ability of the feature vectors Sensitivity : Above 95% and specificity : Above 99%
Sebastian Steger	novel image feature extraction approach	Oral Cancer Prediction Automatically
M. Muthu Rama Krishnan	Wavelet based textureclassificati on of oral histopathological sections	Improves the Accuracy
Ranjan Rashmi	A novel wavelet neural network based pathological stage detection technique	ProtectionAuthentication
K Anu Radha	Detection of Oral Tumor based on Marker – controlled Watershed Algorithm	Accurecy is more than 90% the quality of the image is enhanced using linear contrast stretching.
Woonggyu Jung	Optical Coherence Tomography	Accurecy is more detected oral cancer in 3-D volume images of normal and precancerous lesions
Neha Sharma	Apriori Algorithm	the highest confidence level, thereby, making them very useful for early detection and prevention of oral cancer
Anuradha	Statistical Feature Extraction to Classify Oral Cancers. Transform	The proposed system segments and classifies oral cancers at an earlier stage.
Anuradha	Oral Cancer Detection Using Improved Segmentation Algorithm	Better speed and accuracy

c. Solution Implemented in the project

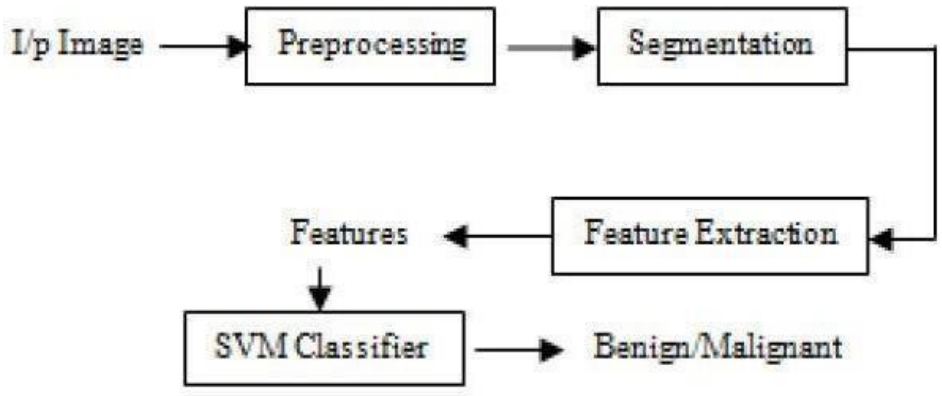
To detect cancer, we first obtain 2 datasets (1 cancerous and 1 non-cancerous. As a part of the next step i.e. Feature Extraction we use the GLCM algorithm and extract the Haralick Features as proposed by Haralick [2]. Once all the features are extracted from the dataset, we segregate the dataset randomly in the following 2 parts

1. Training Dataset (70% of dataset)
2. Test Dataset (30% of dataset)

This marks the end of the feature extraction process
For the purpose of detection we use the SVM Classifier. We then train the SVM Classifier using training dataset and then test the dataset using test dataset
Please note that we first perform the above procedure on MATLAB and then implement on C++. For C++ we have used the OpenCV library for both the image processing and for the SVM classifier.

d. Basic Workflow

Task 1: Obtain a suitable algorithm that detects cancer
Task 2: Confirm the accuracy of the above algorithm on MATLAB.
Task 3: Once satisfactory accuracy is achieved implement the above algorithm in C++



4. IMAGE PROCESSING

a. What is an Image?

An image is nothing more than a two-dimensional signal. It is defined by the mathematical function f (x, y) where x and y are the two co-ordinates horizontally and vertically. The value off (x,y) at any point is gives the pixel value at that point of

128	30	123
232	123	321
123	77	89
80	255	255

an image.

The above figure is an example of digital image that you are now viewing on your computer screen. But actually, this image is nothing but a two-dimensional array of numbers ranging between 0 and 255.

Each number represents the value of the function f(x,y) at any point.In this case the value 128, 230 ,123 each represents an individual pixel value. The dimensions of the picture is actually the dimensions of this two dimensional array.

b. Some Image processing Terminologies-

Bitmap Image

An image composed of black and white pixels.
Brightness- Determines the intensity of the color presented by a pixel in a color image, or the shade of grey presented by a pixel in a greyscale image.

Contrast

The difference between the lightest and darkest regions of an image.

Digital Image

An image captured by an imaging device and represented in a computer as a rectangular grid of pixels.

Greyscale Image

An image composed of pixels that present shades of grey.

Histogram

The histogram of an image visualizes the distribution of the brightness in the image by plotting the number of occurrences of each brightness.

Histogram Equalization

An image-processing technique that reveals detail hidden in images with a poorly-distributed range of brightnesses.

Image

An image records a visual snapshot of the world around us.

Image Processing

The field of computer science that develops techniques for enhancing digital images to make them more enjoyable to look at, and easier to analyze by computers as well as humans.

Kernel

A rectangular grid of convolution weights.

Pixel

A square unit of visual information that represents a tiny part of a digital image.

Pixel Depth

The number of colors or shades of grey a pixel can present. Bitmap pixels have depth two, typical greyscale pixels have depth 256, and typical color pixels have depth 16,777,216.

Primary Colors

The colors red ,green and blue from which all other colors in the RGB color model are mixed.

Resolution

The number of pixels available to represent the details of the subject of a digital image.

RGB

A color model that represents each color with three numbers that specify the amounts of red (R), green (G) and blue (B) that produce the color.

c. Uses of image processing

Image processing techniques provide a good quality tool for improving the manual analysis. Image processing techniques are used in several areas such as military, space research, medical and many more. In this proposed system image processing techniques are used for image improvement in earlier detection and treatment stages. Image quality assessments as well as improvementaredependingontheenhancementstagewherepre-processingtechniqueis used based on principal component analysis and Histogram Equalization. Classification is very important part of digital image analysis. It is computational procedure that sort images in to groups according to their similarities.

d. Library Used for Image Processing

OpenCV- C++ is a widely used and well documented programming language with good performance. It is quite suited for image processing. During the 90's when the common computer became more and more powerful, and digital images started rivalling the analogue camera, Image processing started becoming more useful. And so whole libraries of functions and algorithms were created to aid anyone operating in the field. There are many such computer vision libraries available, some for free and some for a fee, but one of the most popular library is the OpenCV.

With more than 47 thousand people of user community and estimated number of downloads exceeding 14 million, this library is used extensively in companies, research groups and by governmental bodies. It started when a team at Intel decided to create a computer vision library back in 1999 and has grown ever since. As is common in an open source library it has many authors contributing to its development. OpenCV contains different modules that are dedicated to different areas of computer vision. The images are stored in matrices defined by the Mat class, and one very helpful feature of Mat is its dynamic memory management, which allocates and deallocates memory automatically when image data is loaded or goes out of scope respectively.

Also, if an image is assigned to other variables, the memory content remains the same, as the new variables gets passed onto them. This can save resources since algorithms often process smaller parts of the image at a time, and the data need not be copied everytime. This requires OpenCV to keep count of how many variables are using the same block of memory as to not deallocate it the moment one of the variables are deleted. Images can be represented with different data types such as 8-bit or 16-bit or 32-bit integers, and floating points values in both single (32-bit) and double precision (64-bit).

We have used the 3.4.1 version of OpenCV which was the latest as of June 2018.

5. GLCM (Gray level Co-occurrence Matrix) Algorithm for feature extraction

The Grey Level Co-occurrence Matrix, GLCM (also rarely called the Grey Tone Spatial Dependency Matrix)
Definition: The GLCM is a tabulation of how often different combinations of pixel brightness values (grey levels) occur in an image.
The GLCM described here is used for a series of "second order" texture calculations. Second order means they consider the relationship between groups of two pixels in the original image. *Notice that this is not the same thing as "second order equations" which would mean equations with some variables squared.*

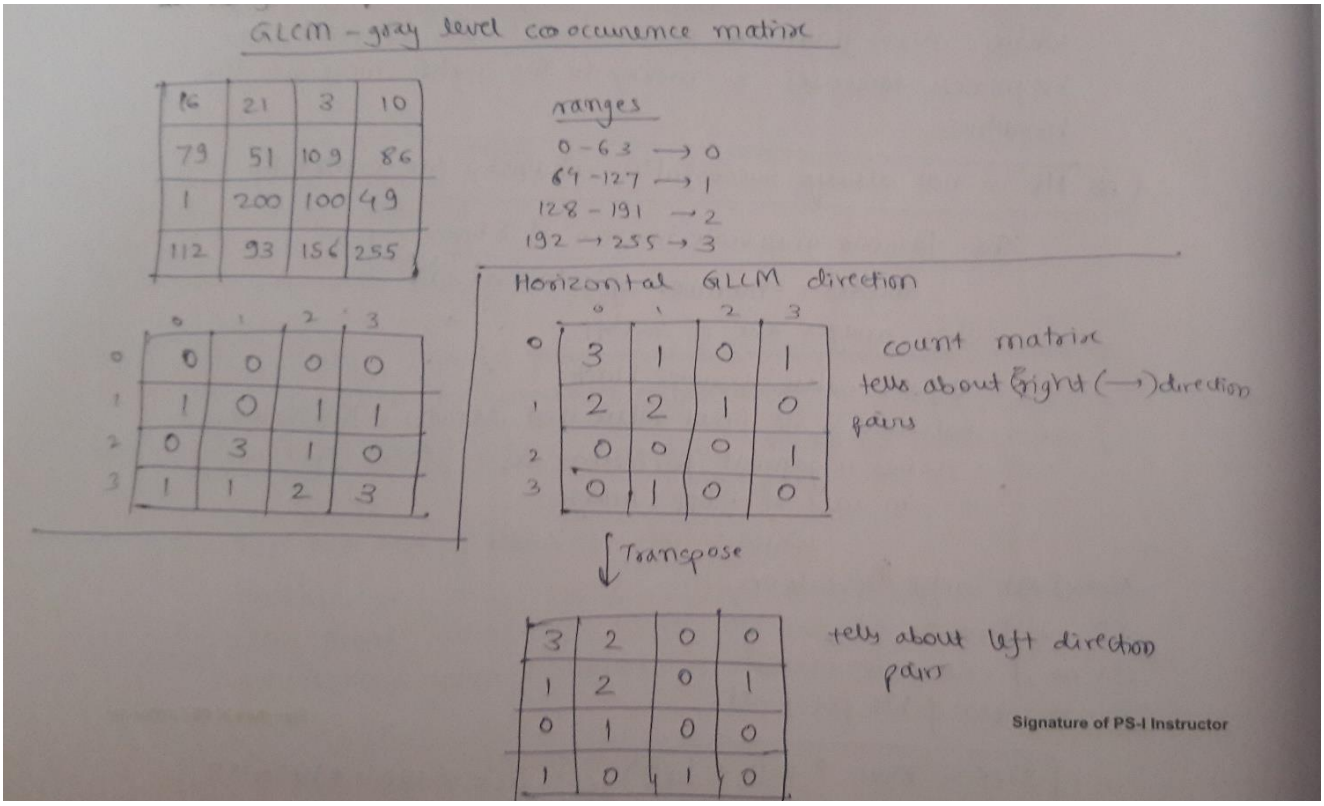
a. Framework for the GLCM:

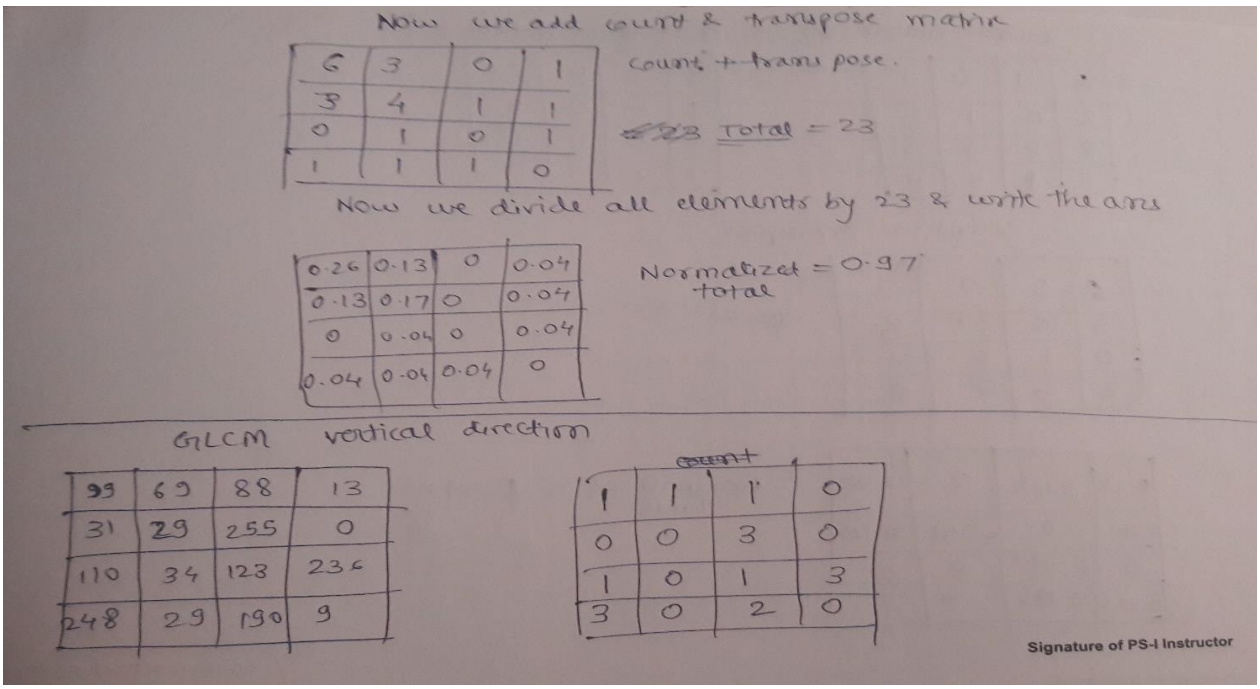
Spatial relationship between two pixels:
GLCM texture considers the relation between two pixels at a time, called the reference and the neighbour pixel. In the illustration below, the neighbour pixel is chosen to be the one to the east (right) of each reference pixel. This can also be expressed as a (1,0) relation: 1 pixel in the x direction, 0 pixels in the y direction. Each pixel within the window becomes the reference pixel in turn, starting in the upper left corner and proceeding to the lower right. Pixels along the right edge have no righthand neighbour, so they are not used for this count. The illustration below shows one such relationship: the pixel value shown in red are reference pixels and the pixels shown in blue are neighbour pixels in a (1,0) relationship to their reference pixel. This shows examples only; all pixels can serve as reference and neighbour pixels. If you are about to object that there is a problem with the left and right edges, you are right.

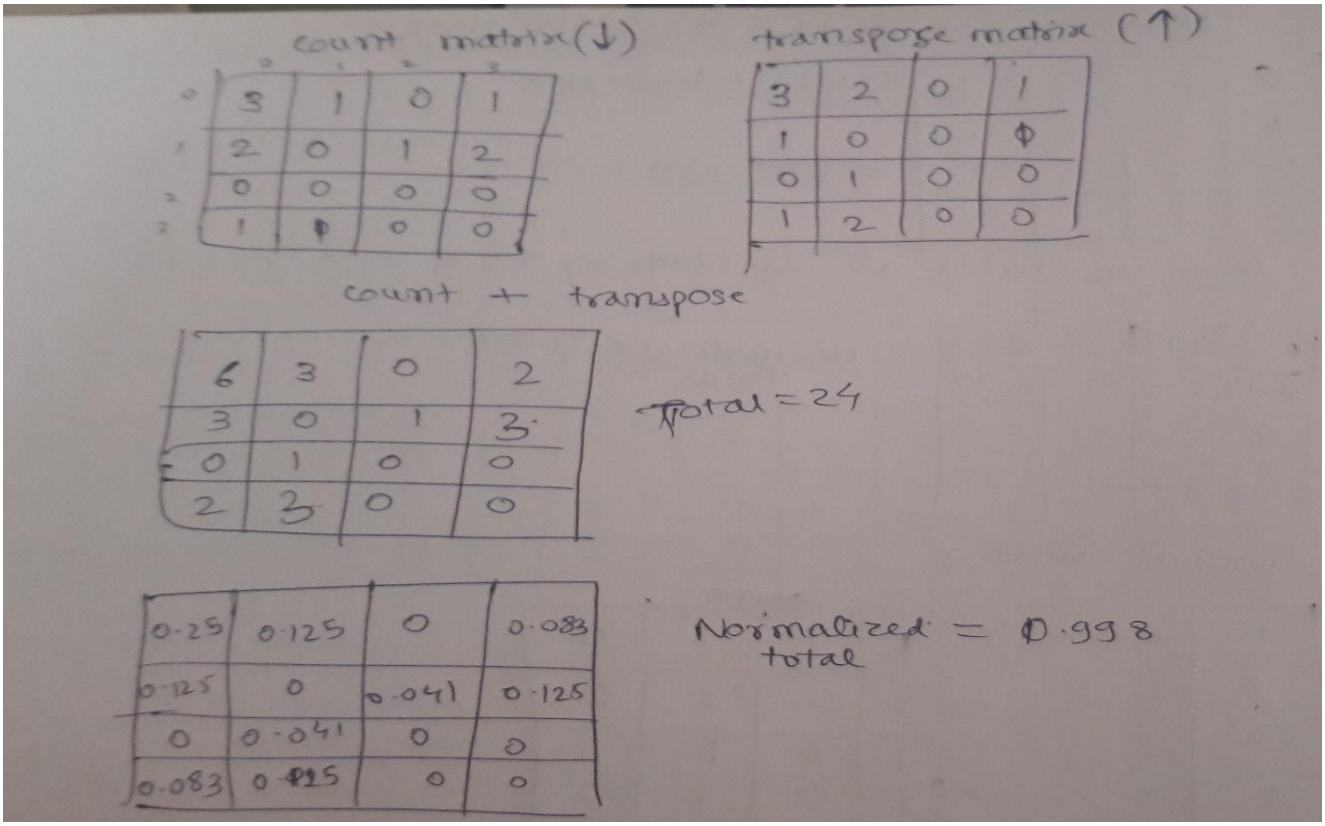
If the window is large enough, using a larger offset is perfectly possible. There is no difference in the counting method. The sum of all the entries in the GLCM (i.e. the number of pixel combinations) will just be smaller for a given window size.
Now we will look at how the GLCM matrix is actually constructed, by counting and tabulating the number of pixel pairs that show a combination of all possible GLCM value pairs. Let's take the first pair of reference and neighbour pixels as shown above, in red (reference) and blue (neighbour). Since the reference pixel has the GL value of 0, and the neighbour pixel the value of 1, we count this as one entry in a table of frequencies.
The top left cell will be filled with the number of times the combination 0,0 occurs, i.e. how many times within the image area a pixel with grey level 0 (neighbour pixel) falls to the right of another pixel with grey level 0 (reference pixel). Each cell is read in this pattern with appropriate changes in

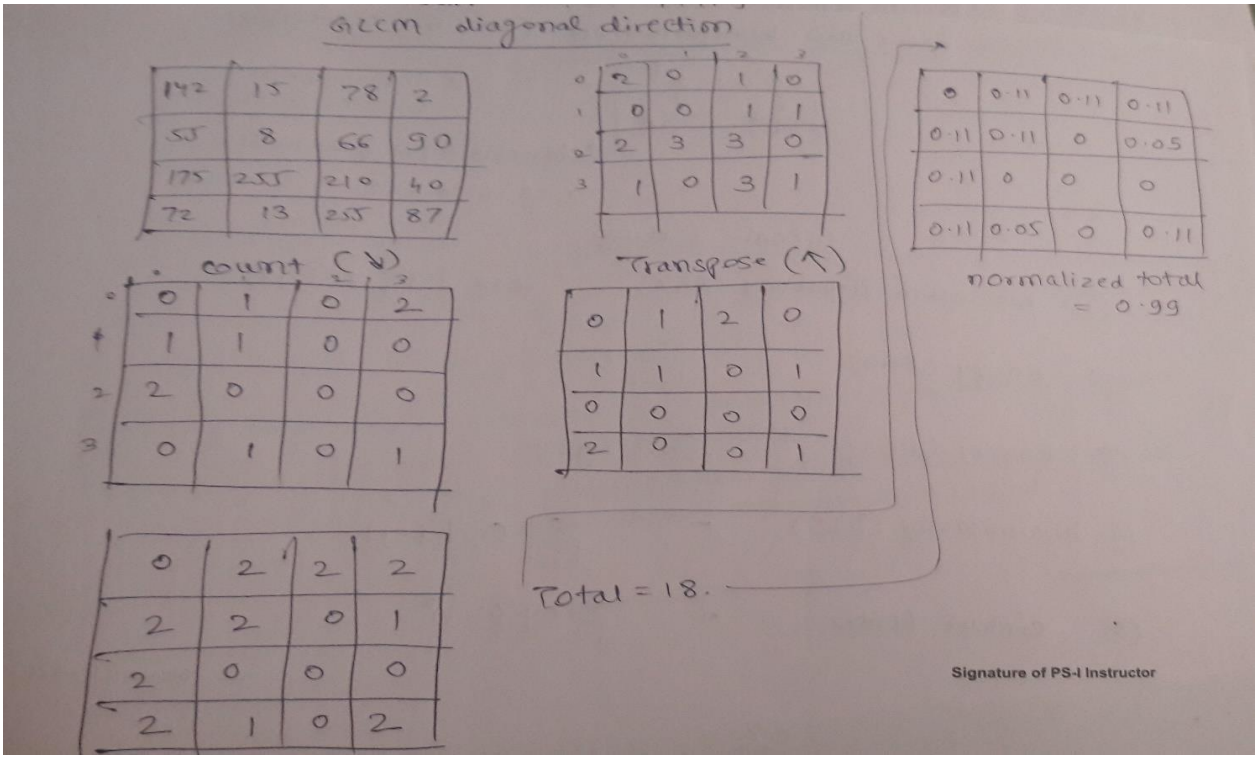
numbers.

A different co-occurrence matrix exists for each spatial relationship. The one produced above was for the (1,0) relationship: a pixel and its neighbour to the right (east). Other possible relationships are above (0,1), next to the west (-1,0), diagonal (1,1) or (-1,-1).






count matrix (↓)
transpose matrix (↑)
count + transpose
Total = 24
Normalized total = 0.998


GLCM diagonal direction
count (↓)
Transpose (↑)
Total = 18.
normalized total = 0.99
Signature of PS-I Instructor

b. Properties of GLCM

Now that we know how to construct the GLCM we can list some generalizations about it that will help us to go to the next step.

1. It is square:

The reference pixels have the same range of possible values as the neighbour pixels, so the values along the top are identical to the values along the side.

2. It has the same number of rows and columns as the quantization level of the image:

The test image is 2-bit has four (2_2) grey level values (0,1,2 and 3). Eight bit data has 256 possible values (2^8), so would yield a 256 x 256 square matrix, with 65,536 cells. 16 bit data would give a matrix of size 65536 x 65536 = 429,496,720 cells!

FAQ: Isn't that too much to handle, even for a computer?

Yes - even for 8-bit data. Most operational programs rescale the image values into 4 bit (16 x 16 matrix with 256 cells) or 5 bit (32x32 matrix with 1024 cells). The rescaling algorithms vary from one software to another, and are usually proprietary, meaning they do not say precisely how they do it.

Until about 2007, almost all image data was 8-bit. Now image data is often in 16-bit or some other value. The same problem exists, and the same solution!

There is another reason for compressing the data into 4 or 5 bit. If all 256 x 256 (or more) cells were used, there would be many cells filled with 0's (because that combination of grey levels simply does not occur on the image). The GLCM approximates the joint probability distribution of two pixels. Having many 0's in cells makes this a very bad approximation. If the number of grey levels

is reduced, the number of 0's is reduced, and the statistical validity is greatly improved. Because users often have no choice (unless writing their own algorithms), the question of the effects of quantization level is often overlooked. In practice, some statistics calculated from the GLCM don't help classification very much when a large number of grey levels are used. Other statistics don't degrade as much.

3. We want the GLCM to be symmetrical around the diagonal:

A symmetrical matrix means that the same values occur in cells on opposite sides of the diagonal. For example, the value in cell 3,2 would be the same as the value in cell 2,3. The matrix we calculated above is not symmetrical. However, texture calculations are best performed on a symmetrical matrix.

The matrix above counted each reference pixel with the neighbor to its right (east). If counting is done this way, using one direction only, then the number of times the combination 2,3 occurs is not the same as the number of times the combination 3,2 occurs (for example 3 may be to the right of 2 three times, but to the left of 2 only once). However, symmetry will be achieved if each pixel pair is counted twice: once "forwards" and once "backwards" (interchanging reference and neighbour pixels for the second count).

Example: A reference pixel of 3and its neighbour of 2 would contribute one count to the matrix element 3,2 and one count to the matrix element 2,3.
Symmetry also means that when considering an eastern (1,0) relation, a western (-1,0) relation is also counted. This could now be called a "**horizontal**" matrix.
Making a matrix symmetrical in this way also neatly gets us around the problem of the window edge pixels: remember the ones on the left could never be a neighbour pixel in an east relationship, and the ones on the right could never be a reference pixel in an east relationship. But in a horizontal relationship, making the symmetrical matrix, each pixel gets to be a reference and neighbour pixel, no matter where it is in the window.

c. Expressing the GLCM as a probability:

Is it more likely to find a horizontal combination of, say, 2,2 in the original image, or is 2,3 more likely? Looking at the horizontal GLCM shows that the combination 2,2 occurs 6 times out of the 24 horizontal combinations of pixels in the image (12 eastern + 12 western). In other words, 6 is the entry in the horizontal GLCM in the third column(reference pixel value 2)and third row (neighbour pixel value 2). The simplest definition of the probability of a given outcome is
"the number of times this outcome occurs, divided by the total number of possible outcomes."

6. Haralick Features

Haralick Features describe the correlation in intensity of pixels that are next to each other in space. Haralick proposed fourteen measures of textural features which are derived from the co-occurrence matrix a well known statistical technique for texture feature extraction. It contains information about how image intensities in pixels with a certain position in relation to each other occur together. Texture is one of the most important defining characteristics of an image. The grey level co-occurrence matrix is the two dimensional matrix of joint probabilities P(i,j) between pairs of pixels separated by a distance ‘ d’ in a given direction ‘r ’. The second order image histogram referred to as the Grey Level Co-occurrence Matrix (GLCM) of an image offers greater information about the inter-pixel relationship, periodicity and spatial grey level dependencies. This matrix is a source of fourteen texture descriptors.

Following some Haralick features that we have used (as they were proposed in the research paper that we are implementing)-

Power law transformation (Gamma correction)

$s = c r^{\gamma}$

s → output, r → input

c = constant = 1 (mostly)

γ → tunable input parameter

r → 0 to 255 ie L = 256.

GLCM features

(1) Maximum Probability (MAX) $\max \{ C_{ij} \forall (i,j) \}$

(2) Energy (ENG) $\sum_{i,j=1}^{G} C_{ij}^2$

(3) Entropy (ENT) $- \sum_{i,j=1}^{G} C_{ij} \log C_{ij}$

(4) Dissimilarity (DIS) $- \sum_{i,j=1}^{G} C_{ij} |i-j|$

(5) Contrast (CON) $- \sum_{i,j=1}^{G} C_{ij} (i-j)^2$

(6) Homogenity (HOM) $\sum_{i,j=1}^{G} \frac{1}{1+|i-j|} C_{ij}$

(7) Inverse difference moment (IDM) $\sum_{i,j=1}^{G} \frac{1}{1+(i-j)^2/G^2} C_{ij}$

(8) Correlation (COR) $\sum_{i,j=1}^{G} \frac{(i-\mu_i)(j-\mu_j)}{\sigma_i \sigma_j} C_{ij}$

where $\mu_i = \sum_{i}^{G} i \sum_{j}^{G} C_{ij}$

$\mu_j = \sum_{j}^{G} j \sum_{i}^{G} C_{ij}$

$\sigma_i = \sum_{i}^{G} (i-\mu_i)^2 \sum_{j}^{G} C_{ij}$

$\sigma_j = \sum_{j}^{G} (j-\mu_j)^2 \sum_{i}^{G} C_{ij}$

Using these features we classify images into 2 classes - cancerous & non cancerous.

Usually all these parameters are found to be high for cancerous images as compared to non-cancerous images.

Signature of PS-I Instruc

7. SVM (Support Vector Machines)

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high-or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

SVMs can be used to solve various real -world problems of Uncertainty in Knowledge-Based Systems:

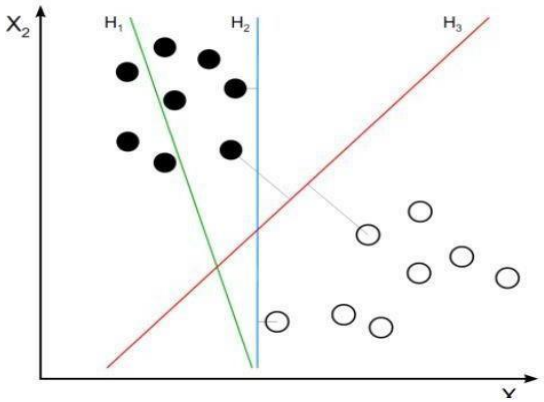
1. SVMs are helpful in text and hypertext categorization as their application can significantly reduce the need for labeled training instances in both the standard inductive and transudative settings.

2. Classification of images can also be performed using SVMs. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback. This is also true of image segmentation systems, including those using a modified version SVM that uses the privileged approach.

3. Hand-written characters can be recognized using SVM.

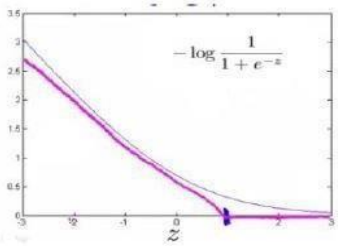
4. The SVM algorithm has been widely applied in the biological and other sciences. They have been used to classify proteins with up to 90% of the compounds classified correctly. Permutation tests based on SVM weights have been suggested as a mechanism for interpretation of SVM models. Support vector machine weights have also been used to interpret SVM models in the past. Post hoc interpretation of support vector machine models in order to identify features used by the model to make predictions is a relatively new area of research with special significance in the biological sciences.

SVM are also known as large margin classifiers, since they will find the boundary with the largest separation between classes. In the figure below, SVM will find H3 as the decision boundary instead of H2.

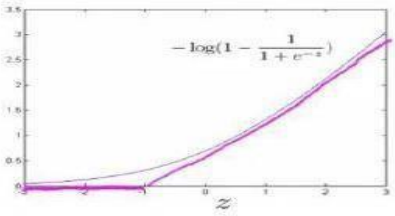


To build our SVM, we have to redefine our cost functions

- When y=1, instead of a curved line create two straight lines (magenta) which acts as an approximation to the logistic regression y = 1 function. Let it be called cost1 function.



- When y = 0, Do the equivalent with the y=0 function plot. Let it be called cost0 function.

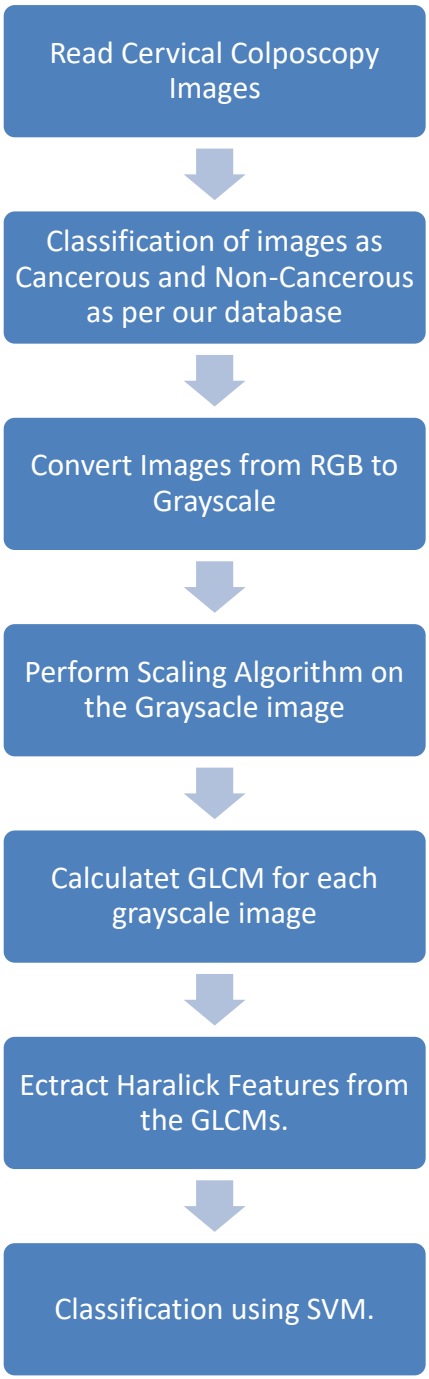


For Support Vector Machines we will use the improvised cost function of logistic regression where the sigmoid function will be replaced by the defined cost functions. We will also get rid of the regularisation term and the (1/m) term and another coefficient C. The cost function for SVM is

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$

8. BASIC WORKFLOW

Algorithm flow of cervical cancer detection method has been shown in figure 1. It includes six steps.



The presented methodology of automated classification of colposcopy cervix images starts with the collection of colposcopy images collected from freely available images on the image was constructed. Once the GLCMs were available we extracted the 13 Haralick features for each GLCM. The SVM classifier was then trained using the 13 features extracted for each image in the training dataset. After training the SVM Model, the images present in the testing dataset were classified by using the above mentioned SVM Model. After predicting the classes on the test dataset, performance analysis was done. The report of the performance analysis is internet. The dataset thus obtained was divided into 2 parts
Training Dataset = 70%of original dataset
Testing Dataset = 30%of original datasets

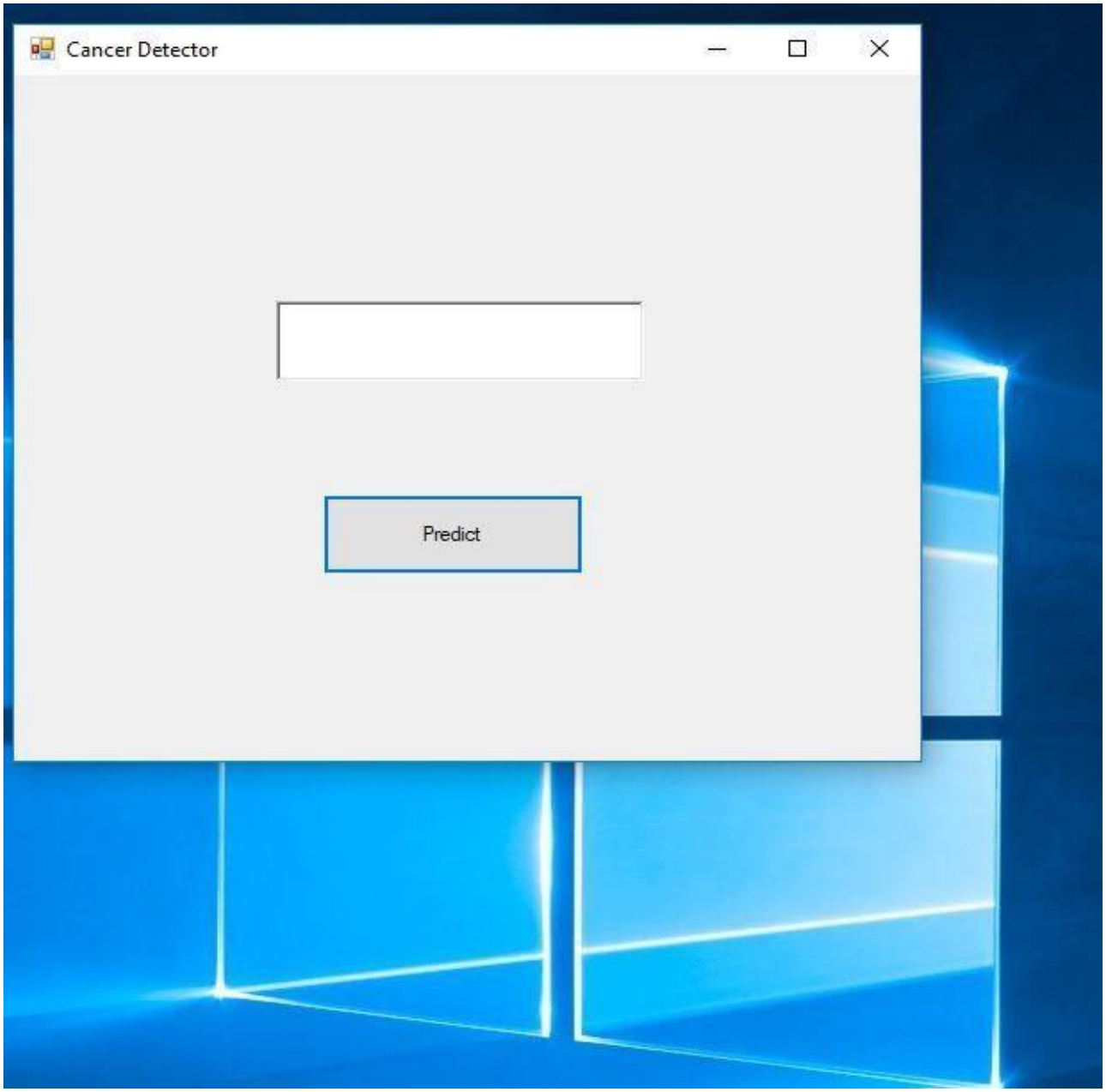
The algorithm works as follows: -
The cervical cancer colposcopy test images were classified and labelled accordingly.
After segregating the images, the images were read by the program
The basic pre-requisite for calculating the GLCM is the presence of only one channeled images as input. This was done by converting the 3- channeled RGB images to the 1- channeled grayscale image. The pixel intensities of the images thus obtained was scaled down from 0-255 to 0-7 uniformly.
The GLCM was calculated using Θ=0 and radius=1
All the second order texture features mentioned by Haralick et. Al [13] were extracted using the GLCM obtained in the above step
SVM was trained and applied to the dataset

First a labeled dataset of images was obtained. The images in the dataset were labelled as cancerous or non-cancerous. These images were given as input to the program which first performed histogram equalization of each of the input images. All the images were then converted to grayscale in order to calculate the GLCM. The transformation of a sample image from an input image to the grayscale image is shown in Fig 2.

The pixel intensities present in the grayscale image were integers ranging from 0-255 (including both values). We then created a mapping from the domain [0,255] to the co-domain [0-7] which distributed the range of the domain uniformly to the range of the co-domain. After applying the above-mentioned mapping to each image in the dataset, the GLCM of each mentioned in experimental results.

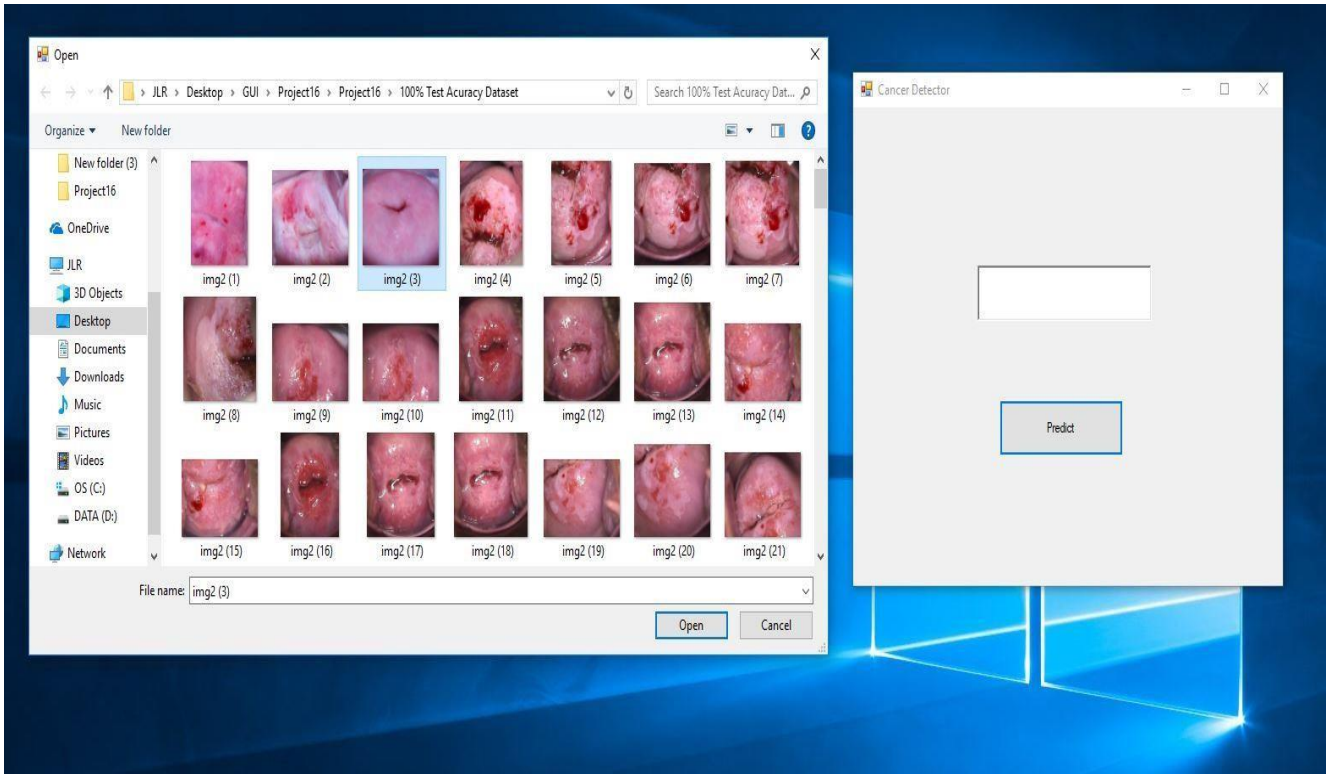
9. GUI DEVELOPMENT

a. The following window opens when the program is executed. It is the basic window form that we created. It contains a simple Text Area and a button. When the button is pressed a browser window opens.



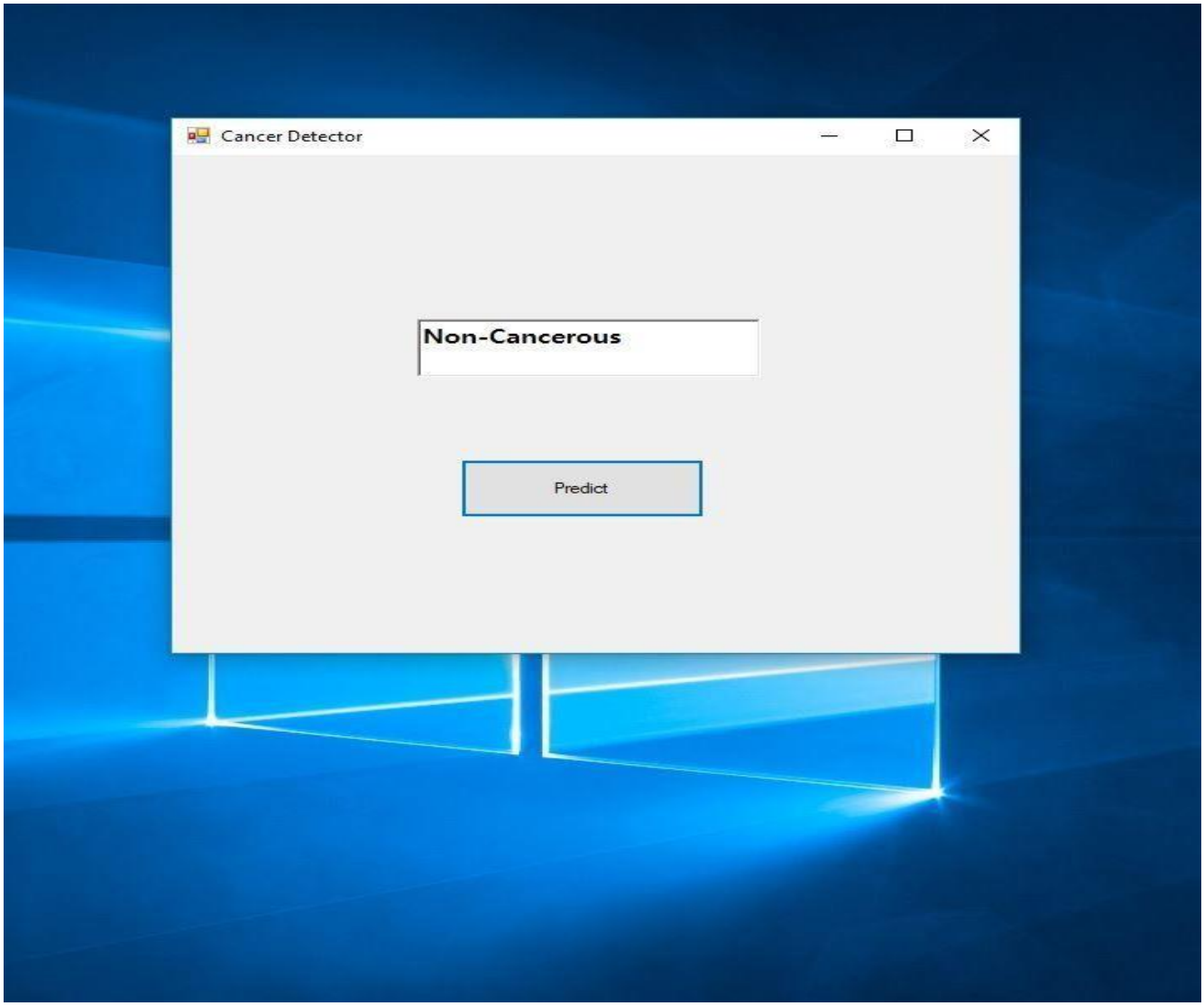
b. Window for selecting Input Image

Following is the window that opens when the predict button is pressed. In the window shown below, a non-cancerous image has been selected as input. Once the button labeled as open is selected, the algorithm predicts the answer in the text area present in the basic windows form



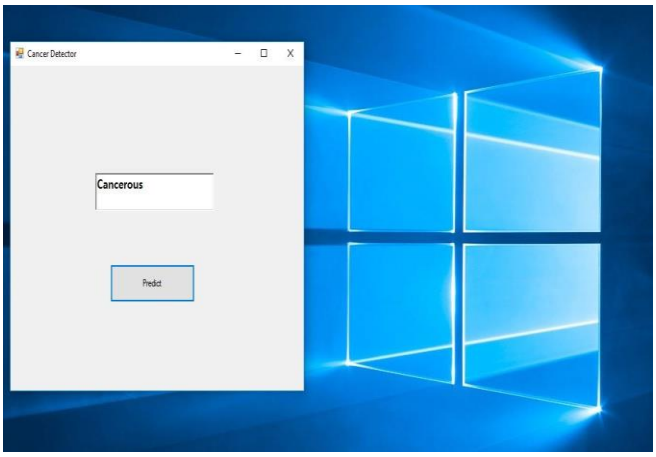
c. Final Output Displayed

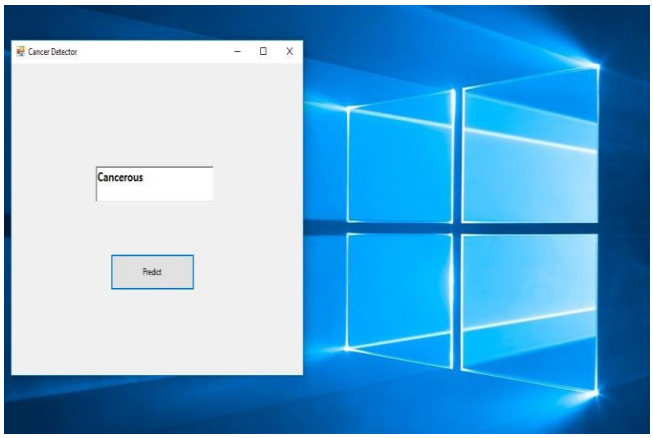
The selected image was predicted as non-cancerous and is displayed in the text box.

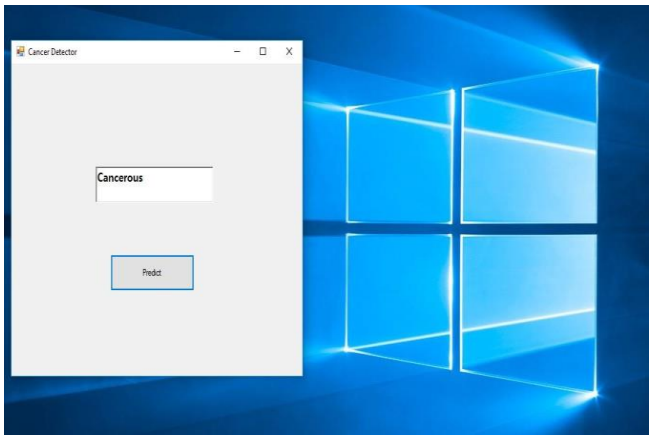


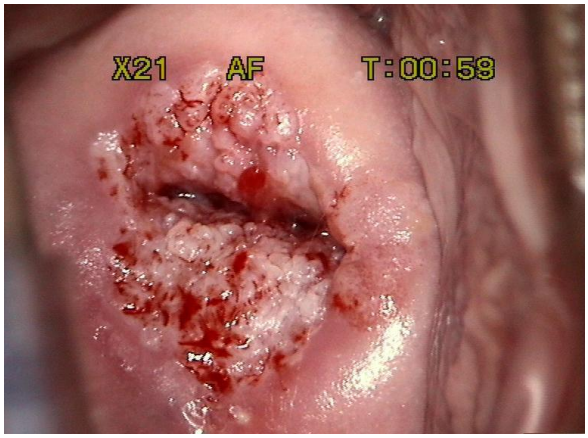
10.EXPERIMENTAL RESULTS

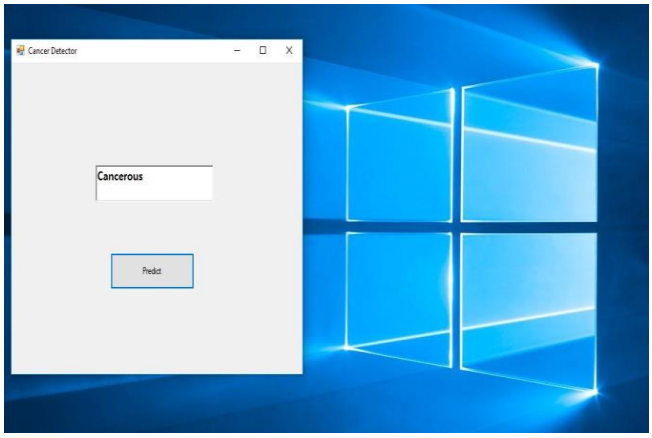
Following are few sample images and the outputs that we got using our code.

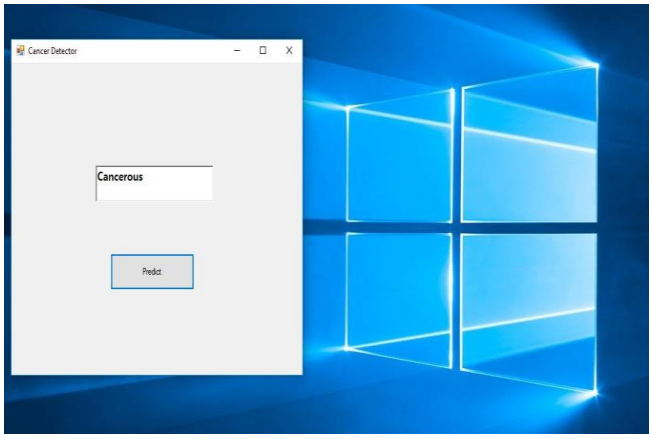


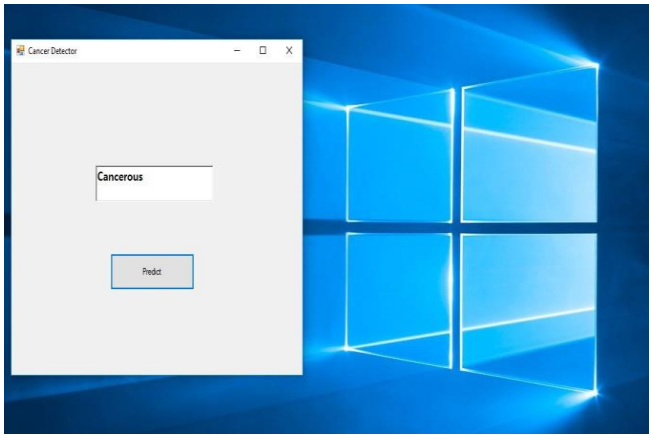


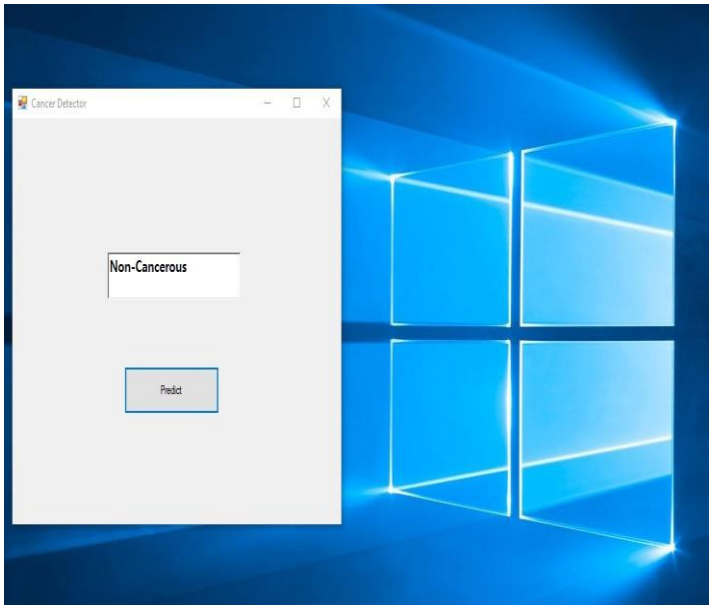


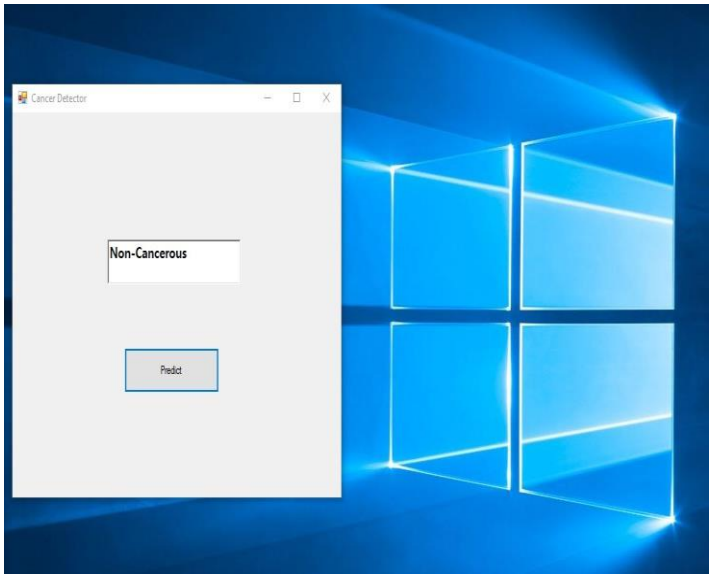

X21 AF T:00:59

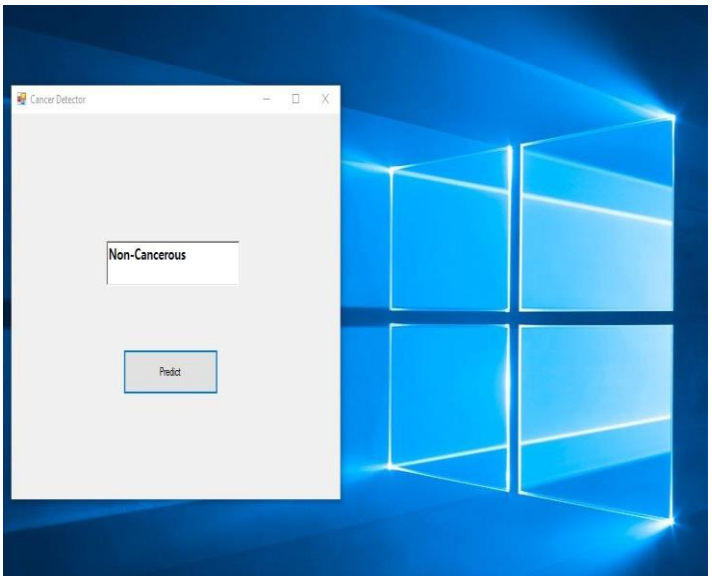

Cancer Detector
Cancerous
Predict

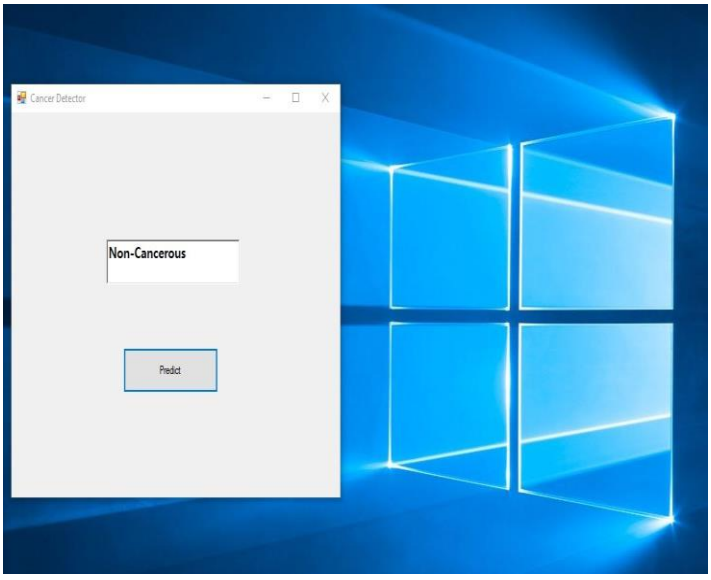

Cancer Detector
Cancerous
Predict

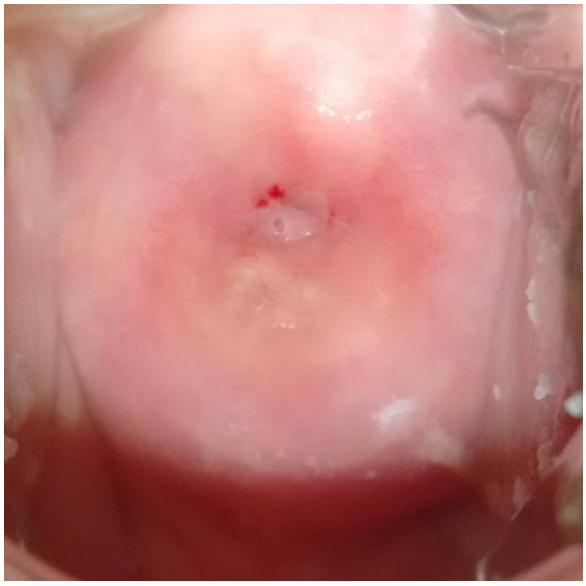

Cancer Detector
Cancerous
Predict

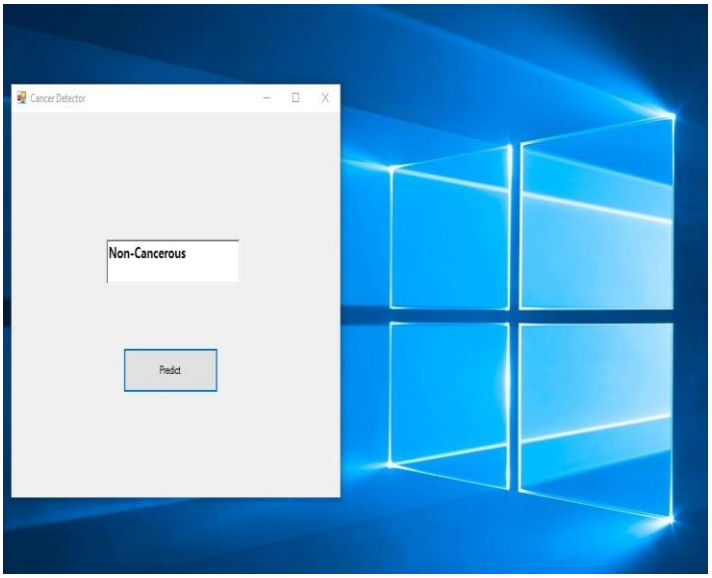

Cancer Detector
Non-Cancerous
Predict

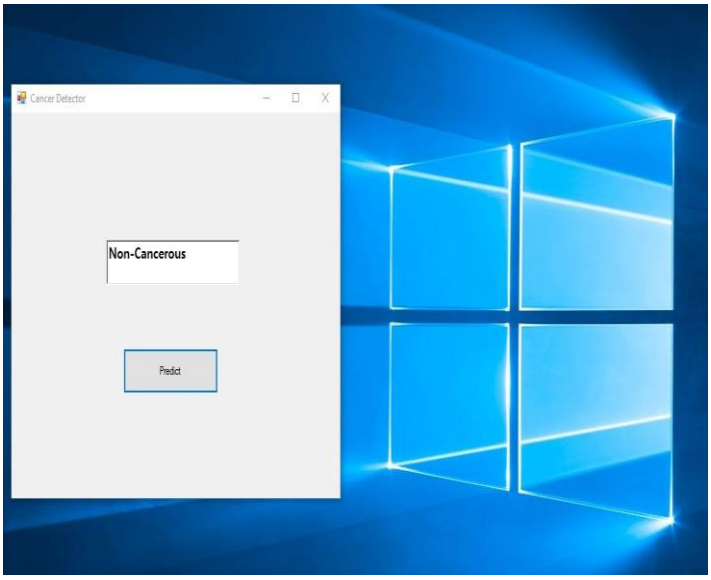

Cancer Detector
Non-Cancerous
Predict


Cancer Detector
Non-Cancerous
Predict


Cancer Detector
Non-Cancerous
Predict

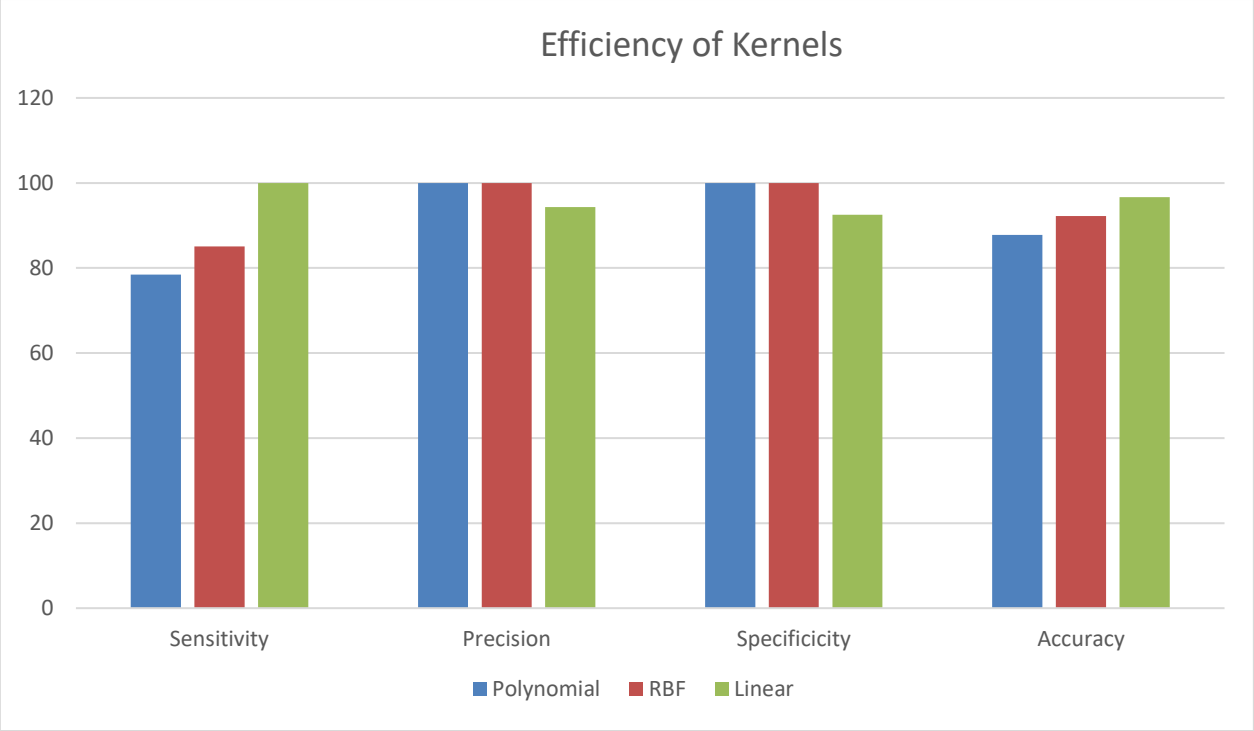



Cancer Detector
Non-Cancerous
Predict


Cancer Detector
Non-Cancerous
Predict

We extracted the Haralick features form the normalized GLCM calculated for each image. Later we calculated the accuracy, specificity, precision and sensitivity for SVM classifier with 3 different kernel functions. The data thus obtained is tabulated below.

	Sensitivity	Precision	Specificity	Accuracy
Linear	78.43%	100%	100%	87.78%
RBF	85.11%	100%	100%	92.22%
Polynomial	100%	94.34%	92.5%	96.67%



The following are the confusion matrices obtained for the individual kernel functions.

Table 2(a)-Confusion matrix for linear kernel function,
(b)-Confusion matrix for RBF kernel function,
(c)-Confusion matrix for polynomial kernel function

(a)Using Linear Kernel function		Predicted	
		Negative	Positive
Actual	Negative	55.56%	0%
	Positive	3.33%	41.11%

(b)Using RBF Kernel function		Predicted	
		Negative	Positive
Actual	Negative	47.78%	7.78%
	Positive	0%	44.44%

(c)Using Polynomial Kernel function		Predicted	
		Negative	Positive
Actual	Negative	43.33%	12.22%
	Positive	0%	44.44%

We need to search for cancerous image, keeping this in mind a positive finding is defined as an image with non-cancerous cervix. A false negative finding is defined as a cancerous image misclassified as non-cancerous Hence it becomes imperative for the SVM Model to minimize the number of false negative findings. The Tables 2 (a), (b), (c) show that the percentage of images classified as false positive are 3.33%, 0%, 0% respectively. These values are clearly small irrespective of the kernel used. One possible method of lowering the chances of a false negative finding is by taking more number of images of the target cervix in order to detect cancer. On the similar lines, a false positive finding is said to be a non-cancerous cervix classified as cancerous. If a false positive finding occurs, the patient can be recalled for the test. In our case the percentage of the false positive findings for the linear kernel, RBF kernel and polynomial kernel are 0%, 7.78%, 12.22% respectively. It can be observed that the percentage of false positive finding is least for the linear kernel.

Original Accuracy	**96.67%**
Feature Removed	**Accuracy Obtained**
Correlation	93.55%
Energy	96.67%
Variance	91.94%
Contrast	93.55%
Homogeneity	93.55%
Sum Average	96.67%
Sum Variance	91.94%
Sum Entropy	93.55%
Entropy	96.67%
Difference Variance	95.16%
Difference Entropy	93.55%
Information Measure of Correlation I	95.16%
Information Measure of Correlation II	93.55%

After achieving an accuracy of 96.67% using the linear kernel function and 13 Haralick features, we selectively removed one feature at a time and trained the SVM model. The trained model was then tested on the same testing dataset as before and we obtained the above-mentioned table of accuracy. The above table suggests that removing entropy and energy and sum average does not affect the accuracy while using the SVM kernel.

11.CONCLUSION

Final Testing

With the SVM trained model thus obtained after performing features extraction, the testing accuracy varies from 98% to 100% on randomly selected cancerous and non-cancerous images. Cervical cancer is a disease that affects millions of women every year. In this paper, we have proposed and automated cervical cancer detection technique. The paper involves creating the GLCM and extracting the aforementioned Haralick Texture Features. 3 different types of kernel functions were used for classification using the SVM classifier. Our analysis showed that the linear SVM classifier gave the highest accuracy of 96.67%.

Problem Remaining

The training time for the program developed is a bit high i.e. the program takes a high time to train the algorithm, this is because the image size of the images in the training image dataset is around 36 MB whereas general size of images should be around 1-2 MB

12.FUTURE SCOPE

Future work may involve image segmentation for segregating useful information from the image and implementing principal component analysis(PCA) for dimensionality reduction. Furthermore an additional glare removal algorithm can also be applied on the image for noise removal of the images. Lange et. al. [11] propose an automated method to remove glare in reflectance imagery of the uterine cervix.

13. REFERENCES

1. R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” IEEE Transactions on Systems, Man and Cybernetics, vol. 3, no. 6, pp. 610–621, 1973.

2. Santhosh B. (2016) "Review on Emerging Techinques to Detect Oral Cancer" International Journal of Electrical Sciences & Engineering (IJESE) Volume 1, Issue 1; January 2016 pp. 41-46

3. K. Anuradha, K. Sankaranarayanan “Oral Cancer Detection Using Improved Segmentation Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 1, January 2015

4. http://www. webmd.com/oral-health/guide/oral-cancer

5. https://in.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html

6. https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

7. https://in.mathworks.com/help/images/ref/graycomatrix.html

8. Ashok, B. and Dr. P. Aruna. “Comparison of Feature selection methods for diagnosis of cervical cancer using SVM classifier.” (2016).

9. D. Kashyap et al., "Cervical cancer detection and classification using Independent Level sets and multi SVMs," 2016 39th International Conference on Telecommunications and Signal Processing (TSP), Vienna, 2016, pp. 523-528.

10. Soumya MK, Sneha K, Arunvinodh C. Cervical Cancer Detection and Classification using Texture Analysis. Biomedical and Pharmacology Journal. 2016 Jun; 9(2):663– 71.

11. Mustafa, N., N. A. Mat Isa, M. Y. Mashor, and N. H. Othman. “New Features of Cervical Cells for Cervical Cancer Diagnostic System Using Neural Network.” IJSSST, 2008; 9(2).

12. “Image Texture Feature Extraction using GLCM Approach” by P. Mohanaiah*, P. Sathyanarayana, L. GuruKumar Professor, Dept. of E.C.E, N.B.K.R.IST, Vidyanagar, Nellore, India Professor, Dept. of E.C.E, S.V University Tirupati, India Asst.Professor, Dept. of E.C.E, N.B.K.R.IST.

13. S. Kaaviya, V. Saranyadevi and M. Nirmala, "PAP smear image analysis for cervical cancer detection," 2015 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, 2015, pp. 1-4.

14. K.Pradeep Chandran, Smt.U.V.Ratna Kumari “Improving Cervical Cancer Classification On MR images Using Texture Analysis And Probabilistic Neural Network” IJSETR, Volume 4, Issue 9, September 2015.

15. Priyank K Malli, Dr. Suvarna Nandyal, Machine Learning Technique for Detection of Cervical Cacner using k-NN and Artificial Neural Network. IJETTCS Volume 6,Issue 4,July-August 2017

16. Chankong T., Theera-Umpon N., Auephanwiriyakul S. (2009) Cervical Cell Classification using Fourier Transform. In: Lim C.T., Goh J.C.H. (eds) 13th International Conference on Biomedical Engineering. IFMBE Proceedings, vol 23. Springer, Berlin, Heidelberg

17. S. Garg, S. Urooj and R. Vijay, "Detection of cervical cancer by using thresholding & watershed segmentation," *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2015, pp. 555-559.

18. Sajeena T A and Jereesh A S, "Automated cervical cancer detection through RGVF segmentation and SVM classification," *2015 International Conference on Computing and Network Communications (CoCoNet)*, Trivandrum, 2015, pp. 663-669.

19. Haralick, Robert & Shanmugam, K & Dinstein, Itshak. (1973). Haralick RM, Shanmuga K, Dinstein ITextural features for image classification. IEEE Trans Syst Man Cybern 3: 610-621. Systems, Man and Cybernetics, IEEE Transactions on. SMC3. 610 - 621. 10.1109/TSMC.1973.4309314.

20. Lange, Holger. (2005). Automatic glare removal in reflectance imagery of the uterine cervix. Progress in Biomedical Optics and Imaging - Proceedings of SPIE. 5747. 10.1117/12.596012.

14. GLOSSARY

Contrast
Contrast measures the quantity of local changes in an image. It reflects the sensitivity of the textures in relation to changes in the intensity. It returns the measure of intensity contrast between a pixel and its neighborhood. Contrast is 0 for a constant image. It is the amount of local variation present in an image. If the amount of local variation is large, the contrast feature also has consistently higher values comparatively. If the gray scale difference occurs continually, the texture becomes coarse and the contrast becomes large. The texture becomes acute if the contrast has a smallvalue.

Correlation
This feature measures how correlated a pixel is to its neighborhood. It is the measure of gray tone linear dependencies in the image. Feature values range from -1 to 1, these extremes indicating perfect negative and positive correlation respectively. and are the means and and are the standard deviations of) and), respectively. If the image has horizontal textures the correlation in the direction of 0° degree is often larger than those in other directions.

Homogeneity
Homogeneity measures the similarity of pixels. A diagonal gray level co-occurrence matrix gives homogeneity of 1. It becomes large if local textures only have minimal changes.

Energy
Energy also means uniformity, or angular second moment (ASM). The more homogeneous the image is, the larger the value. When energy equals to 1, the image is believed to be a constant image.

Entropy
Entropy is a measure of randomness of intensity image.

Angular Second Moment Feature (ASM)
Angular second moment feature is a measure of the uniformity of local gray scale distribution. If P (i,j,d,theta) is centralized near the main diagonal area the local gray scale distribution becomes uniform. It is a measure of homogeneity of the image. In a homogeneous image very few dominant gray tone transitions will be present. The matrix for the image will
have fewer entries of large magnitude.