Machine learning Technique for detection of Cervical Cancer using k-NN and Artificial Neural Network

Priyanka K Malli, Dr. Suvarna Nandyal

1Department of Computer Science & engineering, PDA Engineering college, Kalaburagi, Karnataka, India
2HOD Department of Computer Science & engineering, PDA Engineering college, Kalaburagi, Karnataka, India

Abstract
Cervical cancer along with micro classification are one of the 2 major forms of cancer being observed amongst women across the globe. A cervical cancer results in dead nucleus or change in the morphology of the cells in the cervix. Such cells may have multiple nucleuses, faulty cytoplasm, lack of cytoplasm, dissolved lack of cytoplasm and so on. Detection of cervical cancer in a microscopic smear Test (fluid taken from the cervix) smear is analyzed to microscopic is extremely challenging because such cells does not offer significant color or texture variations from the normal cells. Therefore high level Digital Image Processing technique are required identify abnormalities in human cell related cancer detection system. Therefore an automated, comprehensive machine learning technique has been proposed in this work. The proposed technique gives that color and shape features of nucleus and cytoplasm of the cervix cell. The nucleus and the cytoplasm are separated from the cell using the advanced fuzzy based technique. KNN and Neural network are trained with the shape features and color features of the segmented units of the cell and then an unknown cervix cell samples are classified by this technique. The classification result have shown an accuracy of 88.04% for KNN and 54% for ANN. The proposed system work can be further enhanced by taking other classifiers.

Keywords: Pap smear Images, Feature Extraction, KNN, ANN

1. INTRODUCTION
Cervical cancer is the second most common type of cancer that affects women, ranked after the breast cancer. The acceptability of increased cervical cancer risk and cause of cancer death projected. Cervical screening program has minimize the rate of death in developed countries. Cervical cancer is the one of the deadliest disease, it can be cured if detected in early stage.

Pap Smear Test this was introduced as a screening test for cervical cancer in the year of 1943 by Dr. George Papanicolaou name of the scientist who proposed a method where a spoon like wooden instrument is used to take out a smear to use we need a microscope. Our technique first segments each independent cells in the microscopic image. One or more cells may cancerous cells or non of the cells may be cancerous or all the cells present in the microscopic image may be cancerous. Therefore regression based technique can’t separate the cancers cells from the normal cells therefore our technique should effectively separate every cell present in the image and then extract the features. And the morphology descriptor for each of the cell presents that information to a machine learning system which farther classifies the normalcy or the presence of cancer from each of the cells if one or more cells present in the cervical smear sample has cancer then we say the presence of the cervical cancer.

Fig 1a: Normal cervix cell Fig 1b. Abnormal cervix cell.

2.RELATED WORK
The work related to problem are outlined as follows
[1] Authors have made a survey and analysis on the different types of detection of cervical cancer which is useful for the classification of normal and abnormal cervical cells [2] automated diagnostic system using image processing techniques have been proposed which proves to be an aid to radiologist and play a key role in early detection of cancer using MRI of cervical cancer [3] The authors proposed the computer assisted pap smear Analyzer for cervical cancer screening using quantitative microscopy for detection of normal and abnormal cancer cells [4] In this the cytology images are taken to segment the nucleus and cytoplasm for detection of cervical cancer [5] also two-level cascade classification is used to classify the images captured from thin liquid based cytology slides. [6] And author shown that comparison of classification techniques for pap smear Diagnosis, The classification for pap smear diagnosis aims at classifying the pap smear cells whether it is affected or not.[7] The author proposed semi-automatic segmentation and classification of pap smear cells. The proposed system provides a feasible and effective tool in evaluating cytological specimens. [8] Diagnosis method and multi-thresholding SFS SFFS is used for detecting cervical cancer [9] fully automated method is used for the author

Volume 6, Issue 4, July – August 2017 Page 145

3. MATERIALS AND METHODS

The proposed system is works based on MATLAB 2016 .And it deals with implantation of an automated system which helps us to detect the cervical cancer at the cell level using Digital Image Processing technique

![Figure 2: Block diagram of the Proposed system for detecting cervical cancer](image)

Initially the process includes color conversion for RGB to Gray scale . Enhancement and scaling, segmentation, detecting the cell boundary and detecting the overlapping cytoplasm and nuclei, feature extraction and classification of cells based on the features. Most discriminating set of features were selected and used as input to an artificial neural network (ANN) and KNN. The shape features such as roundness, area, width, height ,texture features and color features are extracted. Both features are trained by ANN and KNN classifier.

2.1. Segmentation

The segmentation is based on the Fuzzy c mean clustering technique.

Fuzzy c- mean clustering

This is one of the data clustering technique which allows one piece of data to belong to two or more clusters. There are two clusters nucleus and cytoplasm .Segmentation is the fundamental step of image analysis it separates the objects form the background to obtain the region of interest so as to get the features . Many types of methods are there among those we are using thresholding method in this work .

![Fig.3.a: Normal cell](image) ![fig.3.b: Segmented cell](image)

2.2. Feature Extraction

Feature extraction is transforming the input data into the set of features . In this Work there are many features are taken in to consideration for detecting the cancerous cell from the sample pap smear image.

**Nucleus and Cytoplasm Morphology**

In this we are taking color features and shape features of the nucleus and cytoplasm. The features of nucleus and cytoplasm such as area, centroid ,size of the nucleus ,shape of the nucleus ,nucleus cell distribution are extracted using region properties. The area of the nucleus is the number of pixels inside the nucleus . In the training phase the area of the nucleus of the normal cells and abnormal cells are calculated . The small areas remaining in the picture which might be looking like a nucleus are removed first . Negative of the segmented image is more suitable as it aids in removing unwanted clusters using morphological erosion faster than the true image . The resultant image has nucleus region along with small suspicious objects with can be removed by performing erosion operation. The eroded image is subtracted from the negative image to get counter of the segmented objects as shown in the figure. After erosion the resultant image consists of nucleus region along with other regions. It is required to select only the nucleus region by some criteria. From this it is observed that the nucleus is bigger region than the other regions. Hence the next step is the labeling the objects followed by selecting the bigger cluster of pixels among the various clusters. After extracting features from the sample image, we apply thresholding for the extracted features. The threshold is calculated using these following color features and shape features

Nucleus area and Cytoplasm area : Calculated by counting the corresponding pixels of the segmented picture. A pixels area is

\[(0.201\mu m)^2\]

N/C ratio : Tells how small the nucleus area is compared to the area of the cytoplasm. It is given by:

\[N/C = \frac{Nucl_{area}}{Nucl_{area} + cyto_{area}}\]
Figure 4: A binary picture of a cell with background(white), cytoplasm(dark blue) and nucleus(light blue). Longest diameter line(L) and shortest diameter lines(s1 & s2) is shown.

**Nucleus and Cytoplasm brightness:** Nucleus and Cytoplasm brightness is calculated as the average perceived brightness, that is a function of the colors wavelength. In our case it is calculated as:

\[ Y = 0.299 \cdot \text{Red}_\mu + 0.587 \cdot \text{Green}_\mu + 0.114 \cdot \text{Blue}_\mu \]

*intensity for each of the colors. They are weighted by the perceived brightness of the human eye.*

**Nucleus and Cytoplasm longest diameter:** This is the shortest diameter a circle can have, when surrounding the whole object. It is measured as the biggest distance between to pixels on the objects border, and forms a line L as shown in figure 4 for the cytoplasm. The names is Nlong and Clong for nucleus and cytoplasm, respectively.

**Nucleus and Cytoplasm shortest diameter:** This is the biggest diameter a circle can have, when the circle is totally encircled of the object. This distance is approximated by the sum of the two lines s1 & s2 shown in figure 4, for the cytoplasm. They are perpendicular to the line L, and is the longest line to each side sitting inside the object. The names of the features is Nshort and Cshort for nucleus and cytoplasm, respectively.

**Nucleus and Cytoplasm elongation:** The elongation is calculated as the ratio between the shortest diameter and the longest diameter of the object.

\[ N_{\text{elong}} = \frac{N_{\text{short}}}{N_{\text{long}}} \quad C_{\text{elong}} = \frac{C_{\text{short}}}{C_{\text{long}}} \]

**Nucleus and Cytoplasm roundness:** The roundness is calculated as the ratio between the actual area and the area bound by the circle given by the longest diameter of the object. They are here named Nroundness and Croundness respectively:

\[ N_{\text{circle}} = \frac{\pi}{4} \cdot N_{\text{long}}^2 \Rightarrow N_{\text{roundness}} = \frac{N_{\text{area}}}{N_{\text{circle}}} \]

\[ C_{\text{circle}} = \frac{\pi}{4} \cdot C_{\text{long}}^2 \Rightarrow C_{\text{roundness}} = \frac{C_{\text{area}}}{C_{\text{circle}}} \]

**Nucleus position:** This is a measure of how well the nucleus is centered in the cytoplasm. It is calculated by finding the distance between the nucleus center and the center of the cytoplasm:

\[ N_{P_{\text{OS}}} = \frac{2a \cdot \sqrt{(x_{\text{N}} - x_{\text{C}})^2 + (y_{\text{N}} - y_{\text{C}})^2}}{C_{\text{long}}} \]

*Here the position is given by the centrum (Xs,Ys) and (Xc,Yc) for the nucleus and cytoplasm respectively.*

**Nucleus and Cytoplasm Maxima/Minima:** This is a count of how many pixels is a maximum/minimum value inside of a 3 pixel radius.

**3. CLASSIFICATION**

Classification is divided into two steps training and testing. Training is the one where we tell the classifier what are the steps to classify images. In testing we give an input and test the class. Here whatever the features are extracting it consist of a trained matrix. After training feature library is created. The trained image is classified using k-NN and ANN classifier. Based on feature extraction matches the trained image is classified for cancer types. Trained matrix consist of mainly 3 parameters F, C, INFO. Here we are taking the size of F (it consist of all 20 features which we previously trained) ‘C’ contains all the class names and
INFO all the file names. After taking all these parameters we are use the classifiers to classify. In this work the classification is made using two classifiers K-nearest neighbor and Neural network methods.

### 3.1 K-nearest neighbor classification

It is the pattern recognition technique used for classification. Here we load the train matrix. Here we have to classify the result of feature of the current image. Against the F and C. It will be going to give us detected class of C. If the loaded image shape features and color features are matches with the trained matrix then it shows the type of Cancer.

### 3.2 Artificial Neural Network

The neural network is simply trained by loading the train matrix and then it expects the classes based on vectorized form and based on feature vectors. Once we train the neural network, here we are classifying the features which are of 20 types. We take the features as an input layer and the number of classes is taken as the output layers. And the hidden layer consists of number of images trained and it will be mapped with number of layers. As many as we train the images the hidden layers also increases.

![Neural Network Training](image7.png)

**Fig 7:** Neural Network Training

In order to test the neural network, we have to load our pre-stored network and we are going to calculate the result by simulating current set of networks. Likewise we are performing the neural network operations.

### 4. Performance measures

**Working of segmentation and feature extraction**

<table>
<thead>
<tr>
<th>Time</th>
<th>Image</th>
<th>Image shows</th>
<th>Image shows</th>
<th>Image shows</th>
<th>Image shows</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td>2</td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td>3</td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
</tr>
<tr>
<td>4</td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
</tr>
<tr>
<td>5</td>
<td><img src="image21.png" alt="Image" /></td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
<td><img src="image25.png" alt="Image" /></td>
</tr>
</tbody>
</table>

**Fig 8:** Feature extraction process

### 4.1 Performance comparison of sample images using k-NN and ANN

**Fig 9:** performance comparison KNN and ANN

<table>
<thead>
<tr>
<th>Cancer Types</th>
<th>k-NN</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carcinoma_in situ</td>
<td>100%</td>
<td>95%</td>
</tr>
<tr>
<td>Moderate Dysplastic</td>
<td>72.73%</td>
<td>62%</td>
</tr>
<tr>
<td>Normal Superficial</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>

**Fig 10:** Classification accuracy of Cancer types using k-NN classifier

From above graph it is absorbed that based on analysis of shape and color features of images the maximum accuracy of 100% is obtained for carcinoma_in suit cancer types and minimum accuracy of 72.73% is obtained for moderate_dysplastic cancer types.

**Fig 11:** Classification accuracy of Cancer types using NeuralNetwork classifier

From above graph it is absorbed that based on analysis of shape and color features of images the maximum accuracy of 95% is obtained for carcinoma_in suit cancer types and minimum accuracy of 25% is obtained for normal_superficial cancer types.

### Total Accuracy of the work

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-NN</td>
<td>88.04%</td>
</tr>
<tr>
<td>ANN</td>
<td>54%</td>
</tr>
</tbody>
</table>
5. Conclusion and future works
In order to analysis whether pap smear test or cytoplasmic test the results analysis is a cells .which is observed in microscope .our technique first segments each independent cells components such as nucleus and cytoplasm and then detect whether one or more cells are cancerous or not through machine learning based technique. There are several technique which are been proposed in the past towards this direction there accuracy has not been found to be significantly accessible .In this work we have developed a machine learning based technique coupled with efficient Fuzzy based segmentation to accurately segment and classify the cell in to normal or cancerous cell and its types . In our work the Average classification result have shown an accuracy of 88.04% for k-NN and 54% for ANN.

The proposed system work can be further enhanced by taking other classifiers. And also it can be further improved by not only classifying the different types of cells but also at the same time that classify the stage of the cancer such as beginning or malignant . Any detection of the cancer at the beginning stage can help significantly to reduce the risk of life by the patient

REFERENCES
Three Algorithms for the segmentation of overlapping Cervical cancer cells ”Journal of Biomedical and Health -11
[12]Erik Martin “Classification of Pap smear images” Technical University of Denmark

AUTHORS
Priyanka K Malli   Recieved B.E degree in computer science and engineering from Visveswaraya Technological University Belgaum Karnataka India in the year of 2014. And at present studying M.Tech in computer science and engineering from PDA Engineering college kalaburagi,karnataka,India.

Dr.Suvarna Nandyal  Recieved B.E and M.Tech degree In Computer science and Engineering from visvesvaraya Technological University Belgaum,Karnataka India and Recieved Ph.D from Jawaharlal Nehru Technological University Hyderabad. At present working as Head of the Department at PDA Engineering college Kalaburagi Karnataka,India.