Spring 2017

Nuclei segmentation using level set method and data fusion for the CIN classification

Ravali Edulapuram

Department:

NUCLEI SEGMENTATION USING LEVEL SET METHOD AND DATA FUSION FOR THE CIN CLASSIFICATION

by

RAVALI EDULAPURAM

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

2017

Approved by

R. Joe Stanley, Advisor
Randy H. Moss
William Stoecker

PUBLICATION THESIS OPTION

This thesis consists of the following article which has been submitted for publication as follows:

Paper I: Pages 4-22 have been submitted to the VISAPP conference.

ABSTRACT

This paper deals with the automation of the detection of the cervical cancer through histology images. This process is divided into two parts, corresponding to segmentation and data fusion. The segmentation and classification of the cervical epithelium images is done using hybrid image processing techniques. The digitized histology images provided have a pre-cervical cancer condition called cervical intraepithelial neoplasia (CIN) by expert pathologists. Previously, image analysis studies focused on nuclei-level features to classify the epithelium into the CIN grades. The current study focuses on nuclei segmentation based on the level set segmentation and fuzzy c-means clustering methods. Morphological post-processing operations are used to smooth the image and to remove non-nuclei objects. This algorithm is evaluated on a 71-image dataset of digitized histology images for nuclei segmentation. Experimental results showed a nuclei detection accuracy of 99.53 percent. The second section of this thesis deals with the fusion of the 117 CIN features obtained after processing the input cervical images. Various data fusion techniques are tested using machine learning tools. For further research, the best algorithm from Weka is chosen.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. R. Joe Stanley, for his continuous support and guidance throughtout this graduate program. I thank him for introducing me to the research topic, advising me and teaching me how to analyze a problem. I also thank my committee members, Dr. Randy H. Moss and Dr. William Stoecker, for analyzing my research work, encouragement and support all through the process.

This research was supported [in part] by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC). In addition we gratefully acknowledge the medical expertise and collaboration of Dr. Mark Schiffman and Dr. Nicolas Wentzensen, both of the National Cancer Institute's Division of Cancer Epidemiology and Genetics (DCEG).

I would also like to thank Haidar Almubarak and Peng Guo for their immmense knowledge and great technical suppport.

I thank my friends for their enthusiasm and support all through the process. I want to express my sincere gratitude to my parents and my sister, Chandana, for their love and support.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

NOMENCLATURE

Symbol	Description
U_k	Membership function of the k^{th} pixel
Y_i	Observed intensities of the pixel
X_i	True intensities of the pixel
G_i	Gain field for the i^{th} pixel of the image
Q	General Cost function
Q'	Updated Cost function

1. INTRODUCTION

The second most common cancer found in women is cervical cancer. Cervical Intraepithelial Neoplasia (CIN) is the abnormal growth of squamous cells on the surface of the cervix. A small percentage of the abnormal growth cases lead to cervical cancer. Early detection of the disease might cure the cancer completely. Pathologists diagnose by examining the cell and tissues under the microscope. The level of the disease can be obtained by the thickness of the squamous epithelium on the surface of the cervix tissue. CIN has been categorized into four grades Normal, CIN-1, CIN-2 and CIN-3, where Normal indicates no sign of cancer and CIN3 indicates high possibility of pre-cancerous condition. The classification of the CIN grades is shown in the Figure 1.1.

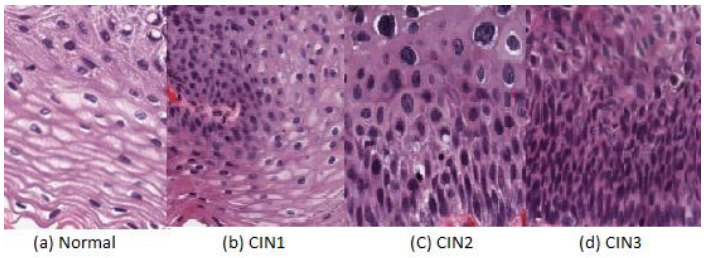

Figure 1.1: Classification of CIN grades

From Figure 1.1 it can be seen that the immaturity of the cells increases from bottom to the top as the severity of the disease increases. CIN1 is the low possibility of pre cancer and as the grade of CIN increases (2 and 3), images show a substantial increment in the number of immature cells which is composed of undifferentiated cells.

The results obtained from the pathologists are extremely accurate and provide a gold standard, but the time is required for each diagnosis and the degree of expertise to make the diagnosis are both scarce resourses. Thus automation of the process is desirable. To achieve automation, many researchers have combined and implemented various machine learning algorithms to automate the process of cancer detection and classification. Detection of the cancer and the level of the disease can be obtained by using image processing techniques. Many machine learning algorithms have been implemented for the estimation of CIN grades based on input images obtained from the

pathologists. The computation technique for this process includes preprocessing, segmentation, feature extraction and then classification using data fusion techniques. Algorithms involved in feature extraction from the processed images and then one of the data fusion methods are implemented to obtain the final classification label. Selection of good features contributing to the final image label plays an important role in obtaining the best classification results. Various features include visual texture, color, Gabor and nuclei features which all contribute to the classification process. Of all features, nuclei features play a crucial role in deciding the final image label, providing important data for the final classification. Therefore nuclei segmentation plays an important role in cervical epithelium grading.

Various segmentation methods have been introduced including convolutional neural networks, fuzzy clustering, support vector machines (SVM), watershed segmentation, level sets, probabilistic models, graph cuts etc. to obtain accurate segmentation results. (Malpica et al. 1997) has implemented a watershed algorithm to detect clustered nuclei using brightness and morphology features. This is limited to fluorescent-stained nuclei clusters. (Ghosh et al. 2010) have used leukocyte images instead of fluorescent-stained images and have implemented the fuzzy divergence algorithm for estimation of the threshold. (Sinha and Ramakrishnan 2003) have used a hybrid method for segmentation of nuclei and the cytoplasm from the image background. The method implemented was the two-step color image segmentation combined with k-means clustering. (Sarrafzadeh and Dehnavi 2015) have implemented a generic segmentation process for the pathology images for the nuclei and cytoplasm segmentation using K-means clustering and region growing techniques. (Yang, Meer, and Foran 2005) have implemented a gradient vector color flow active contour method for the segmentation of the images in the luv color space. This method used an unsupervised segmentation approach. As it can be observed that there are various approaches for the segmentation based on the input images. In this paper a hybrid algorithm is introduced for the nuclei segmentation from the background of the cervical images. A combination of active contour level set segmentation with fuzzy c-means clustering has been implemented for the segmentation of the nuclei. This approach is based on initialization using fuzzy clustering; the algorithm stops after all the edges of the nuclei are discovered.

Similarly once the segmentation is done, features are extracted using various techniques such as SVM, LDA, logistic regression and so on. The next problem is to combine the data in such a way that the final output of the algorithm corresponds to the final image class label. Features can be combined using various data fusion techniques. A review (Castanedo 2013) describes three categories for the data fusion process including data association, state estimation and decision fusion. Various frameworks for data fusion such as Whyte's classification, Dasarathy's classification and JDL data fusion framework are presesnted. (Crowley and Demazeau 1993) discuss a specific data fusion problem which helps in perceptual fusion of the sensor data where the procedure for the data fusion is derived from the construction of the sequence of systems. In this paper, data fusion of the features obtained from nuclei segmentation is obtained by implementing various techniques and finally a best technique is chosen and implemented which will be discussed in the later sections.

This thesis deals with segmentation of the nuclei from the background and fusion of the cervical image features. The second section presents nuclei segmentation based on fuzzy clustering and the active contour method. The third section explains various data fusion techniques and their behaviour with respect to the features.

PAPER

I. NUCLEI SEGMENTATION USING LEVEL SET METHOD AND FUZZY C-MEANS CLUSTERING

ABSTRACT

Digitized histology images are analyzed by expert pathologists in one of several approaches to assess pre-cervical cancer conditions such as cervical intraepithelial neoplasia (CIN). Many image analysis studies focus on detection of nuclei features to classify the epithelium into the CIN grades. The current study focuses on nuclei segmentation based on level set active contour segmentation and fuzzy c-means clustering methods. Logical operations applied to morphological post-processing operations are used to smooth the image and to remove non-nuclei objects. On a 71-image dataset of digitized histology images, the algorithm achieved an overall nuclei segmentation accuracy of 96.47%. A simplified fuzzy spatial cost function has been proposed that may be generally applicable for any n-class clustering problem of spatially distributed objects.

1. INTRODUCTION

The abnormal growth of squamous cells on the surface of the cervix leads to cervical cancer. Quantitative study of the cervix tissue helps in the early detection of cancer. The thickness of the squamous epithelium on the surface of the cervix and the various nuclei features have been examined in previous studies including (Krishnan et al. 2012) to determine the grades of cervical intraepithelial neoplasia (CIN), a cervical cancer precondition. CIN grades include normal, CIN-1, CIN-2 and CIN-3. CIN-1 grade showcases the initial stage of the pre-cancerous condition and CIN-2 and CIN-3 reveal a greater density of nuclei. Examples of the CIN grades are demonstrated in Figure 1.

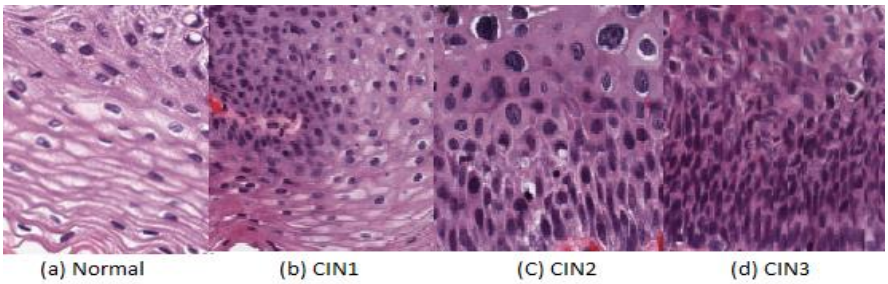

Figure 1: Examples of CIN grades indicating increase in the number of immature cells

Many algorithms are implemented for detection of the nuclei features over the cervical tissue. (Song et al. 2015) has used convolutional nets and graph partitioning for segmentation of the nuclei and the cytoplasm. (Lu, Carneiro, and Bradley 2013) implemented K-means for the initial segmentation and superpixels for the segmentation of cytoplasm and nucleus. (Walker et al. 1994) used fast morphological gray-scale transforms for nuclei segmentation. There have been several studies investigating approaches to segment nuclei and nuclei related features. Accurate nuclei features are achieved from the ideal nuclei segmentation algorithms. Most of the previous algorithms are related to deep learning techniques. Therefore this paper introduces a level set algorithm in combination with the fuzzy clustering algorithm for the segmentation of the nuclei.

1.1. FUZZY CLUSTERING AND LEVEL SET ALGORITHM

The level set Algorithm is often combined with an additional algorithm for accurate results.(Wang and Pan 2014) has used the local correntropy-based K-means along with the level set algorithm. This algorithm helps in eliminating the complex noise present. Similarly this paper uses spatial fuzzy c-means clustering for the initialization of the level set parameters. Those parameters change with respect to the type of input image.

Initially the level set method uses the level set function which evolves from the zero level set function and ideally stops at the boundaries of the object. This function is restricted by the driving force. This force can be either the inverse of the gradient of the image or a Gaussian function which can be a constant positive or the negative force based on the input image. The driving force function is generally the gradient function because

it detects the sharp intensity changes in an image. Therefore the value of the gradient is high at the edges which indicate the sharp intensity change. The contour is obtained based on the driving force. High driving force inside the object allows the contour to expand and low force at the edges allows the evolution to stop at the edges of the object (Phillips 1999). In this paper the driving force is controlled by the membership function. This helps in controlling the evolution of the level set. These are the main steps of the algorithm which are explained below. The sequential steps for the proposed algorithm are shown in Figure 2.

Project Specifications: The algorithm to segment the nuclei as proposed in this paper includes level set segmentation which uses fuzzy c-means clustering for the initialization of parameters. A brief description of the algorithm is given below.

- Level set segmentation is used since the initial contour (zero level set) function starts evolving and stops at the edges of the nuclei. The starting zero level set function is initialized with respect to the fuzzy c-means output. This algorithm is divided into the following steps.
- Images from the 71-image dataset are given as input; input images are modeled so as to include the spatial information into each pixel. This modelling is done with respect to the gain field function.
- This modelled input is used to calculate the new cost function for the algorithm. Minimization of updated cost function gives the new membership function, centroids and the bias field. These new membership functions can be used as the update laws for the fuzzy c-means clustering.
- The updated membership function is used to obtain the energy function. This energy function helps in deriving the driving force which controls the motion of the level set function.
- Minimizing the energy function gives the driving force and this driving force helps in evolution and evolution termination. This method divides the whole image into two regions, one for nuclei and one for non-nuclei regions. This helps in obtaining the contour which surrounds all the nuclei by separating the image using the values of the data points.
- Contour obtained from the previous step is used to obtain the mask of the nuclei.

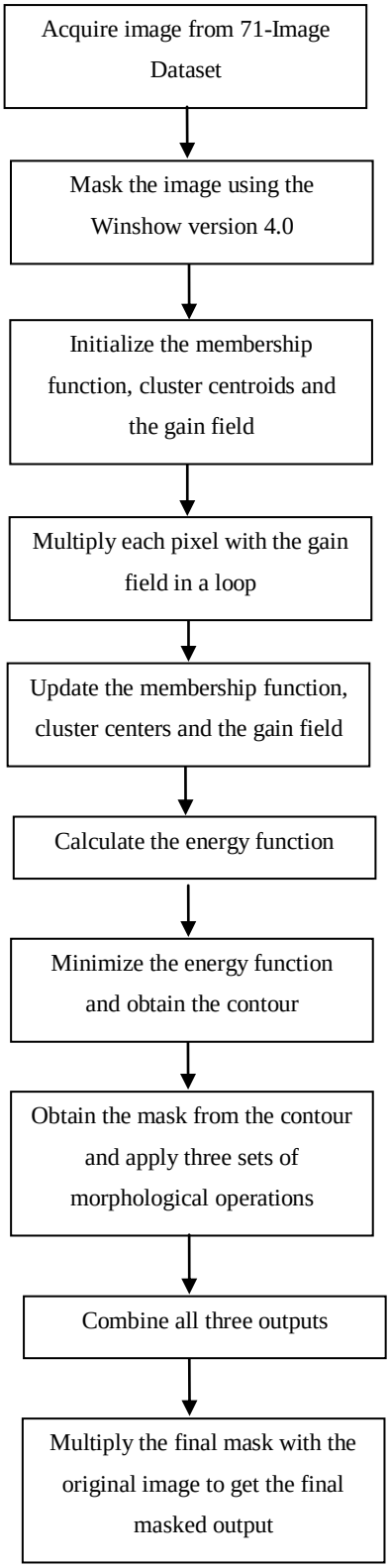

Figure 2: Sequential steps of the proposed algorithm

- Morphological operations are applied on the previously obtained mask to clean the image by removing unwanted masked objects. Three sets of morphological operations are applied so as to remove the unwanted objects while retaining the desired objects.
- The three outputs obtained from the previous operations are combined to give the final mask. Three sets of combination of morphological operations are applied since the data included in one output might be present in another output. So as to avoid the data losses, all three outputs are combined.
- This final masked output is evaluated based on the small algorithm which calculates the number of objects detected in the mask and by visually calculating the number of false positive and true negative objects in the mask. The percentage of algebraic sum with respect to the number of nuclei detected gives the accuracy of segmentation of nuclei.

2. PROPOSED ALGORITHM

The images used for the segmentation of the nuclei are from a 71-image dataset obtained from the database of the National Library of Medicine, NLM. These images are the digitized histology images of hematoxylin and eosinophil (H&E) preparations of tissue sections of regular cervical tissue (Guo et al. 2015). These images are initially masked to eliminate the non-epithelium regions. This masking is done manually using the application Winshow software version 4.0 (Erkol et al., 2005; Kasmi et al., 2005). The input and the output images are shown in the Figure 3.

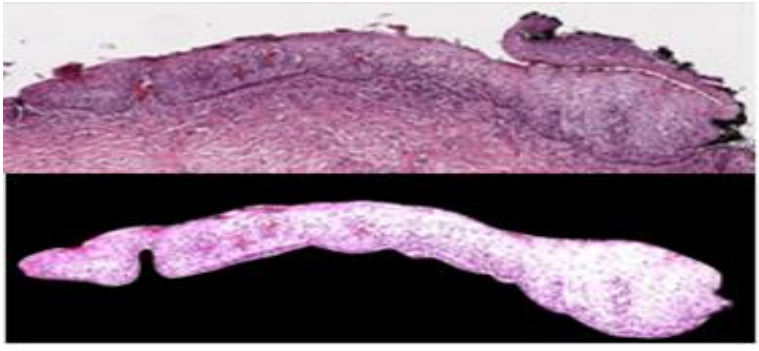

Figure 3: The unmasked and the masked image of the epithelium

2.1. SPATIAL FUZZY C-MEANS CLUSTERING

As discussed above spatial information is included in the membership function. The input image obtained after masking is modeled to include the spatial information. A gain field is introduced and is multiplied with each pixel of the input image. This helps in including the spatial information in each pixel rather than using a confined window. The equation for the modeling of the input image is shown below (Balla-Arab, Gao, and Wang 2013).

$$Y_i = X_i G_i, \forall i \in \{1,2,3, \dots, N\} \tag{1}$$

where Y_i and X_i are the observed and the true intensities of the pixel and G_i is the gain field for the i^{th} pixel of the image. N is the total number of pixels present in that image. This modeled image is then used for further analysis in place of the input image. The cost function of this algorithm is modified by introducing the modeled output in place of the general input. The general cost function Q and the updated cost function Q' are shown below.

$$Q = \sum_{j=1}^{k} \sum_{i=1}^{N} u_{ji}^{f} \left\| x_i - v_j \right\|^2 \tag{2}$$

$$Q' = \sum_{j=1}^{k} \sum_{i=1}^{N} u_{ji}^{f} \left\| y_i / g_i - v_j \right\|^2 \tag{3}$$

As it is observed that the parameter ‘f’ indicates the amount of fuzziness to be included for each cluster in the cost functions. This value is obtained by applying the algorithm on various inputs with various fuzziness values. The final fuzzy value which is used in this algorithm is 2. This helps in introducing the spatial information into the membership function. In general, the minimization of the cost function gives the final clusters and their centers which are the values obtained after the convergence. Here the updated cost function is minimized to get the converged cluster centers. Here the minimization is done based on the gradient descent method. While minimizing, the first derivatives of the cost function are calculated with respect to the membership function, cluster centers and the gain field. The first derivatives obtained are then equated to zero

which gives the points where the cost function is minimized. These derivatives when equated to zero give the final update laws for the membership function, gain field and the cluster centers. The equations for the updated membership function, cluster centers and the gain fields are given below (Balla-Arab, Gao, and Wang 2013).

$$U_k(x,y) = \frac{1}{\sum_{l=1}^{a}\left(\frac{\|Y(x,y) - B(x,y) - v_k\|}{\|Y(x,y) - B(x,y) - v_l\|}\right)^{\frac{2}{(f-1)}}} \tag{4}$$

$$v_k(x,y) = \frac{\int_{\omega}^{\emptyset} U_k^f(x,y)\left(Y(x,y) - B(x,y)\right)dx\,dy}{\int_{\omega}^{\emptyset} U_k^f(x,y)\,dx\,dy} \tag{5}$$

$$B(x,y) = Y(x,y) - \frac{\sum_{j=1}^{k} U_j^f(x,y)v_k}{\sum_{j=1}^{k} U_j^f(x,y)} \tag{6}$$

where U_k indicates the membership function of the k^{th} pixel and $Y(x,y)$ is the modeled input image at that particular location. $B(x,y)$ is the bias function which is obtained using gain field of the modeled image at that location. f indicates the amount of fuzziness to be included in each cluster. $\emptyset$ indicates the whole image whereas ω indicates the part of the image. v_k indicates the centroid of the k^{th} pixel. a indicates the number of clusters.

2.2. SIMPLIFIED SPATIAL COST MEMBERSHIP FUNCTION FOR OPTIMAL CLUSTERING

As it can be seen, the exponent in the membership function is $\frac{2}{(f-1)}$. Here if the value of f is greater than 2, the membership function increases gradually which might lead to over clustering and if the value of f is less than 2 and greater than 1, the membership function decreases and the pixels which are supposed to have high membership function will have low membership function which might lead to under-segmentation. So as to balance the segmentation, the fuzziness parameter is taken as 2,

optimized by running the algorithm on various input images. So when f is taken as 2, the membership function reduces to the equation given below.

$$U_k(x,y) = \frac{1}{\sum_{l=1}^{a} \left(\frac{\|Y(x,y) - B(x,y) - v_k\|}{\|Y(x,y) - B(x,y) - v_l\|} \right)^2} \tag{7}$$

This is the final membership function which is used to derive the energy function of the image and also the driving force which controls the motion of the level set function. Since just need two clusters are needed, one for nuclei and one for non nuclei, the number of clusters is assigned as two. This assumption will give rise to the following equation.

$$U_1(x,y) + U_2(x,y) = 1 \tag{8}$$

2.3. LEVEL SET ACTIVE CONTOUR METHOD

The membership function obtained from the above equation includes the spatial information. This membership function is used in the energy function to obtain the driving force. This driving force generally is taken as the inverse of the gradient of the image or the Gaussian function of the image. This can be either positive or negative based on the initial contour position with respect to the required object, because the gradient of the image gives the information about the edges present in the image which can be utilized to stop the contour evolution exactly at the edges. The value of the gradient of the image is high at the edges. The level set function starts at the zero level set function and evolves until the edges of the object are found. The zero level set function in general is obtained from the level set function intersected with a constant plane. This intersection gives a contour in two dimensional space. This is demonstrated in Figure 4. $\emptyset(x,y,t)$ is the level set function and $\emptyset = 0$ is the equation for the zero level set function. The red contour obtained is the intersection of the level set function and the plane which is the zero level set function. The evolution of the level set function starts from the zero level set function and the evolution is ideally stopped at the edges of the required nuclei.

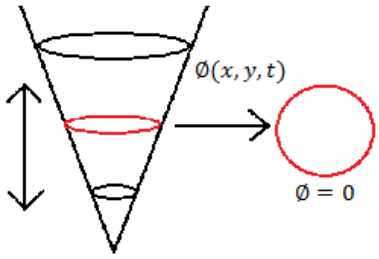

Figure 4: Represents zero level set function of the level set method

In this paper the driving force is obtained from fuzzy c-means clustering by using the membership function. The parameters of the level set algorithm which control the evolution are shown in Table 1.

Table 1: Parameters of level set

Parameters	Description
τ	Time step of evolution
f	Fuzziness parameter
V1	Cluster centre
V2	Cluster centre
λ	Multiplicative factor

'τ' represents the time step. A larger time step may reduce the time of evolution but may reduce accuracy. This value is optimized after the algorithm is applied on various input mages. This value varies with the type of input images and may remain consistent among the same type of images. 'f' represents the amount of fuzziness induced in the membership function. This value is taken as 2 as discussed above. 'V1' and 'V2' are the cluster centres, one for the nuclei and one for the non-nuclei regions and is initialized accordingly. 'λ' is a constant which is multiplied to the force function. This value is taken as 2 for these images. This value is obtained after the algorithm is applied on several input images. The membership function is introduced into the driving force; the equation for the driving force is given below (Balla-Arab, Gao, and Wang 2013).

$$F = \lambda(U_1^f(x,y)\|Y(x,y) - B(x,y) - v_1\|^2 - U_2^f(x,y)\|Y(x,y) - B(x,y) - v_2\|^2) \quad (9)$$

where λ is a parameter which enhances or reduces the controllability of the driving force. In this paper the value of λ is considered as 1. If the segmentation is not done for λ =1, then the λ value can be increased for better segmentation. This driving force contains the modeled input image, gain field information, membership function and the cluster centers. These values are obtained from the previously derived equations and are substituted in the equation of the driving force. Minimization of the driving force divides the whole image into the two regions, one for non-nuclei region and one for nuclei region. This helps in obtaining the contour which stops evolving at the edges of the nuclei. This contour is obtained by dividing the image based on the force obtained in the previous method. The input image is the output obtained from the Winshow version 4.0 application and the contour superimposed on the input image is shown in figure. As it can be seen, the most of the nuclei are detected and along with the nuclei, the red regions are also detected which are non-nuclei.

2.4. MORPHOLOGICAL OPERATIONS

Various morphological operations are applied to obtain a clean mask. Three functions are implemented to clean the mask while retaining the data. These morphological outputs help in retaining the data by preserving the information from the input. Three functions are applied since the data present in one output may not be present in the other output. So combining all the three outputs gives the clean and realistic final result. The three functions are demonstrated below.

i. Small nuclei are retained while removing the large area objects and very small area objects
ii. Large area nuclei objects are retained while removing the small nuclei objects
iii. Difference image between i and ii. .

The pictures below demonstrate the morphological outputs. It is evident that in Figure 5, nuclei with the comparatively small size are retained and in Figure 6 nuclei which have large area are retained. Figure 7 is the morphological output of the difference

of the previous two images. This helps in retaining the medium sized nuclei present. Figure 8 is the combined output and is free of noise which is the final nuclei mask. This nuclei mask when multiplied with the input image, gives the masked nuclei output. The masked nuclei output is shown in Figure 9.

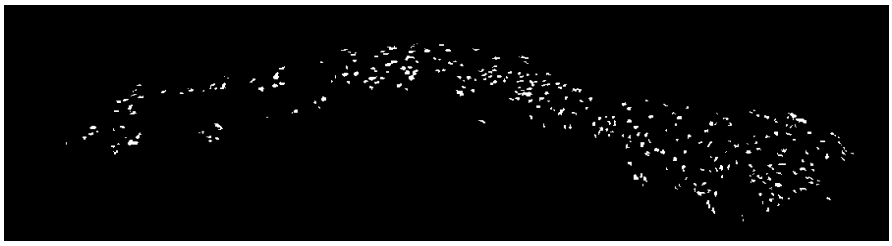

Figure 5: Retaining the small nuclei by eliminating large area objects

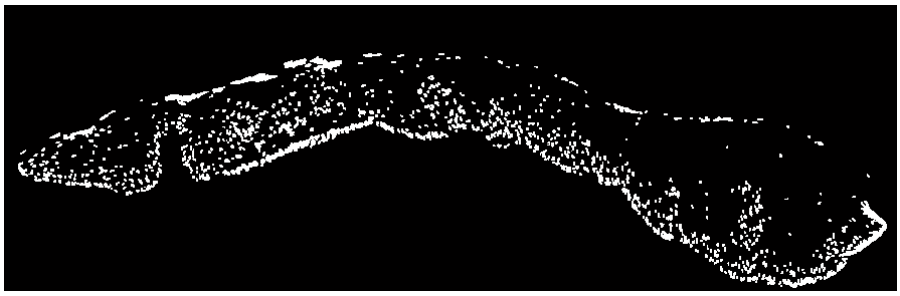

Figure 6: Retaining the large area objects

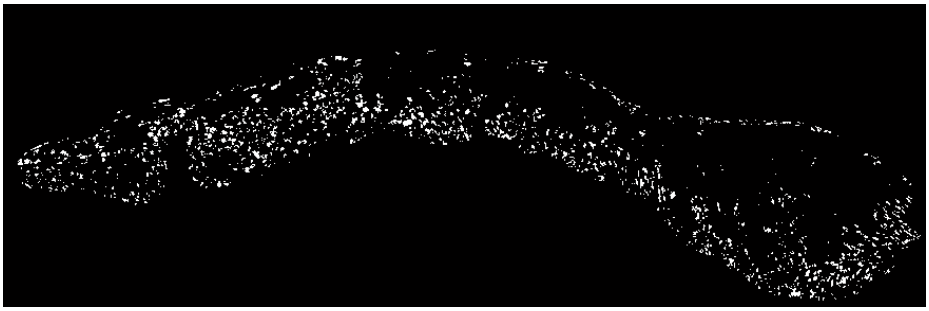

Figure 7: Output of the difference

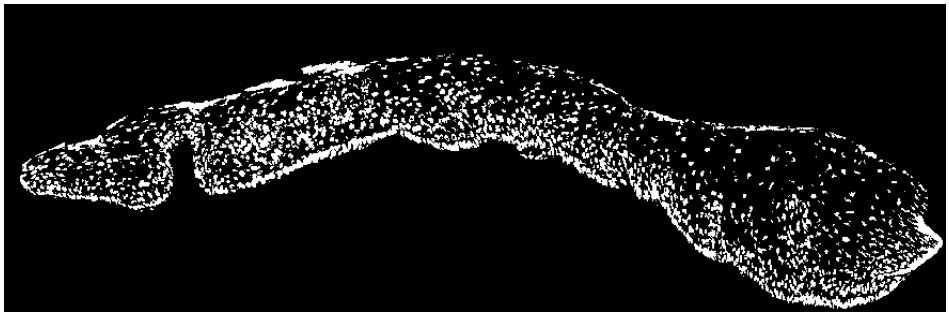

Figure 8: Combined output of the Morphological functions

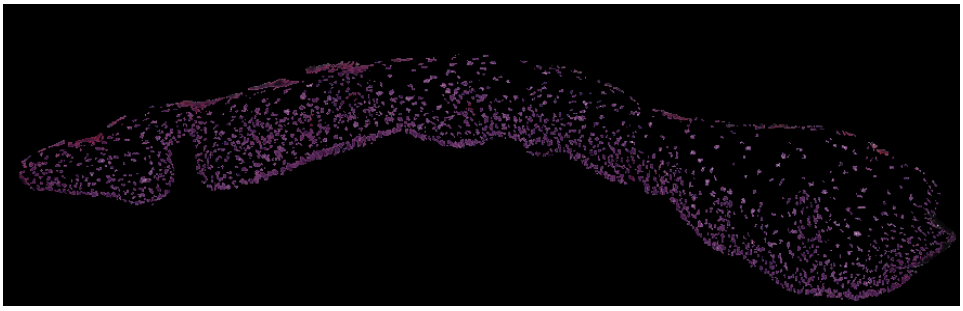

Figure 9: Nuclei Mask

3. EXPERIMENTS AND RESULTS

This algorithm is applied on various images and the accuracy for all the 71-image dataset is calculated. The examples of the true negative and the false positive cases are shown in Figure 10.

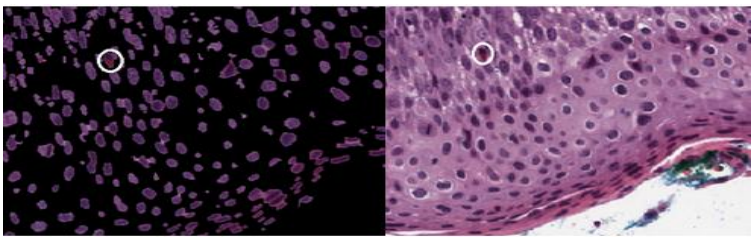

Figure 10: False Positive

As it can be seen the circled area in the masked image is not a nuclei but it is detected as the nuclei. Therefore this considered as false positive result. This is deduced

by comparing the masked image with the original image. Another example illustrates the above the nuclei which are not detected. This example is demonstrated in Figure 11.

As it can be seen, the nucleus which is supposed to be identified is not detected and hence this is false negative result. The true positive results are the cases where the nuclei are truly detected and false negative is the case where the nuclei are falsely detected. These false negative and false positive cases lead to the reduction of the accuracy. Based on these visual deductions, the accuracy is calculated. The best and the worst cases of the algorithm are shown below.

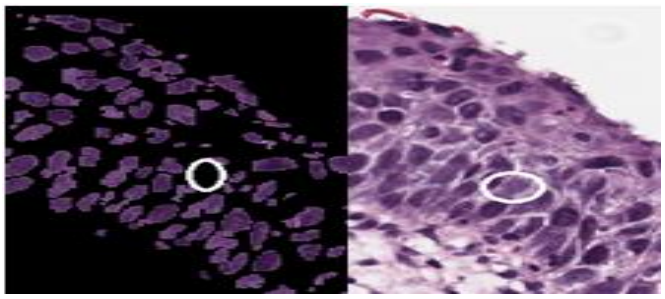

Figure 11: False Negative

Best Case: One of the best case outputs among the experimental data for the 71-Image dataset is demonstrated below. The input image for the algorithm is shown in Figure 12. The nuclei mask generated using the proposed algorithm is shown in Figure 13. The final output of the morphological operations is shown in Figure 14.

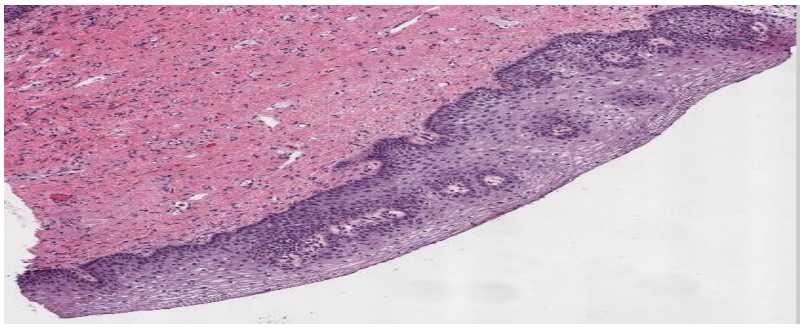

Figure 12: Input image

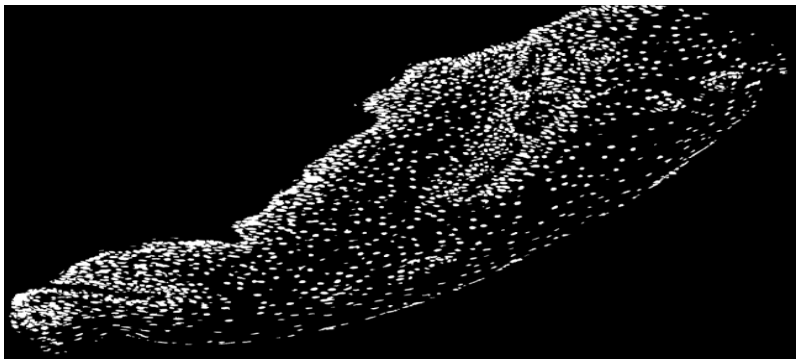

Figure 13: Mask generated

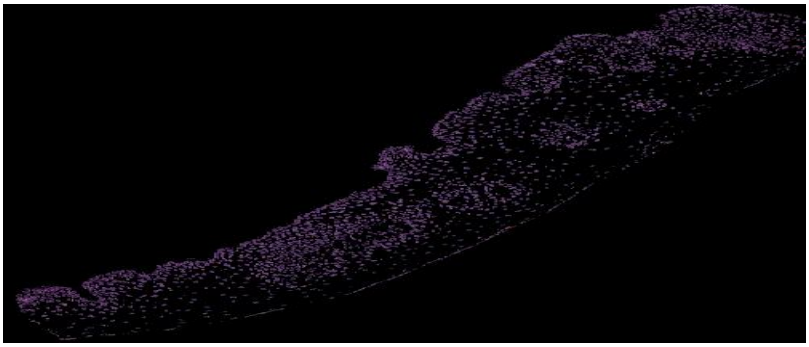

Figure 14: Masked nuclei output

Worst Case: Similar to the previous case, this section demonstrates the worst case of the experimental study. Figure 15 displays the input image of the poor segmentation output. The masked nuclei output is shown in Figure 16 and the final output of the morphological operations is displayed in Figure 17.

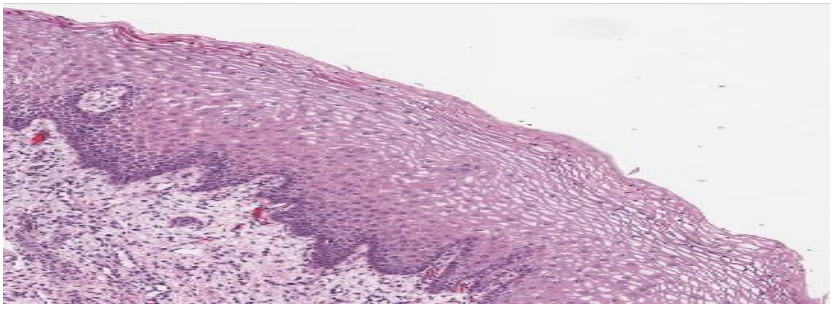

Figure 15: Input image

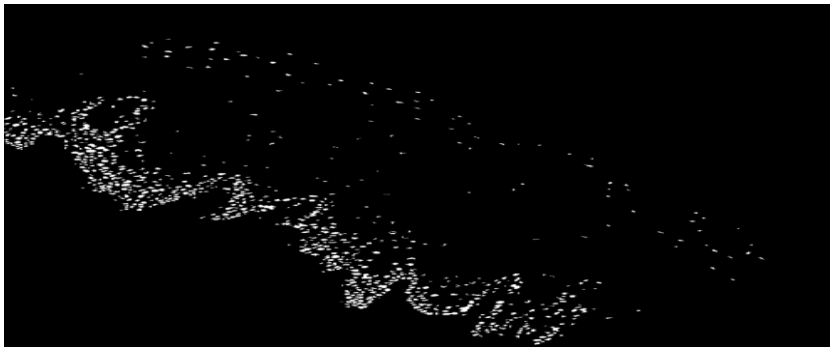

Figure 16: Masked image

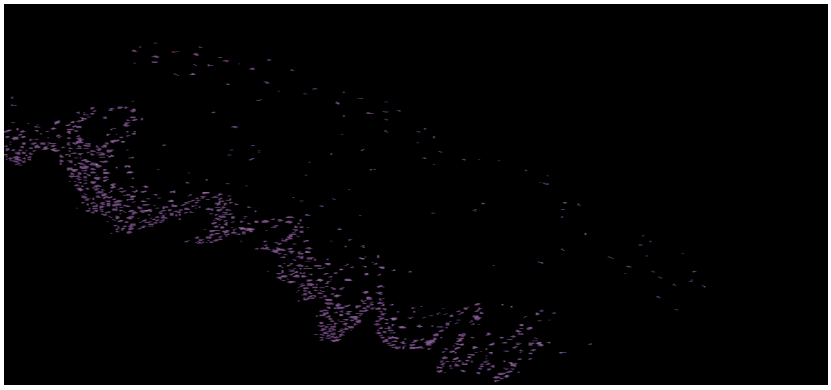

Figure 17: Masked nuclei output

In most of the cases, the nuclei are detected with good accuracy other than for a few input images. This is because of the reduced size of the nucleus. The algorithm is able to detect the nuclei even when they are of smaller size. Morphological operations when applied on the resultant binary mask, eliminate the smaller sized nuclei since they are considered as noise. This can be avoided by improving the functionality of morphological operations.

4. EVALUATION

In order to know how well the nuclei are segmented, accuracy of segmentation is calculated which helps to evaluate the code. Accuracy is calculated by initially calculating the number of objects detected as the nuclei in the final mask. This is done using a small algorithm in MATLAB. Once the number of nuclei is found, the false

negative and the false positive values are obtained from the image (Szénási, Vámossy, and Kozlovszky 2012), these values are used for the determination of the accuracy. Accuracy is calculated using the below given equation. (Szénási, Vámossy, and Kozlovszky 2012)

$$\text{Accuracy} = T_P/(\text{Total Number of Nuclei}) \tag{10}$$

where T_P indicates the total number of nuclei detected i.e. the true positive values, T_N indicates the true negative values and F_P indicates the false positive values from the image. For the image used for demonstration the accuracy obtained is 99.53%. In the experimental results, the accuracy for the best case image is 100% and the accuracy for the worst case image is 84.98%. This accuracy is calculated for all the 71-image dataset.

5. COMPARISON

Much research has been done for the segmentation of the nuclei. As per (Guo et al. 2015), part of their research deals with the nuclei segmentation where the K-means algorithm and a few other morphological operations are used for the segmentation of the nuclei. They have achieved as great as 88.5% labelling and classification accuracy. This paper segments the nuclei based on the fuzzy c-means and level set segmentation method and performs the segmentation of the nuclei on the 71-image dataset obtained from the NLM database. The accuracy of the segmentation achieved is 99.53% on an average. Table 2 demonstrates example values of the accuracy of the input images.

Table 2: Accuracy of Nuclei Segmentation

Total No. Of Nuclei	T_P	F_p	F_N
74731	73791	1662	346

6. DISCUSSION

The results obtained above are the best case and the worst cases obtained using this algorithm. The 100% detection of all the nuclei is achieved when the nuclei are not touching each other and also when they have larger dimension nuclei. In the worst case most of the nuclei are detected that is 84.98% but the other 15% of the nuclei are not detected since the size of the nuclei is too small and while applying the morphological operations, these small masked objects are removed, being considered as noise. If the algorithm is modified which allows the small objects to be retained, then the accuracy of this particular image increased by 10%, but the average percentage accuracy is reduced from 95% to 87%. So to avoid this, the algorithm is not modified and an alternate solution has to be found which doesn't alter the overall accuracy.

REFERENCES

Balla-Arab, Souleymane, Xinbo Gao, and Bin Wang. 2013. “A Fast and Robust Level Set Method for Image Segmentation Using Fuzzy Clustering and Lattice Boltzmann Method.” *IEEE Transactions on Cybernetics* 43 (3): 910–20. doi:10.1109/TSMCB.2012.2218233.

Guo, Peng, Koyel Banerjee, R Stanley, Rodney Long, Sameer Antani, George Thoma, Rosemary Zuna, Shellaine Frazier, Randy Moss, and William Stoecker. 2015. “Nuclei-Based Features for Uterine Cervical Cancer Histology Image Analysis with Fusion-Based Classification.” *IEEE Journal of Biomedical and Health Informatics*, no. c. doi:10.1109/JBHI.2015.2483318.

Krishnan, M. Muthu Rama, Mousumi Pal, Ranjan Rashmi Paul, Chandan Chakraborty, Jyotirmoy Chatterjee, and Ajoy K. Ray. 2012. “Computer Vision Approach to Morphometric Feature Analysis of Basal Cell Nuclei for Evaluating Malignant Potentiality of Oral Submucous Fibrosis.” *Journal of Medical Systems* 36 (3): 1745–56. doi:10.1007/s10916-010-9634-5.

Lu, Zhi, Gustavo Carneiro, and Andrew P. Bradley. 2013. “Automated Nucleus and Cytoplasm Segmentation of Overlapping Cervical Cells.” In *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8149 LNCS:452–60. doi:10.1007/978-3-642-40811-3_57.

Phillips, Cl. 1999. “The Level-Set Method.” *The MIT Undergraduate Journal of Mathematics*, 155–64. http://diyhpl.us/~bryan/papers2/frey/levelsets/Phillips C., The level-set method.pdf.

Rahmadwati, G.N. & Ros, M. & Todd, C. & Norahmawati E., 2011. Cervical cancer classification using Gabor filters. In First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology, pp. 48-52.

Song, Youyi, Ling Zhang, Siping Chen, Dong Ni, Baiying Lei, and Tianfu Wang. 2015. “Accurate Segmentation of Cervical Cytoplasm and Nuclei Based on Multiscale Convolutional Network and Graph Partitioning.” *IEEE Transactions on Biomedical Engineering* 62 (10): 2421–33. doi:10.1109/TBME.2015.2430895.

Szénási, Sándor, Zoltán Vámossy, and Miklós Kozlovszky. 2012. “Evaluation and Comparison of Cell Nuclei Detection Algorithms.” In *16th IEEE International Conference on Intelligent Engineering Systems (INES2012)*, 469–75. doi:10.1109/INES.2012.6249880.

Walker, R F, P Jackway, B Lovell, and I D Longstaff. 1994. “Classification of Cervical Cell Nuclei Using Morphological Segmentation and Textural Feature Extraction.” In *Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, 297–301.

Wang, Lingfeng, and Chunhong Pan. 2014. “Robust Level Set Image Segmentation via a Local Correntropy-Based K-Means Clustering.” *Pattern Recognition* 47 (5): 1917–25. doi:10.1016/j.patcog.2013.11.014.

SECTION

2. DATA FUSION

2.1. DATA FUSION

Segmentation of the nuclei has shown a comparatively good result and this section discusses various data fusion techniques and determines the best algorithm for the features present. These features are called the confidence values obtained from two machine learning algorithms such as SVM and LDA. The dataset used is a 117-image dataset obtained from NLM. Support Vector Machine (SVM) is a classifier which helps in obtaining a hyperplane with largest minimum distance from the data points (support vectors) whereas LDA (Linear Discriminant Analysis) divides the feature space based on the class discriminatory information. These algorithms are implemented and features acquired are the confidence values which are obtained from (Peng Guo, 2016). To implement this algorithm, an input histology image is divided into 10 vertical segments by obtaining the median of the epithelium in the image (De et al. 2013)**.** These 10 vertical segments are fed to the machine learning algorithms, both SVM and LDA respectively. These 10 vertical segments were fed to both the algorithms respectively to obtain the output features. These are the confidence values which are required to be merged to obtain a final classification label. Each vertical segment will have four confidence values where each value represents the probability of the vertical segment belonging to that particular label. Therefore the sum of the four confidence values for a vertical segment will always be unity. Overall for an image there will be 40 confidence values, four for each vertical segment with values ranging from 0 to 1.

Accurate fusion of this data is deciding factor for the final classification label. Voting method is taken as base method for this process and all the accuracies are compared based on the threshold accuracy which is the accuracy of the voting method output. This section starts with PSO (Particle Swarm Optimization) and also tests various machine learning algorithms in weka. Methods with significant accuracy are identified and implemented to obtain the final class label with good accuracy.

2.2. PARTICLE SWARM OPTIMIZATION

This method is a stochastic optimization technique developed based on the observation of the pattern of bird flocking. This helps in obtaining the optimum solution for an objective function. The behaviour of the data points is obtained using simple rules such as Separation, Alignment and Cohesion. Separation allows the particles to avoid collision among the direction of the points with the neighbouring agents, Alignment maintains constant velocity for all the data points and cohesion allows all the particles to stay near the neighbouring agents. This method has various parameters to be balanced. The list of parameters to be balanced is shown below.

Number of particles

Fitness function

Target value/condition

Particle positions

Particle velocities

Personal best and Global best

Stopping value

Number of particles gives the information about the target value and it should be accurate to search the data and locate the target value. The target condition has to be specified by the user. Velocity value indicates the rate of change of data value. Personal best value indicates the closest distance from the target value a particle has reached. Global best is the minimum of all distance values of all particles from the target value. Personal best and Global best values are obtained from the algorithm whereas the remaining parameters have to be configured for a minimized error. This threshold for minimized error is obtained using Stopping value. This helps in terminating the algorithm when the target is not exactly located. The values used in this paper are shown in Table 2.1. Using the computational method for the PSO, initially the current position of the particle is evaluated and then it is compared to the previous best and the global best position. Then based on the result, either this particle imitates or is followed by the other particles. This is a direct search algorithm instead of a gradient search. For each iteration velocities, global best, local best and positions of all the data points are updated.

Table 2.1: Parameters of PSO

Parameter	Value
Velocity parameters	(0.01, 0.0001, 1, 1.6, 0.6)
No. of iterations	400
No. of Particles	40
Hidden layer	18

The values are obtained by a trial and error method and after enough iterations, these values gave an accuracy of 70.09% for SVM confidence values and 71.79% for LDA features. As this accuracy is not a significant value, other methods were implemented on the features to obtain a better accuracy. To obtain a best algorithm, these features are tested in weka against various methods.

2.3. INTRODUCTION TO WEKA

Weka stands for Waikato Environment for Knowledge Analysis. This is a machine learning tool used to test features using various methods for the selection of best method. It helps in preprocessing, classification, clustering and many other algorithms. It is a user-friendly tool that allows the user to change the value of the parameters based on the input requirements. The input format for the features is specific as each image should have a flat line of fixed features as an input. The values of the features can be numeric or nominal. Its output contains an enormous amount of data which contains the accuracy, f-measure, confusion matrix and so on. This data is helpful for the analysis of the output and can also determine if the method is suitable for implementation. This eliminates the time to code various algorithms in order to select the best algorithm. Thus this tool was selected to determine the best algorithm.

2.3.1. Algorithm Selection Using Weka. Features are initially converted to the required input format and are given as an input to Weka. As discussed earlier, feature sets derived from two algorithms (SVM and LDA) are to be analyzed. Both features sets are tested against various methods in Weka. If a particular feature set consistently gives low accuracy when compared to the other feature set then the low accuracy feature set can be

ignored for further analysis. Table 2.2 gives the information about the algorithm and its corresponding accuracy obtained for both SVM and LDA features. As discussed earlier, voting method is the base method for the fusion process. It is necessary to determine if any other algorithm provides a better accuracy when compared to the voting method. Weka enables the user to run 34 algorithms on the feature set. Out of 34 algorithms, 5 algorithms demonstrated comparable accuracy with the accuracy obtained from the voting method. Table 2.2 shows the algorithms and the corresponding accuracy obtained for both feature sets.

Table 2.2: Algorithms and its accuracy using weka

Algorithm	LDA feature accuracy	SVM feature accuracy
Voting Method	79.49	71.79
Naïve Bayes Multinomial	78.63	73.5
BayesNet	74.36	69.23
LogitBoost	74.36	60.6
Naive Bayes Theorem	73.5	72.65
Hoeffding tree	73.5	72.64

As seen from the table the voting method has the highest accuracy and also is the base method for the data fusion. It has an accuracy of 79.49%. As the accuracy obtained here is not higher than the accuracy obtained using voting method, a better way has to be determined to improve the accuracy. One way to obtain better accuracy is to determine the contribution of each segment to the final label. Then the segments with lower contribution can be ignored. The next subsection discusses about the contribution of each segment based on the final label.

2.3.2. Contribution of Vertical Segments. To improve the accuracy, the contribution of each vertical segment classification to the final label is calculated. Initially four confidence values are obtained for each vertical segment by dividing the input feature vector into ten segments. For each segment, the index of the maximum confidence value is taken as the individual class label. This index indicates the

classification of the vertical segment label. The contribution of each vertical segment is calculated by comparing it with the whole image label. This can be obtained by counting the number of times the classification label of that particular vertical segment matches the whole image label. The same is implemented for all ten vertical segments. The results are shown in Table 2.3.

Table 2.3: Contribution of 10 vertical segments

Segment	Percentage (LDA)	Percentage (SVM)
1	39.31	35.89
2	44.44	36.75
3	45.29	42.73
4	43.58	38.46
5	41.88	36.75
6	43.58	43.58
7	48.71	41.88
8	41.88	41.02
9	43.58	44.44
10	47.01	41.02

From the Table 2.3 it is observed that few segments contribute more towards the final class label when compared to the other vertical segments. As seen from the table, LDA features show considerable variance in the contribution of each vertical segment to the classification. As for the SVM features, the significance of each vertical segment doesn't show much variance. Hence the LDA features are used for further analysis. Based on the results obtained using LDA features, the segments 2, 3, 4, 6, 7, 9, 10 are considered for the classification results. These specific segments are used for the classification on LDA features.

2.3.3. Using Weka for Selected Segments. The selected segments are remodeled to obtain the required input format and are tested against various methods in Weka. The methods with best accuracies are shown in Table 2.4.

Table 2.4: Methods with best accuracies

Classifier name	Accuracy of classification
Bayesnet	76.92
Naïve Bayes	75.21
Naïve Bayes Multinomial	80.34
Simple Logistic	75.21
SMO	75.21
IBK	76.0684
Bagging	74.359
RandomSubspace	74.259
Hoeffding Tree	75.21
LMT	75.21

As seen from Table 2.4, the accuracy obtained using Naïve Bayes multinomial gives a significant accuracy of 80.34%. This accuracy can further be increased by implementing (coding) this algorithm and by tweaking the parameters. Therefore from the results obtained above, Naïve Bayes classifier is implemented for the classification of the features.

2.4. NAÏVE BAYES CLASSIFIER

This type of classifier is one of the generative probabilistic models. This method is based on the Bayes rule which calculates conditional probability of a hypothesis given data which is shown below.

$$P(h/D) = P(D/h) * P(h)/P(D) \quad (1)$$

where P(h), P(D) represent prior probabilities of the hypothesis and the data D, P(D/h) indicates the probability of data D given hypothesis h which is also called *likelihood function,* P(h/D) indicates the probability of hypothesis given the data which is also referred to as *posterior probability.* Once the posterior probabilities are calculated, then maximum a posteriori hypothesis is calculated using the formula mentioned below.

$$h_{ML} = \arg\max P(D/h) \tag{2}$$

This hypothesis is obtained assuming P(h_i) = P(h_j) for all $h_i, h_j \in H$. This assumption is called uniform prior and the hypothesis is called maximum likelihood hypothesis. These probabilities are updated automatically with the new training dataset and test dataset. This helps in reduction in the time and space complexity. The outputs obtained are not only the classification result but also gives the probability of the data to be in that particular class. Also this classifier can be used to combine various classifier outputs and obtain the best result among all the classifiers. Therefore this classifier is used on our feature set of confidence values.

The classifier considers the features for a particular image are independent of each other. The implementation described in this paper assumes that all features are independent of each other. It initially calculates the values of the conditional probabilities for all the possible cases on the training dataset such as $P(D/l)$ where l varies from 0 to 3 (indication all class labels) and D indicates a particular data point. The conditional probability is calculated for all data points. Training and testing of the data is done based on the leave-one-out method, which helps in obtaining training and test dataset. For all the iterations, test probability is calculated for the test dataset.

The output obtained for the LDA features using Naïve Bayes method is shown in Figure 2.1. This result is obtained from MATLAB which outputs confusion matrix, accuracy and the total elapsed time for the code to run. The accuracy obtained is 80.34%. Also the confusion matrix suggests that most of the misclassified images are just off by one.

```
confusion matrix:
    37     4     0     0
     5    12     5     1
     0     1    19     2
     0     0     5    26

accuracy = 80.3419%
Elapsed time is 4.882075 seconds.
```

Figure 2.1: Output of Naive Bayes

These results can be analyzed further by obtaining the misclassified samples and comparing the obtained output with the ground truth. As it can be seen, there are 23 misclassified samples, out of which 22 are off by one. In these samples 10 samples are misclassified as a level higher and 11 samples are misclassified as a grade lower than expected. This can be explained based on the input features. For further analysis these misclassified images and their features are examined. Based on the observations it can be concluded that all the segments tend to have the misclassified label instead of its original label. This observation can further be explained by dividing the input features into ten segments and by obtaining the index of the maximum confidence value. Maximum number of indices obtained for all the vertical segments of the image matches the misclassified label. This implies that Voting method as well as Naive Bayes method has most of the misclassified images common. Therefore the input features have to be modified or another method has to be identified to identify the underlying pattern of these confidence values. Solving this problem will be a future work of this project.

2.5. EVALUATION

Evaluation of the Naïve Bayes method is done based on the accuracy and the confidence matrix obtained from the MATLAB output. Validation of the method is done using the leave-one-out validation testing method. The leave-one-out model uses $n-1$ samples (where n is the total sample size) of the data as the training dataset and the remaining one sample is the test dataset. This is iterated for over n times, until all the samples are tested. Once the testing is done, the accuracy is calculated using the below given formula.

$$Accuracy = \frac{(sum\ of\ all\ diagonal\ elements\ in\ a\ confusion\ matrix)}{(sum\ of\ non-diagonal\ elements\ of\ the\ confusion\ matrix)} \quad (3)$$

In a confusion matrix, all the diagonal elements indicate the correctly classified samples and all the non-diagonal elements indicate the misclassified samples.

2.6. COMPARISON

This section started with the Particle Swarm Optimization method for the data fusion technique and concluded with the Naives Bayes Classifier. Naïve Bayes is preferred over PSO because of the increase in accuracy and also the simplicity of the algorithm. Naïve Bayes involves fewer parameters to optimize when compared to the Particle Swarm Optimization. This helps in obtaining global solution when compared to the other algorithm.

3. CONCLUSION

The results obtained above show that the accuracy has significantly increased all through the process. This was obtained by initially selecting the appropriate algorithm for the classification and later by also eliminating the features which do not contribute much to the final output. This can also be done using the *Lasso Regression* which helps in automatically eliminating the irrelevant features from the input. This will be the future work of this data fusion project. Also good input features result in an output with accuracy higher than the present obtained accuracy. Therefore for the future work the feature extraction and the lasso regression for selecting appropriate features will be next step.

REFERENCES

Castanedo, Federico. 2013. “A Review of Data Fusion Techniques.” *TheScientificWorldJournal* 2013: 704504. doi:10.1155/2013/704504.

Crowley, James L., and Yves Demazeau. 1993. “Principles and Techniques for Sensor Data Fusion.” *Signal Processing* 32 (1-2): 5–27. doi:10.1016/0165-1684(93)90034-8.

De, Soumya, R. Joe Stanley, Cheng Lu, Rodney Long, Sameer Antani, George Thoma, and Rosemary Zuna. 2013. “A Fusion-Based Approach for Uterine Cervical Cancer Histology Image Classification.” *Computerized Medical Imaging and Graphics* 37 (7-8): 475–87. doi:10.1016/j.compmedimag.2013.08.001.

Ghosh, Madhumala, Devkumar Das, C. Chakraborty, and Ajoy K. Ray. 2010. “Automated Leukocyte Recognition Using Fuzzy Divergence.” *Micron* 41 (7): 840–46. doi:10.1016/j.micron.2010.04.017.

Guo P, Almubarak H, Banerjee K, Stanley R J, Long R, Antani S, Thoma G, Zuna R, Frazier SR, Moss RH, Stoecker WV. Enhancements in localized classification for uterine cervical cancer digital histology image assessment. J Pathol Inform 2016;7:51

Malpica, Norberto, Carlos Ortiz De Solrzano, Juan Jos Vaquero, Andrs Santos, Isabel Vallcorba, Jos Miguel Garca-Sagredo, and Francisco Del Pozo. 1997. “Applying Watershed Algorithms to the Segmentation of Clustered Nuclei.” *Cytometry* 28 (4): 289–97. doi:10.1002/(SICI)1097-0320(19970801)28:4<289::AID-CYTO3>3.0.CO;2-7.

Sarrafzadeh, Omid, and Alireza Mehri Dehnavi. 2015. “Nucleus and Cytoplasm Segmentation in Microscopic Images Using K-Means Clustering and Region Growing.” *Advanced Biomedical Research* 4: 174–79. doi:10.4103/2277-9175.163998.

Sinha, Neelam, and A G Ramakrishnan. 2003. “Automation of Differential Blood Count.” *The IEEE Region 10 Technical Conference on Convergent Technologies for the Asia-Pacific Region (TENCON 2003)* 2 (i): 547–51. doi:10.1109/TENCON.2003.1273221.

Yang, Lin, Peter Meer, and David J Foran. 2005. “Unsupervised Segmentation Based on Robust Estimation and Color Active Contour Models.” *IEEE Transactions on Information Technology in Biomedicine :* 9 (3): 475–86. http://www.ncbi.nlm.nih.gov/pubmed/16167702.

VITA

Ravali Edulapuram was born in Telangana, India. After finishing high school in 2008, she attended Amrita School of Engineering and graduated with a Bachelor's degree in Electronics and Communication Engineering in the month of May 2014. She also worked for a year and half with Robert Bosch Engineering and Business Solutions. She attended Missouri University of Science and Technology between 2015 and 2017 with Graduate Teaching Assistantship. She received a Master's degree with Missouri University of Science and Technology in Electrical Engineering in May 2017.