Tumor Cell Identification using Features Rules

Bin Fang[1,2] Wynne Hsu[1,2] Mong Li Lee[1]

[1]Singapore – MIT Alliance
National University of Singapore

[2]School of Computing
National University of Singapore
Singapore 117543

{smafangb,whsu, leeml}@comp.nus.edu.sg

ABSTRACT

Advances in imaging techniques have led to large repositories of images. There is an increasing demand for automated systems that can analyze complex medical images and extract meaningful information for mining patterns. Here, we describe a real-life image mining application to the problem of tumour cell counting. The quantitative analysis of tumour cells is fundamental to characterizing the activity of tumour cells. Existing approaches are mostly manual, time-consuming and subjective. Efforts to automate the process of cell counting have largely focused on using image processing techniques only. Our studies indicate that image processing alone is unable to give accurate results. In this paper, we examine the use of extracted features rules to aid in the process of tumor cell counting. We propose a robust local adaptive thresholding and dynamic water immersion algorithms to segment regions of interesting from background. Meaningful features are then extracted from the segmented regions. A number of base classifiers are built to generate features rules to help identify the tumor cell. Two voting strategies are implemented to combine the base classifiers into a meta-classifier. Experiment results indicate that this process of using extracted features rules to help identify tumor cell leads to better accuracy than pure image processing techniques alone.

Keywords

identification, features rules, local adaptive thresholding, dynamic water immersion, meta classifier, majority vote, weighted vote

1. INTRODUCTION

The mechanism of tumor cell metastasis has been the subject of research for many years in pathology. Tumor cells first migrate from the primary tumor, penetrate into the circulation, and eventually colonize distant sites. Knowledge regarding the dissemination of tumor cells has been considered very important in clinical studies of pathology. The quantitative analysis of tumor cells in the filed of pathology forms the fundamental element to characterize the dissemination activity of tumor cells. In traditional pathology, tumor cells are first stained. Then, they are being introduced into experimental animals such as mice.

SIGKDD '02, July 23-26, 2002, Edmonton, Alberta, Canada.

After a few days, tissue specimens containing the stained tumor cells are prepared. A highly trained medical professional will then analyze the tumor cell metastasis by manually counting the number of stained tumor cells in the specimen. This process is very time-consuming and not objective. Figure 1 shows some examples of typical images of tissue specimens with the stained tumor cells.

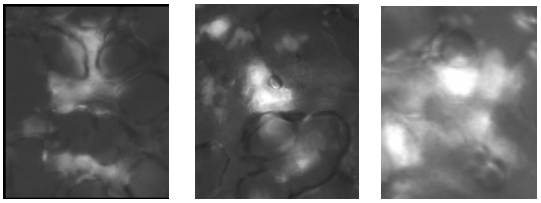
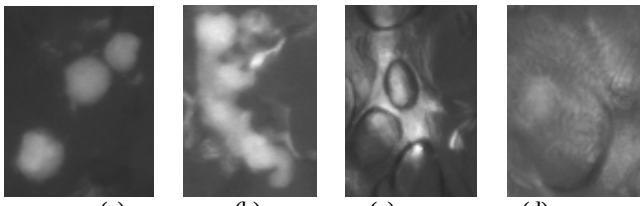


Figure 1. Small portions of a tissue-section histological image

Many of these images exhibit poor contrast with non-uniform background illumination. In particular, the tumor cell boundaries are not sufficiently sharp to be readily extracted and many tumor cells, in fact, overlap with each other to form cell groups. The following observations have been made with regard to the problem of tumor cells identification:

1) High intensity pixels in the form of white patches scattering in the image may have a number of interpretations: they could be the true tumor cells (see Figure 2a), or they could be clumped cell groups where the tumor cells have grown to form a colony (Figure 2b), or they could be histologial noise such as reflection of light on the spherical surface of normal cells (Figure 2c), or they could be psedopods of tumor cells (Figure 2b).
2) The non-uniform background illumination may result in the intensity of the tumor cells being darker than the background intensity at another portion of the image (see Figure 2d).
3) The presence of cell fragments (see Figure 2b) complicates the problem further as their cell boundaries are typically not sharp enough to be readily extracted. In addition, there are many situations where the cells are touching one another that resulted in highly irregular shape (Figure 2b).

There have been a number of attempts to automate the process of tumor cell counting [1-10]. Many of these attempts focus on using pure image processing techniques to identify the tumor cells. However, due to the highly complex nature of the tumor cells images as described above, these approaches have limited success.



(a) (b) (c) (d)

Figure 2. a) Individual tumor cells, b) clumped tumor cell groups, fragments and pseudopods of tumor cells, c) histological noise, d) 'brighter' background.

In this paper, we propose a robust local adaptive thresholding and dynamic water immersion algorithms to segment regions of interest from the background. These regions of interest may correspond to true tumor cells or histological noise such as white light reflection regions and the extended psuedopods of tumor cells. Based on the segmented regions of interests, meaningful features are extracted. Three base classifiers are then built to generate features rules that will differentiate the tumor cells from the histological noise. Further, it is observed that the three different classifiers may potentially offer complementary information about the patterns to be classified. This could be integrated to improve the overall performance of the identification system. Two integration strategies, namely majority vote and weighted vote, have been applied to unify the three base classifiers into a meta classifier. Experimental results indicate that this process of using extracted features rules to help identify tumor cells leads to better accuracy than pure image processing techniques alone, and meta-classifier with weighted vote has the best predictive accuracy.

2. RELATED WORK

The design of an automatic system for medical image analysis in pathology has been the main research objective for many years. There have been many cell segmentation and identification methods that employ image processing techniques to count the number of tumor cells in digitized histological images. They are categorized into statistical classification methods [6,7,9], region growing methods [3,4], and boundary methods [5,8,10].

Awasthi *et al.* [3] used a combination of multiple thresholding, dilation morphology operation and region growing methods to perform cell segmentation. Berns *et al.* [4] used a combination of median filter, local histogram, and morphology filter with watershed method to achieve the same goal of cell segmentation. Gauthier *et al.* [5] used global thresholding, component labeling, morphology filter. Jeacocke *et al.* [6] used a multi-resolution method which contains quadtree smoothing, lowest level classification and boundary re-estimation by water immersion. Chen *et al.* [7] used spatial adaptive filter, watershed and refining of the labeled image. Anoraganingrum [8] used median filter and mathematical morphology operation for edge detection based cell segmentation. Kovalev *et al.* [9] used pattern recognition based on intensity of G image plane and the balance between G and B intensity for color images. Wu *et al.* [10] used cost function based optimal method to obtain a parametric reconstructed image which approximates the original image and threshold such image to segment cells from background.

Although these methods can be simply implemented and have satisfactory experiment results with relatively simple medical images, they work poorly on complex images. No further processing was suggested to improve the ability of automatic cell segmentation. This is the reason that we propose a robust image processing technique to cope with complex histological images and some data mining techniques to help improve system performance.

Recently, research in data mining and knowledge discovery is rapidly gaining popularity in the field of medical image classification. Some of the research and applications carried out use various data mining classification techniques for image categorization. Antonie *et al.* [11] investigated the use of neural network and association rule mining for classifying digital mammongrams into two categories: normal and abnormal images. Hsu *et al.* [12] combined 12 image attributes extracted for each individual vessel segment and fed them into an association based data mining classification tool, CBA to classify the input vessel segments as normal or abnormal in the application of an Integrated Retinal Information system. However, as far as we know, there has been no application of data mining techniques to the problem of tumor cell identification.

3. OVERVIEW OF OUR APPROACH

First, we apply a robust local adaptive thresholding to segment white patches with histological meaning from non-uniform illumination background. Second, we use a dynamic water immersion approach to detach touching cells in clumped cell groups and extract the regions of interesting for further investigation. For each segmented regions, we extract three relevant features of size and shape that accurately describes the characteristics of the region. Next, data mining techniques are applied on the extracted features to derive a number of base classifiers, namely C4.5 classifier [13], Bayes classifier [14], and CBA classifier [15]. Finally, to improve the accuracy of the tumor cell detection problem, a meta classifier with two voting strategies of majority vote and weighted vote is built upon the three base classifiers.

4. FEATURE EXTRACTION FOR MINING

The most crucial step in this tumor cell identification system lies in the accurate segmentation of the regions of interest. Segmentation refers to the process of extracting meaningful regions out from the image background. Such regions typically correspond to objects of interest or their parts. The segmentation of tissue-section histological images into regions that correspond to meaningful biological structures, such as cells and clumped cell groups, is a difficult problem as described in Section 1. In this section, we study the use of adaptive local thresholding method and dynamic water immersion algorithm to perform the segmentation.

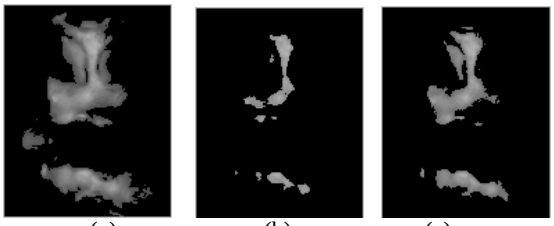
Traditionally, segmentation techniques are categorized into two classes: those that employed region-finding algorithms versus those that employed contour-detection algorithms. Most classical region-finding algorithms involve partitioning the grey level histogram in such a way that the appropriate thresholds for segmentation can be easily determined. However, studies indicate that for complex images such as the histological images, simple thresholding techniques based on a globally determined value will

not work well. One way to overcome this problem is to adopt a global adaptive thresholding method based on the analysis of pixel intensity distribution. According to the mean and variance of histogram, since we want to segment white patches from the background, the threshold can be set as follows,

$$TH = mean + a \times std \quad (1)$$

where *TH* is the adaptive threshold, *mean* and *std* represents the mean and standard deviation of the gray level distribution of all pixels respectively. a is a constant.

Global adaptive thresholding method can effectively segment white patches with histological meaning from background. However, when the images have non-uniform background where white patches at one location in the image can be 'darker' than the background at other locations, then global adaptive thresholding technique is unable to effectively extract all the white patches that are scattering over the image.



(a) (b) (c)

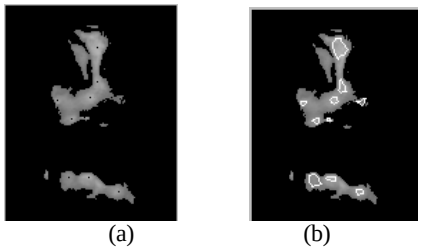
Figure 3. Segmentation results by a) a global low threshold, b) a global high threshold, c) a local adaptive thresholding.

Figure 3a and 3b shows the results of applying the global adaptive thresholding method with a low and a high threshold respectively. Clearly, a relatively low global threshold will lead to a large number of unexpected histological noise as shown in Figure 3a. To prevent extracting the unexpected histological noise, a relatively high threshold is preferred. However, this may lead to some histological meaning patches being excluded because they are even 'darker' than many background patches somewhere else not displayed in Figure 3b. To deal with this problem, we propose the use of a local adaptive thresholding method. The main idea is to apply the global thresholding method to one small area at a time. First, the image is divided into N×N sub-images of equal size. Then, the histogram of each sub-image is computed and the local threshold is determined based on the intensity distribution of pixels in the sub-image. Figure 3c shows the result of applying adaptive local thresholding method.

Having obtained the segmented image, our next task is to separate the clumped cell groups into individual cells so that meaningful features for each cell can be extracted for mining purposes. This problem is made complicated by the fact that the boundaries of the cells are fuzzy and cannot be readily extracted. Furthermore, the shapes of cells are varied due to the presence of cell fragments. Edge detection techniques such as Laplace of Gaussian (LoG) method perform poorly as it tends to lead to broken and disjoined edges which require additional efforts to analyze the output of an edge detection image before boundaries information can be extracted.

Watershed or water immersion algorithm is considered to be a powerful technique for object contour detection. Water immersion algorithm works by grouping pixels with similar gradient information. Direct application of water immersion method to the digitized histological images typically produces over-segmentation of the individual cells. Instead, we propose a dynamic water immersion algorithm to cope with the situation. Details of the algorithm follow.

First, a N×N window is used to locate the local maxima points in the image. For each segmented white patches, we place the center of the window over each pixel in the white patches. If the intensity of the center pixel is the highest with respect to all the other pixels in the window, we say that the center pixel is a local maxima; otherwise, the window will move to be centered at another pixel to continue the search for all local maxima points. At the end of this phase, all the maxima are marked and they will be treated as the starting seeds for water immersion method. One advantage of using the sliding window approach is that with the appropriate window size, it is possible to eliminate a large amount of maxima points that correspond to the light reflection regions thus removing false detection. This is because the intensity level of the maxima points corresponding to the light reflection patches are generally lower than that of potential tumor cells or even the extended pseudo-pods of cells. Given that the distances between the maxima points of the light reflection patches and the nearest maxima points of the neighboring tumor cells are generally less than that between two touching tumor cells, it is possible to set the window size in such a way that these false maxima points are 'absorbed' by the neighboring tumor cell maxima points while the true maxima points corresponding to the touching cells are not affected. Figure 4a shows the effect of choosing a suitable window size on the detection of the true maxima points.



(a) (b)

Figure 4. a) Local maxima marked by dark dots, b) regions of interesting extracted marked by continuous white line.

Having identified the true maxima points, water immersion process starts from the detected maxima points and progressively floods its neighboring pixels. The neighboring pixels are defined to be the 8-direction neighbors. These neighbors are placed in a growing queue structure sorted in descending order of the intensity level of the pixels. The lowest intensity pixel in the growing queue will be 'immersed' first and it is marked as belonging to the same region label as the current seed. The marked pixel is then removed from the growing queue. All neighboring pixels whose intensity level is lower than the marked pixel are added to the growing queue. This progressive immersion process continues until the growing queue is empty.

Unfortunately, simple application of the water immersion technique have the tendency of over-flooding that leads to incorrect shape and size of the potential tumor cells. This can seriously undermine the accuracy of the classifiers in the next step. To overcome this tendency of over-flooding, we apply an additional stopping criterion. In addition to ignoring all

neighboring pixels with intensity level lower than the last placed pixel, we also ignore all those pixels whose intensity level is too low as compared to the last placed pixel. To implement this, we use a dynamically set seed-to-pixel contrast threshold. This threshold is larger for the 'brighter' seed and smaller for the 'darker' seed. This is because from a priori knowledge, we know that the variation in intensity level of pixels neighboring to a 'brighter' maxima point is larger than that to a 'darker' maxima point. The seed-to-pixel contrast threshold is determined using the equation as follows:

$$Con\_th = A \times \frac{I_{\max}}{255} \times e^{\frac{I_{\max}}{255}} \qquad (2)$$

where *Con_th* is the seed-to-pixel contrast threshold, *A* is a constant determined by analyzing the intensity level variation of the tumor cells , and I_{max} is intensity level of the seed.
Figure 4b shows that the proposed algorithm is able to effectively extracted accurate contours of the potential tumor cells, and is able to separate those touching tumor cells. At the same time, a large number light reflection region has been eliminated. However, the algorithm is still unable to eliminate false regions corresponding to the extended psuedopods of cell segments. In order to obtain an accurate count of tumor cells in the histological images, it is necessary to adopt a different strategy. In the next section, we discuss how data mining techniques can be used to help classify these regions of interest into either tumor cells or non-tumor cells.

5. MINING THE TUMOR CELL FEATURES

At this point in the analysis process of histological images, we have segmented the original image into individual regions of interest with different sizes and shapes. For each segmented region, we need to extract relevant information for meaning mining to take place. Although touching cells together with cell fragments make cells in highly irregular shapes, because the process of dynamic water immersion method to extract individual regions of interesting does not follow the irregular shapes of cells but extract potential tumor cell regions with rounder and larger appearance than extracted noise regions such as extended psuedopods, we focused on features related to the size and the shape of each region.

Shape measurements are physical dimensional measures that characterize the appearance of an object. The goal is to use the least number of measurements to characterize an object adequately so that it may be unambiguously classified. Human judgements of shape complexity depend on several factors. Of course, topological factors play an important role, the number of components and holes in the object affect its judged complexity. In our tumor cell identification problem, we assume that the regions are made up of connected set of pixels with no holes inside. In order to guarantee this property, a closing morphological operation has been used to fill all the holes that may possibly exist in the regions of interest. This ensures that each region of interest has a single border. For such region, the size and shape attributes are area of the region, the roundness and the elongation of the region.

The first feature we have found useful relate to the size of the region in the form of the area of the region. Area of the region of interest is defined as the total number of pixels within the region, up to and including the boundary pixels. Another useful feature is the perimeter of the region. It is defined as the arc length of the digital boundary of the region under consideration. The arc length is obtained from its chain code representation [16] with 8-direction connection whereby the horizontal and vertical moves are counted as 1 and the diagonal moves are counted as $\sqrt{2}$. With this definition, we define the roundness of a region as:

$$roundness = \frac{4 \times p \times Area}{Perimeter^2} \qquad (3)$$

Note that the roundness measurement is a value between 0 and 1. The greater the ratio is, the rounder is the object. If the ratio is equal to 1, the object is a perfect circle. As the ratio decreases from 1, the object becomes elongated or irregular. In the real plane, the 'isoperimetric inequality' rule holds that roundness = 1 for any shape. However, in a digital measurement, it turns out that the roundness may be greater than 1 for small area regions.

Another important feature to characterize morphology of a region is the elongation measure. The elongation of a region is defined as the ratio of the width of the minor axis to the length of the major axis. This ratio is computed as the minor axis width distance divided by the major axis length distance, giving a value between 0 and 1. If the ratio is equal to 1, the region is roughly a square or is circularly shaped. As the ratio decreases from 1, the region becomes more elongated. Major axis is the longest line that can be drawn through the region. The two end points of the major axis are found by selecting the pairs of boundary pixels with the maximum distance between them. This maximum distance is also known as the major axis length. Major axis angle is the angle between the major axis and the x-axis of the image. The angle can range from 0° to 360°. Similarly, we define the minor axis length and minor axis angle where the minor axis must maintain perpendicularity with respect to the major axis at all time. The result is a measure of the degree of elongation, *Elong*

$$Elong = \frac{L_{MAJOR}}{L_{MINOR}} \qquad (4)$$

where L_{MAJOR} is the major axis length of the region and L_{MINOR} is the minor axis length of the region.

With the extracted features of each regions of interest, a number of base classifiers are built to help classify these regions into either tumour cells or non-tumour cells. We have selected the three most commonly used data mining techniques to build our team of base classifiers. The first base classifier is the statistical Bayes classifier [14] which incorporates the full covariance matrix. The size and shape features of the regions of interest are not necessarily independent of each other and maintaining the correlation information among them may help to improve the final classification performance. In the training phase, the mean vector and the covariance matrix are computed from the training data to model the Bayes classifier. When a test data comes in, it is assigned to the class with the highest posterior probability. Our second base classifier is based on the decision tree techniques. We use the most widely cited decision tree classification tool, C4.5 [13], to build this base classifier. Our third and final base classifier is based on association rules mining technique. This classifier is built using the Classification-Based on Association (CBA) tool that was first proposed in 1998 [15].

While the individual classifiers do give reasonably accurate predictions of tumor versus non-tumor cells, to improve the predictive accuracy further, we propose the use of two voting

strategies to integrate the base classifiers into a meta-classifier. The first voting strategy is the majority vote strategy. In the majority vote strategy, the meta classifier will output the majority class among all the base classifiers as the final class. For example, in our problem, we have two classes: tumor versus non-tumor. Once test data comes in, if the Bayes classifier identifies it as class tumor, but both the C4.5 and CBA classifiers classify it as class non-tumor, then the majority-vote meta-classifier will assign the test data to class non-tumor. This technique is simple but it does not take full advantage of the strengths of the different classifiers.

It is observed that different classifiers have different abilities in classifying the different classes. For example, the Bayes classifier may be very accurate in identifying the tumor class, whereas the CBA classifier may be very good in identifying the non-tumor class. To take full advantage of their different strengths, a weighted vote strategy for building the meta classifier is proposed. In this strategy, different weights are assigned to the individual classifier's outputs for different classes. Assume the weight of the Bayes classifier when it results in class tumor is α_1, and when it results in class non-tumor is α_2. Similarly, the weight of the C4.5 classifier when it results in class tumor is α_3, and for non-tumor class, the weight is α_4. Finally, the weight of the CBA classifier when it results class tumor is α_5, and for the non-tumor class, the weight is α_6. All these weights are in the range from 0.0 to 1.0. When a test data comes in, if the Bayes classifier identifies it as belong to class tumor, the score of this test data for class tumor is α_1, and at the same time, the score of the test data for class non-tumor is $1-\alpha_1$. If the Bayes classifier identifies it as belong to class non-tumor, then the score of the test data for class tumor is $1-\alpha_2$, while the score for class non-tumor is α_2. In the same manner, we are able to obtain the scores for class tumor and non-tumor from the other two base classifiers. At the meta classifier level, the total scores for the two classes are computed. If the total score for class tumor is greater than the total score for class non-tumor, the meta-classifier will classify the particular test data to be class tumor.

6. EXPERIMENTAL RESULTS

The images used in our experiments are derived from the histological sections containing tumor cells of experimental animals' lungs which had been stained using GFP - green fluorescent protein. Female mice, 4-8 weeks of age, were given injections with a single cell suspension of 2 x 10^6 cells in 100 µl into the tail vein. After 24 or 48 h, the mice were sacrificed and the lungs were frozen in Histo Prep. Ten-micrometer frozen sections were made. The slides were scanned using a digital micrometer (Microcode II; Boeckeler Instruments, Tucson, AZ) to ensure that all areas were counted only once. Green fluorescent cells were confirmed by overlay with a DAPI-stained nucleus. The tissue sections were observed by a Leica inverted fluorescent microscope. A Hamamatsu Orca digital camera was connected to the microscope and linked to a Mac G4. A ×4 objective was used during acquisition. Histological images were captured directly using the digital camera as 8-bit gray-level 1024×1022 TIFF files and stored on hard disk of Mac G4 computer. The histological images were characterized by high intensity pixels in the form of white patches corresponding to potential tumor cells expressing GFP. The proposed methods were evaluated on a database of 20 tissue-section histological images captured under same environment such as subjective magnification, exposure and data format, etc. The resolution of these images is 1024×1022 in 8-bit grey level TIFF format.

During the local adaptive thresholding stage, the images are divided into 3×3 sub-images. After performing some initial studies, we set α in Eqn (1) to 2.35. This value is good enough to retain as many white patches as possible while removing most of the background noise. For the dynamic water immersion algorithm, a window size of 7x7 pixels is used to locate the local maxima. The *Con_A* in Eqn (2) was determined by initial investigation to be 42 for extracting contours of regions of interesting. Figure 4 show an example of the segmented white patches and local maxima marked by dark dots. Contours of regions of interesting extracted by using the dynamic water immersion algorithm are highlighted by continuous white lines.

In our experiments, there are 2850 regions of interest in total. A medical professional had labeled 1704 regions to be tumor cells. This implies that using the proposed image processing technique alone, we are only able to achieve an accuracy of 59.9%, which is far from satisfactory. Next, we perform experiments on the predictive accuracy of the three base classifiers. Here, we use the 10 fold cross-validation testing strategy. The results obtained for 10 splits of the database are summarized in Table 1.

Table 1. Error rates for the 10 splits with the three base classifiers.

	Bayes	C4.5	CBA
Split	Error rate (%)	Error rate (%)	Error rate (%)
1	15.8	15.1	17.5
2	17.5	15.1	15.4
3	23.2	20.4	18.2
4	21.1	20.0	17.9
5	22.1	20.7	20.7
6	14.7	16.8	18.9
7	17.9	17.9	15.8
8	26.0	24.6	21.1
9	30.8	23.2	20.7
10	25.3	26.0	22.1
Average (%)	**21.4**	**20.0**	**18.8**

On average, the CBA classifier has the lowest average error rate of 18.8%. We also noticed that the classification error rates of the CBA classifier for 10 splits of database are more compact and consistent than that of Bayes and C4.5 classifiers. Table 2 summarizes the results of the experiments for evaluating the performance of meta-classifiers using the two voting strategies. In the table, false-positive error rate refers to the misclassification of tumour cell as non-tumour cell while false negative error rate relates to misclassification of non-tumour cell as tumour cell. Here, the average error rate for the majority-vote meta-classifier is 19.0%. On the other hand, the weighted-vote meta-classifier has the lowest average error rate among all the individual base classifiers with an average error rate of 18.7%. This result is achieved at $\alpha_1 = 0.9$, $\alpha_2 = 0.6$, $\alpha_3 = 0.7$, $\alpha_4 = 0.6$, $\alpha_5 = 0.6$, $\alpha_6 = 0.9$. The Bayes classifier has a stronger ability to identify tumor cells, hence it receives greater weight for classifying class tumor cell. Statistical significance evaluations of paired t-tests between the weighted-vote meta-classifier and individual base classifiers

have also been conducted. It was found that the improvement of the meta-classifier over Bayes classifier is significant at 78% confidence level. However the improvement over the CBA classifier is not really statistically significant.

Table 2. Comparison of classification performances of three base classifiers and the meta-classifiers using both majority and weight voting combination strategies.

Classification method	False-positive error rate (%)	False-negative error rate (%)	Average error rate (%)
Bayes classifier	6.6	43.5	**21.4**
C4.5 classifier	12.0	31.8	**20.0**
CBA classifier	14.1	25.8	**18.8**
Meta-classifier (Majority Voting)	10.4	31.7	**19.0**
Meta-classifier (Weight Voting)	14.1	24.8	**18.7**

7. CONCLUSION

There is an increasing demand for automated systems that can analyze complex medical images and extract meaningful information for mining patterns. The quantitative analysis of tumour cells is fundamental to characterizing the activity of tumour cells. In this paper, we describe a real-life image mining application to the problem of tumour cell counting. We propose a robust local adaptive thresholding and dynamic water immersion algorithms to segment regions of interesting from background. Our studies indicate that image processing alone is unable to give accurate results. Therefore, meaningful features are extracted from the segmented regions and the use of extracted features rules is examined by building a number of base classifiers to help identify the tumor cell. Two voting strategies are also implemented to combine the base classifier into a meta-classifier to improve identification accuracy. Experiment results indicate that this process of using extracted features rules to help identify tumor cell leads to better accuracy than pure image processing techniques alone. Meta-classifier with weight voting has the lowest average error rate among all the classifiers.

8. ACKNOWLEDGEMENT

We would like to thank Dr Christopher Wong of Functional Genomics Laboratory, Genome Institute of Singapore, Singapore, for providing us the histological images and valuable explanations of medical background related.

REFERENCES

[1] L. Lam, C.Y. Suen, 'Application of majority voting to pattern recognition : an analysis of its behavior and performance,' *IEEE Trans.Systems Man Cybern.Part A*, Vol. 27 (5), pp. 553 –568, 1997.

[2] J. Kittler, M. Hatef, R. Duin, J. Matas, 'On combining classifiers,' *IEEE Trans. Pattern Anal. Mach. Intell*., Vol. 20 (3), pp. 226 –239, 1998.

[3] Awasthi, K. Vikas, W. Doolitle, G. Parulkar, J.G. MC Nally, 'Cell tracking using a distributed algorithm for 3d image segmentation,' *Bioimaging*, vol. 1, pp. 98-112, 1994.

[4] G. S. Berns, M.W. Berns, 'Computer-based tracking of living cells,' *Experimental Cell Research*, Vol. 142, pp. 103-109, November 1982.

[5] D. Gauthier, M.D. Levine, P.B. Noble, 'Principles of object detection for an automated cell tracking system,' *Image Analysis in Biology*, CH. 2, pp.9-28, CRC press, Florida, 1991.

[6] M.B. Jeacocke, B.C. Lovell, 'A multi-resolution algorithm for cytological image segmentation,' *Proc. 2nd Australian and New Zealand Conf. on Intelligent Information Systems*, pp. 322 –326, 1994.

[7] YM Chen, K. Biddell, AY Sun, P.A. Relue, J.D. Johnson, 'An automatic cell counting method for optical images,' *Proc. BMES/EMBS*, Vol. 2, pp. 819, 1999.

[8] D. Anoraganingrum, 'Cell segmentation with median filter and mathematical morphology operation,' *Proc. Intl. Conf. Image Analysis and Processing*, pp. 1043 –1046, 1999.

[9] V. K. Kovalev, A. Y. Grigoriev, H.-S. Ahn, N.K. Myshkin, 'Segmentation technique of complex image scene for an automatic blood cell counting system,' *Proc. SPIE - Medical Imaging, Image Processing*, Vol. 2710, pp. 805-810, 1996.

[10] H.-S. Wu, J. Gil, J. Barba, 'Optimal segmentation of cell images,' *IEE Proc.-Vis. Image Signal Processing*., Vol. 145, No. 1, pp. 50-56, February 1998.

[11] M.-L. Antonie, O. R. Za¨ýane, A. Coman, 'Application of Data Mining Techniques for Medical Image classification,' *Proc. 2nd Int. Workshop Multimedia Data Mining* (MDM/KDD'2001), pp. 94-101, San Francisco, USA, August 26, 2001.

[12] W. Hsu, L. M. Lee, G. K. Goh, 'Image Mining in IRIS: Integrated Retinal Information System,' *Proc. ACM SIGMOD*, pp. 593, Dellas, Texas, U.S.A., May 2000.

[13] J. R. Quinlan, 'C4.5: program for machine learning,' Morgan Kaufmann, 1992.

[14] K. Fukunaga,'Introduction to Statistical Pattern Recognition,' 2nd Ed, Boston: Academic Press, 1990.

[15] B. Liu, W. Hsu, Y. Ma, 'Integrating Classification and Association Rule Mining,' *Proc. 4th Int. Conf. KD. and DM.* (KDD-98, Plenary Presentation), New York, USA, 1998.

[16] H. Freeman, "On the encoding of arbitrary geometric configurations," *IEEE Trans. Electr. Comput.*, EC-10, pp. 260-268, 1961.