Study on histopathological based on clustering of cervical cancer used minimum spanning tree

SHANG Linjing, Lydia
Shang.bio.sci.ms@lydiashaw.asia

1. Introduction

Histopathology takes an overwhelming role in clinic. Although there are some non-invasive testing and imaging methods to detect cancer at present, histopathology is still an unavoidable means of testing, which is regarded as the 'golden standard' of diagnosing tumors by clinical doctors. Histopathologists, however, use microscopes to observe tissue slices, analyze their structures and diagnose them based on their medical knowledge. It must take a long period to develop an excellent and experienced histopathologist. Moreover, the objectivity of this way of observing pathological images with naked eyes is unstable, which doctors are inevitably affected by experience, workload and emotion. To this end, an increasing number of computer-aided diagnosis (CAD) techniques have been applied to the field of histopathological image analysis.

In this paper, a method based on graphs theory is informed, which can identify information of different organization structures in images by using the characteristics of graphs to help doctors quickly find the lesion area and improve the accuracy of diagnosis. Cervical cancer is one of the most common cancers in woman, hence its research is very representative. In order to apply the important topological information hidden in pathological images to solve the clustering problem

of cervical cancer tissues, the first stage of rough clustering is managed used color features and k-means based on graph theory in this paper. Secondly, applying the skeletonization method, the generated nodes are approximately regarded as the distribution of cervical nuclei. Then, the minimum spanning tree is constructed based on the generated nodes and the geometric features are extracted. Subsequently, k-means clustering algorithm is applied again, then the function of manual correction using mouse is provided to the doctor. In the experiment, the dataset of histopathology of cervical cancer used in this paper has ten whole scanned images, which have achieved considerable results, and has great potential in cancer prevention. For using unsupervised learning, traditional ways to assess the system are not suitable for this study. Therefore, a new judgment of independence test would be assessed the accuracy of computer recognition.

Key words: Cervical cancer; Histopathology image; Graph theory; Unsupervised learning; Independence test

2. Research Background

It is undeniable that, cervical cancer is a globally frequent female cancer. In 2019 the latest cancer report in China released that cervical cancer incidence rate ranked second in Chinese female malignant tumors, which was located after breast cancer(Zhou Hui et al.,2019) . What's worse, the incidence rate of cervical cancer has been increasing consistently in recent years, some developing countries are suffering from it deeply(R Siegel et al.,2017). It is notable that the age of patients is getting younger and younger(Gong Jiaomei, 2013). As for the reason, irregular work and frequent pregnancy may be the causes of this cancer (Feng Ling, 2013).

Therefore, because of the high incidence rate and it is a very typical type of uterine malignant tumor so that the research of cervical cancer has considerable significance, especially its high survival rate and good cure effect would safe more women. Considering of that, an increasing number of researchers are beginning to study it.

The traditional method of pathological image diagnosis is based on the doctor observing the pathological image with naked eyes. At that process, they may just make diagnosis according to their own judgment, which leads to the diagnosis process more time-consuming and energy-consuming, and finally result that there are many uncertain factors affecting judgement. Considering this situation, diagnosis results tend to be very different between individual doctors, and false negative and false positive often occur (Ding Haiyan et al.,2000).

With the rapid development of modern science and technology, computer technology improved continuously so that it is possible to use computer to analyze and diagnose medical

images. Therefore, computer-aided diagnosis technology set off a new round of research upsurge, which was introduced to study cervical cancer histopathology and to help doctors improve the efficiency and accuracy of medical treatment.

However, the application of computer-aided diagnosis in cervical cancer is still a new field. When cervical cancer cells become cancerous, the shape of the cells will change, and the nucleus will become larger shape with the increase of the degree of lesions, which means the morphology, distribution and topological structure of cancer cells will also change(P Sukumarand et al.,2016). For this reason, this feature is helpful to distinguish normal tissue from cancerous tissue so that this research studies its topological structure and uses graph theory to find the lesion area to help doctors diagnose would be effective.

3. Current Study

Before I illustrate my study method specifically, some relative studies are supposed to be retrospected. First of all, the imaging equipment used in pathological images has been further improved, which is the basement of medical imaging. Thanks for the invention of high-precision digital microscope, doctors no longer need artificial cell measurement, and the machine can automatically complete scanning pathological sections and other functions (Yang Yutao et al.,2016). In 2017, Hoogendam P Jacob et al. studied the cervical cancer imaging used High-resolution T2-weighted 7.0-T MRI and proved that it was better than CT imaging (Hoogendam P Jacob et al.,2017).

In Chemical dyeing field, Capillary liquid based cell thin layer staining (Liu Xinmei,

2016),folate receptor-mediated staining (Lin Zuandi et al.,2017), Pap smear staining (Li Xiaolin et al., 2015) and other different staining methods have been successfully used in the relevant detection of cervical cancer. In recent years, researchers have found that the first exon of p16 gene is closely related to cervical cancer. P16 gene expression products increased with the degree of cervical cancer canceration (Zheng Xiaojuan et al.,2011), which has a certain auxiliary effect on the dyeing effect of hematoxylin eosin staining method.

Meanwhile, in medical therapy, the effect of photodynamic therapy in eradicating cervical cancer has been worked out (Chizenga et al., 2019). (Qiang J et al.,2020) studied biotherapeutics developed cancer immunotherapy based on tumor infiltrating lymphocytes technology.

As for studies of machine learning, (Anousouya Devi, M et al.,2016) Combined the artificial neural network and learning vectorization to realize the classification and detection of cancer cells. Mustafa N fused Ann and SVM to identify the pathological image of cervical cancer(Mustafa N et al.,2007). Dong Jianjun(Dong Jian jun et al.,2006) and Ma Jin(Ma Jin et al.,2013) have also made corresponding research on the application of artificial neural network in cervical cancer detection and recognition. Li Wenjie constructed three single classifiers, namely support vector machine (SVM), k-nearest neighbor (k-nearest neighbor) and artificial neural network (ANN), and fused them with fuzzy integral (Li Wenjie,2016). Rahmadwati et al. Used K-means clustering and color segmentation algorithm to classify cervical cancer cells (Rahmadwati R et al.,2012). In 2020, a decision tree combined with SVM,KNN,MLP three methods were used to assess the risk factor of cervical cancer(Jiayu Lu et al.,2020).

As the matter of fact, these papers mentioned are really worth-reading and these authors received a considerable result. However, the missing link is still exist so that the next part I will

illustrate my study design and experiment result.

4. Missing Link

In this paper, Graph theory used to cluster different density area, which involves a lot of mathematical geometry knowledge that has demonstrated by some authors (Bondy J A et al.,1978). These include Normalized Cut (Shij Maljkj, 2000), Minimum Spanning Tree (Felzen Szwalb P F et al., 2004) and principal set based methods (Pavan M et al.,2003).

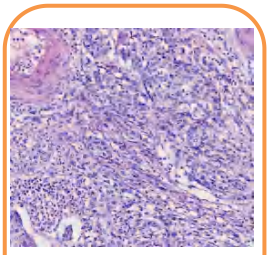
Meanwhile, k-means method would also be used in this paper. Before that, some authors used k-means method which are relevant to my work such as clustering for extract cell edge (Zhao Yinghong et al.,2014), they combined with seed region growth algorithm, successfully segmented cancerous nucleus and cytoplasm. The average gray level of the image as the characteristic parameter was also be used for K-means clustering to obtain an adaptive threshold segmentation algorithm (Guan Tao, 2012). A level set segmentation algorithm was proposed in 2018 (Yu Tingting, 2018). In 2019, an automatic segmentation of cervical region was also used K-means method(Bing Bai et al.,2019).

However, the combination of graph theory and K-means clustering method applied to the study of cervical cancer pathological images has not been published. Although some classifiers can achieve very high accuracy, it mainly trained with the constructed database which means if some new pathological images are input at this time, whether the accuracy can still be sustained is not sure. What's more, patients tend to believe more in the doctor's diagnosis rather than computers. Therefore, a semi-computer and semi-doctor combination is used in my study so that this paper proposes a research method based on graph theory, which uses graph to find the

relationship between points and extracts the structural information of different organizations by its features, and then divides the points by clustering method.

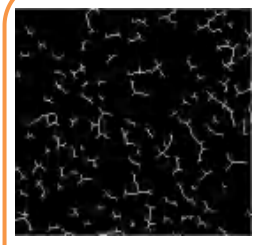
5. Study design

In this paper, the application of computer-aided diagnosis in cervical cancer pathological tissue image was studied. Firstly, due to the pathological images obtained from doctors are very valuable, I expanded data using mirror transform and rotation. Then, some easily image processing in morphological ought to be processed. Rather than thresholding segmentation I chose to use K-means. Color feature of pathological image is extracted, and then the image is processed by K-means for the first time, and then nodes are generated by skeletonization, which is approximately the distribution of cervical nucleus. Then, the minimum spanning tree is constructed based on the generated nodes and geometric features are extracted. I chose four feature values and combined them to a vector which used for cluster. Then, K-means algorithm is applied again and computer can recognize different density area and tagged in different colour. Meanwhile, the function of manual correction is provided to doctors, who can mark the image with the mouse. In order to assess the precision, for unsupervised learning, traditional assessment is not suitable, the experiment is supposed to judge the correct results for data analysis by independent test subsequently. At that time, it is necessary to combine the judgment results of human and computer in this assessment, and get accurate evaluation results through SPSS software. The specific research content is shown in Figure 1 flow chart below.

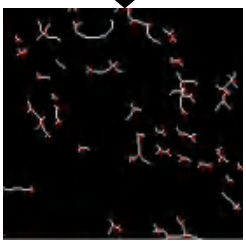


(a)original image

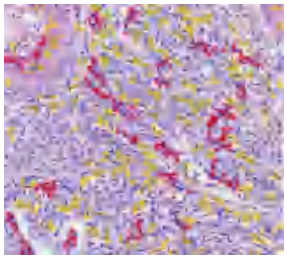
(b)k-means clustering



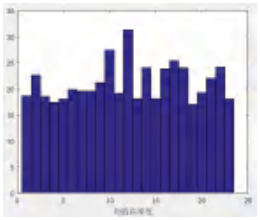
(c)skeletonization



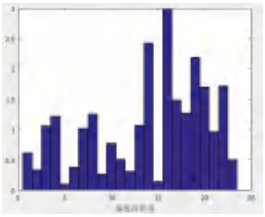
(d)nodes generating



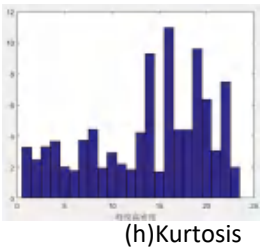
(e)MST



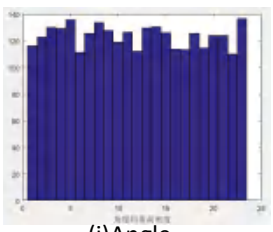
(f)Mean



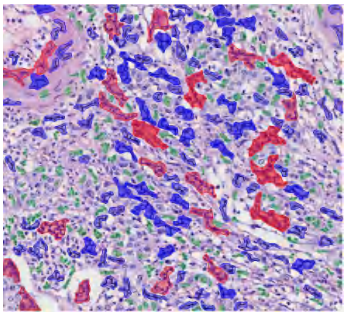
(g)Skewness



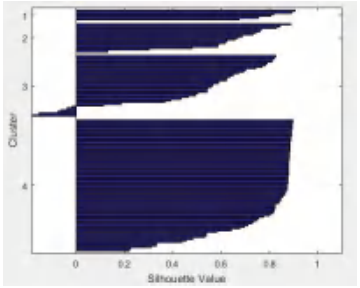
(h)Kurtosis



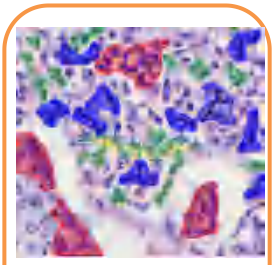
(i)Angle



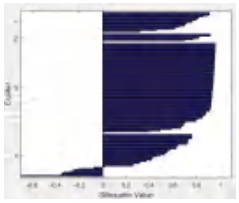
(j)clustering



(k)shadow map



(l)manual annotation



(m)clustering figure

doctors / computers	positive	negative
positive	ture positive H_0	false positive $H_0(\beta)$
negative	false negative $H_0(\alpha)$	true negative H_{0^-}

(n)data analysis

(a) is the original pathological image of cervical cancer. In order to extract color features, K-means is used to make the first clustering result(b). (c)is the result of skeletonization based on (b). (d) It is to generate nodes on the basis of skeletonization.(e) The minimum spanning tree image is constructed by using the generated nodes. Using the minimum spanning tree to extract the proximity relationship and spatial arrangement between the nodes. (f) (g), (H), (I) are the statistical values of some calculated graphs, and these statistical values are taken as the characteristics of graphs to describe the topological spatial structure of images. (j) (k) using graph features and K-means clustering again get more obvious results. (l) (m) manually label the new regions of nuclear formation and classify them into known clusters. (n) is the independent test form.

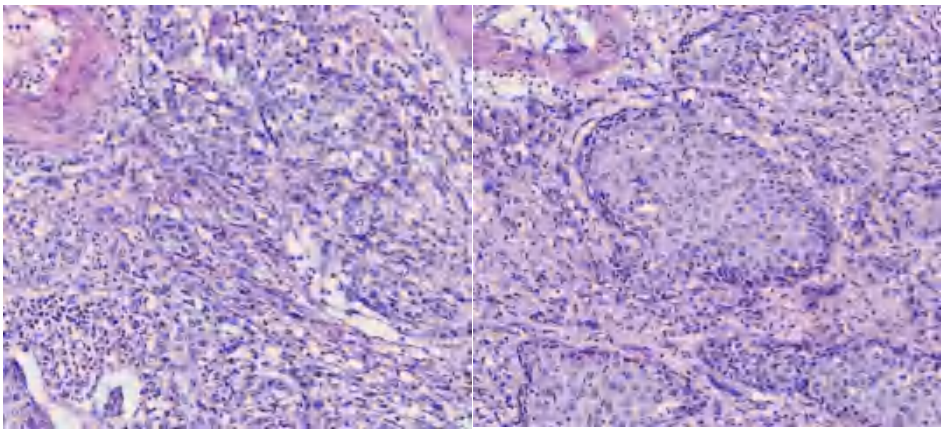
Figure 1 whole process of study design

5.1 Histopathological images of cervical cancer

In tumor research, the diagnosis of histopathology is very important. Doctors use a microscope to observe the tumor section of cervical cancer and analyze the cell distribution to diagnose. Staining the cells helps doctors observe the cells. At present, there are three main staining methods in cervical cancer cells:

HE staining is also known as hematoxylin eosin staining. Hematoxylin can make substances that are easy to react with alkaline substances blue and purple. Eosin acts on eosinophils and dyes them pink. That is, the nucleus will be dyed blue and purple, and the cytoplasm will be purplish red. The dyes used in pasteurization are hematoxylin and orange. Orange can penetrate into the small molecular structure and is suitable for cell plasma. It is a simple method to change eosin and eosin into neutral dye by oxidation (Wei Ce,2015).

In this study, the pixel of cervical cancer pathological image was 976 × 881, and the staining method was hematoxylin eosin staining. It can be seen that the nuclei stained purple in the image present a better staining effect. At the same time, in the second cluster, the adhesion cells were divided into high, medium and low density, and labeled with different colors.



(a) (b)

Fugure 2 Histopathological image of cervical cancer

5.2 Expanded data

The cervical cancer tissue section was prepared and photographed and provided by two histopathologists from Liaoning Cancer Hospital Research Institute. In the experiment, hematoxylin eosin staining was used to color the cells, and then artificial sealing was carried out. Each slice was collected by Olympus digital scanning microscope.

The image storage of Olympus is between 200M and 2024M and pixel is between 30000×30000 and 60000×60000. Because each scanning image is very large and has different sizes, before the experiment, each image is cut into several sub images, and the storage of each sub image is between 1M and 2M, and the pixel value is 976 × 881. At the same time, ten representative sub images are selected as the main research object to test and evaluate the clustering results. Meanwhile, to improve accuracy, the dataset of image is expand by reversing and rotating the image so that more pictures are available to input computer. Figure 3 shows some images expanded in the experiment.

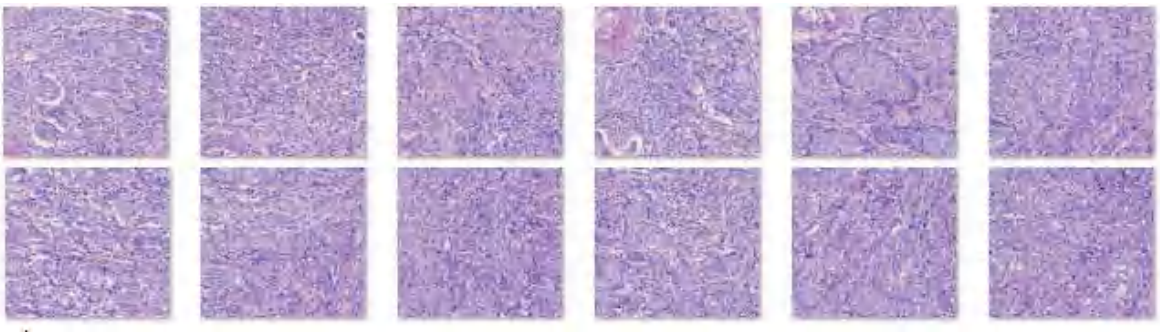


Figure 3 images expanded

5.3 Image processing

In this paper, the image processing used k-means belongs to unsupervised learning, which is often used in association, clustering and dimension reduction. Its input data does not need to be labeled, and there is no definite result. Therefore, the sample data category is often unknown, which is the difference between unsupervised learning and supervised learning.

In the process of unsupervised learning, the samples do not need to be labeled in advance. In fact, in the process of practical application, the label of samples can not be known in many cases. Even the category of training samples may not be available, so unsupervised learning has a wider application direction than supervised learning. The commonly used unsupervised learning algorithms mainly include principal component analysis, isometric mapping, local linear embedding, etc(Yin Ruigang et al.,2016).

Clustering is a typical example, which is used to gather similar things together without any concern about what this category is. The main clustering algorithms are partition method and hierarchical method. Typical partition methods include k-means algorithm, k-medoids algorithm(Kaufman, 1987) and clarans algorithm(Raymond T. Ng et al.,2002). Partition clustering algorithm(Guha, S et al.,1998) needs to divide the data set into k parts and input K as a parameter.

The typical hierarchical clustering methods are birch algorithm and DBSCAN algorithm. Among them, K-means is widely used because of its fast convergence speed and easy implementation(Stricker M et al.,1995).

The k-means algorithm was first proposed by Macqueen(Macqueen J B et al.,1980). This method selects several points in all samples as the initial cluster center, and the selection process is random. Then, the method will calculate the distance between the remaining points and the initial selected points, and assign each sample to the nearest cluster center. In each clustering, the more appropriate cluster center will be updated according to the points in all clusters, and all the samples and centers formed a cluster.

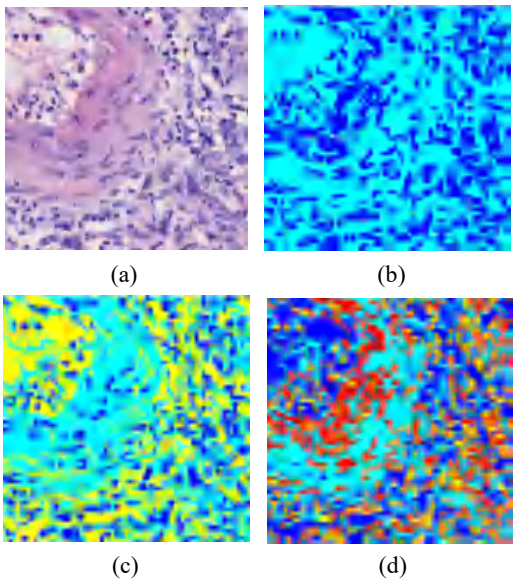


Figure 4 the segmentation in k=2,3,4

In K-means image segmentation, each pixel in RGB color space is regarded as a three-dimensional vector. In K-means clustering, the number of clusters K and the maximum number of iterations n have a certain impact on the image segmentation results. In this paper, different cluster numbers such as k = 2, k = 3, k = 4 have been tried in the experiment, and the maximum number of iterations is 50. In order to make the results more obvious, the pixels of

different clusters are represented by different colors, and the results are shown in Fig. 4

In this study, the histopathological images were stained with hematoxylin eosin. The nucleus was blue purple and the cytoplasm was pink. The tissue fluid without protein and nucleic acid was white. As this study tends to extract features from the nuclear structure distribution to detect lesions, it is not necessary to separate the cytoplasm and tissue fluid. It can be seen from the image that the clustering effect is very good when k = 2, and the structure distribution of tissue fluid is also segmented when k = 3, while the structure information of k = 4 is chaotic and greatly disturbed by noise. After comparing the results obtained by changing the number of clusters and the maximum number of iterations, the image segmentation result with cluster number k = 2 is finally selected and transformed into binary image by simple morphological processing, and the results shown in Fig 5 are obtained.

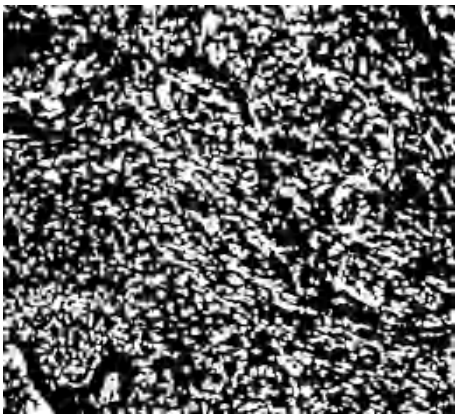


Figure 5 Binary image after morphological processing

As can be seen from Fig. 5, the binary image segmentation effect after morphological processing is more obvious, filling in the small holes in the cell during segmentation, and converting into binary image for subsequent skeleton extraction and other operations.

5.4 Minimum Spanning Tree

Graph theory is a new research hotspot in recent years, which involves a lot of geometry theory knowledge and is a branch of mathematics (Yin Jianhong et al.,2003). It takes graph as research object to study the properties of graph. Graph is a set of multiple nodes V and edge E connecting nodes, which can be expressed as an ordered pair of G = (V, E).

The algorithm of graph theory provides a simple and systematic model, which can remove specific edges and divide the graph into several subgraphs for image segmentation, which plays an important role in the computer field(Chen Xing et al.,2016). Many problems can be transformed into graph theory problems and solved by algorithms. Normalized cut method (Martijn van den Heuvel et al., 2008), principal set based method and Minimum Spanning Tree are commonly used graph theory methods.

The Minimum Spanning Tree problem is the most common network topology problem in graph theory. It often involves identifying the spatial location of a single core, and then constructing the graph as a group of connected core nodes, so that the topological structure or spatial information of the image can be described (Miranda G H B et al.,2012). In addition, more descriptions of the graph information can be obtained from the density and spatial arrangement of these individual nuclei relative to each other(Park M et al.,2009) . Graham and Hell have made a comprehensive analysis on all aspects of MST (R I.Graham et al.,1985). For dense graphs, Prim algorithm(R C Primm,1957) is more suitable, on the contrary, for sparse graphs, Kruskal algorithm(J B Kruskal, 1956) is more suitable.

In this study, thanks for the research of the smallest connected subgraph of Delaunay triangulation, which contains all nodes in the original graph, and the sum of weights is the

minimum (Cruz-Roa A et al.,2015). On the basis of the research, I selected the Minimum Spanning Tree method and applied prim algorithm.

In the adherent cells, they were artificially divided into three categories: high density, medium density and low density. Assume L, M and N are matrices with four connected regions marked on the processed binary image. Set the number of pixels in the marked area as S(i).

(1) Eliminated $S(i) > 500$ in L are regarded as low density regions.

(2) Eliminated $S(i) < 500 \,||\, S(i) > 1700$ in M are regarded as medium density regions.

(3) Eliminated $S(i) < 1700$ in N are regarded as high density regions.

Then, the binary images are skeletonized. It is notable that one must refine the skeleton structure on the basis of ensuring that the skeleton structure is not broken. Subsequently, the three connected region matrices mentioned above are skeletonized as well in Fig.6 are the cytoskeletons extracted from L, M and N respectively connected region matrix.

Figure 6 the cytoskeletons of L,M and N

There are two kinds of shape feature extraction, one is to use edge features, that is, the contour of the shape, the other is to use the whole shape region (Haralick R M et al.1973). The main methods include boundary feature method, Fourier descriptor, shape independent moment and geometric parameter method.

The boundary feature method extracts the required parameters from the boundary features,

and the most classical method is Hough transform(Leavers, V. F,1993) line detection method. In this method, the outer edges of the graph are connected to form a closed shape, and the image is transformed into a parameter space for parameter extraction.

Fourier descriptor method has good geometric invariance. First proposed by Zahn and roskies, it is a widely used edge based description method (Zahn C T et al.,1972). It transforms the outer edge of the figure and transforms the image to the frequency domain to extract features. The curvature function, centroid distance and complex coordinate function can be obtained from Fourier descriptor.

The geometric parameter method is more convenient and direct to describe the regional characteristics. It uses more quantitative measurement of shape, and can get shape area, perimeter and roundness. Eccentricity and other geometric parameters (Zhang Yue,2012). It should be noted that the extraction of shape parameters requires image preprocessing,image processing and segmentation. Therefore, the geometric parameter method is also affected by the effect of image processing.

In this paper, the centroid and skeleton of cell is extracted firstly, and then the shape features of the region perimeter and area are applied, finally combined with the color features which would demonstrate following for the second clustering.

The extracted skeletonized image is attached to the binary image of pathological cells. From figure 7, it can be observed that the bifurcation point is similar to the position of nucleus, which can be used to replace the position of nucleus, and the bifurcation point has certain characteristics and can be extracted.

At the time, the extracted skeleton is traversed to find the point in its eight neighborhood

which is also in the skeleton. When more than three points in the eight neighborhood are also points in the skeleton, then this point must be a bifurcation point, that is, a node. Figure 7 also shows the result of the node generation, and the location of the node is marked in red.

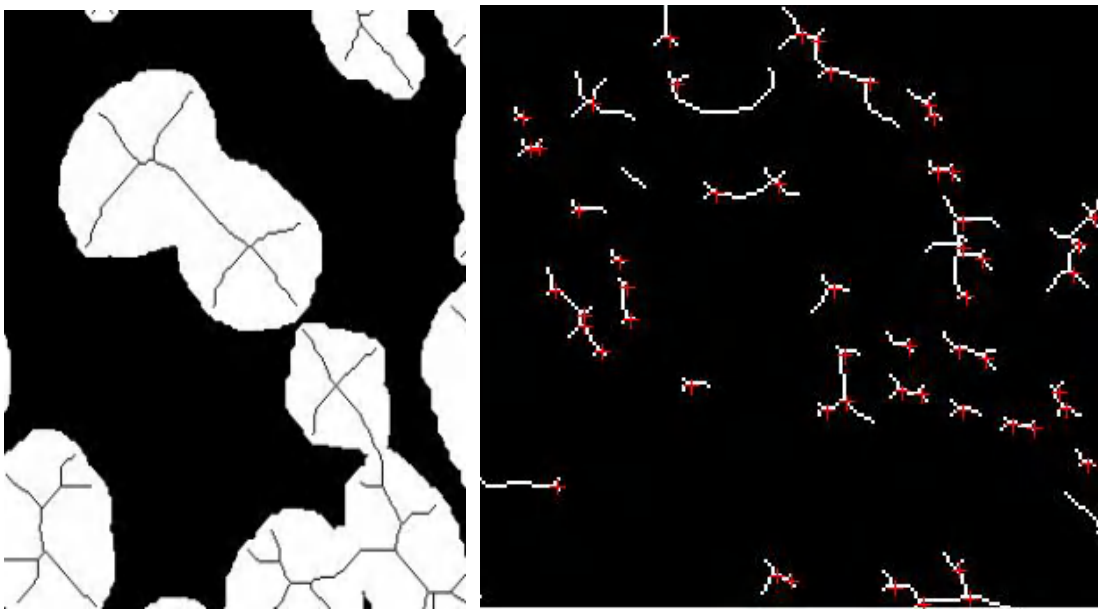


Figure 7 extracted skeleton and nodes generation

Then, Minimum Spanning Tree finally can be constructed. Since most of the low density adherent cells screened are a small number of cell adhesion, in this section, only a large area of adhesion in medium density and high-density is constructed, and the visualization results of minimum spanning tree construction are shown in Fig. 8. In this section, the Minimum Spanning Tree is constructed for high density and medium density areas respectively. This method of separation processing and separate construction can help distinguish different types of organizations. In order to make the construction results more intuitive and recognizable, images finally represented by two different color lines. The black line builds the minimum spanning tree for the medium density area, and the blue line builds the minimum spanning tree for the high density area.

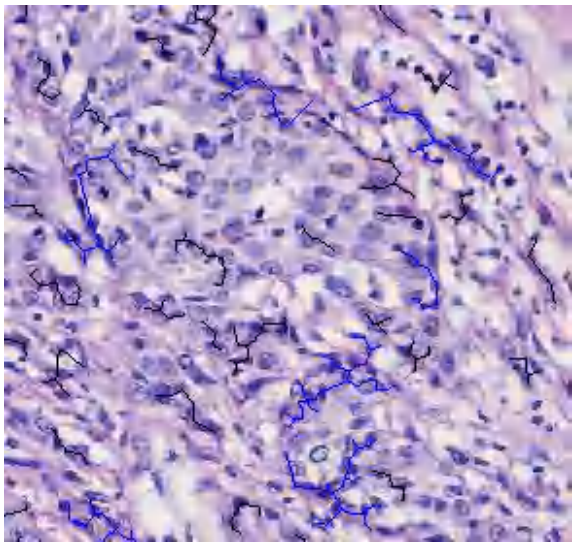


Figure 8 Minimum Spanning Tree

5.5 Feature extraction

In this study, the color moment originates from the probability density matrix. It was first proposed by stricker and orengo. It is a simple and effective representation method(Stricker M et al.,1995). The information distribution of images can be described by multi-order moments, which not only contain color information, but also contain more image information.

However, actually, the color information mainly exists in the low order moments, so the color moments mainly use the first three moments. The advantage of this method is that it does not need to count the pixel values in the space like the color histogram, and the dimension of the feature vector is also low.

Suppose that there are n pixels in the image, the mathematical definition of color moment is as follows:

Mean is used to measure the average intensity of a color.

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} p_{i,j}$$

Variance is used to measure the uniformity of image color distribution.

$$\sigma_i = \left(\frac{1}{N} \sum_{j=1}^{N} (p_{i,j} - \mu_i)^2 \right)^{\frac{1}{2}}$$

Skewness is used to measure the symmetry and distortion of the overall distribution.

$$s_i = \left(\frac{1}{N} \sum_{j=1}^{N} (p_{i,j} - \mu_i)^3 \right)^{\frac{1}{3}}$$

Kurtosis is used to describe the steepness of image color distribution.

$$\gamma_i = \left(\frac{1}{N} \sum_{j=1}^{N} (p_{i,j} - \mu_i)^4 \right)^{\frac{1}{4}}$$

Where $p_{i,j}$ is the i-th color component of the j-th pixel in the image. i can be taken as 1,2,3, which represent RGB color components respectively. Because the multi order moments of color moments have more geometric meanings besides color information, the extracted features such as mean, slope and kurtosis can also be used as geometric features. At the same time, the multi-level matrix can still be used in the gray image with color component 1.

In this study, firstly, color histogram was used to extract color features, and then sparse nuclei were clustered to separate adherent nuclei. Then, the fourth moment is used to extract more image information in the minimum spanning tree, and the second clustering is carried out. Therefore, color moment is not only for color feature extraction, but also can be used for deep-seated geometric feature extraction.

Based on the graph generated by the Minimum Spanning Tree, I calculated various statistics used the functions illustrated above and used them as shape and geometric features to describe different organizations. The shape features extracted in this study include the mean, variance, skewness and kurtosis of edge length, as well as the angle of each figure. Geometric features

include the perimeter of the organization and the independent arrangement of nodes in each organization.

Sparse cells are extracted in the first clustering. In the subsequent processing of adherent cells, the graph features and set features extracted by the Minimum Spanning Tree are used. Afterwords, K-means clustering is used again to set K values for two types of tissues, and then more detailed results are obtained through the second stage clustering to predict the cancer risk of tissues and quickly find the lesion location. According to the clustering results of the graph features extracted from the minimum spanning tree, k = 2 is selected to cluster the high density and medium density, and the results are shown in Fig. 9

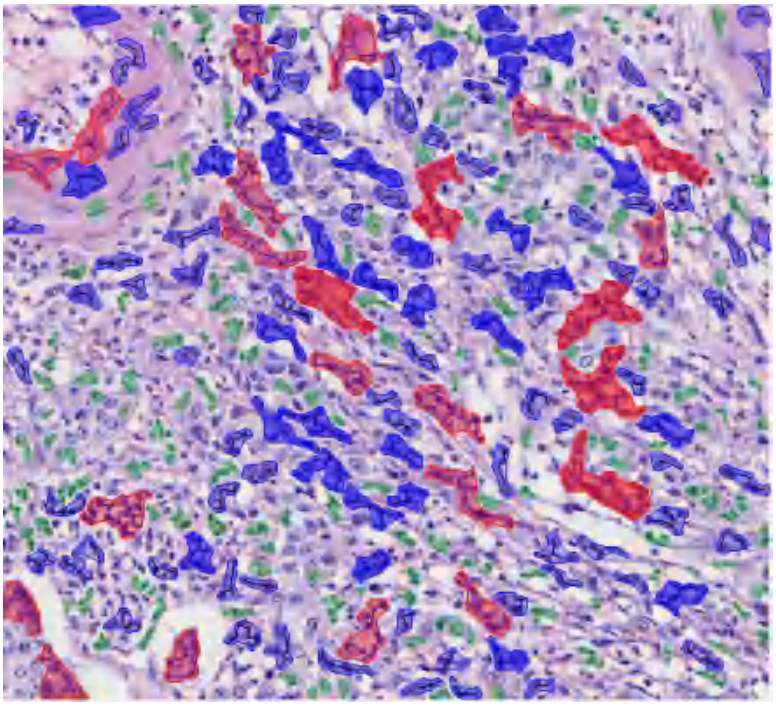


Figure 9 Clustering result

5.6 Manual annotation

The clustering results shown in Figure 9 can help doctors to quickly detect lesions and predict cancer risk. In addition, this paper also provides doctors with the function of manually

labeling nuclei to mark dense areas or suspicious lesions. The yellow area shown in Figure 10 is the result of manual annotation. After the doctor points out the nucleus, it will generate a color marked area, and the nucleus will be connected by lines to form a graphic structure, so as to extract features and add it to the cluster of adherent cells.

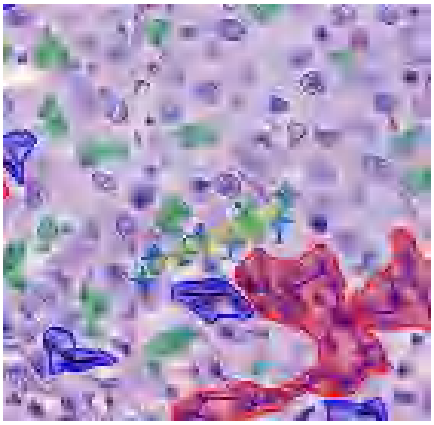


Figure 10 Manual annotation

6. Data analysis

As figure 9 shown, there are three different colours. The red one means cell high density areas, which are affirmed to a tumor area by computers. The blue one means medium density areas, which are suspected to cancer cells field. The green one means low density area which means it is potential to become cancer cells. Using these colours, doctors can quickly focus on the density area and save much more time.

6.1 Result analysis

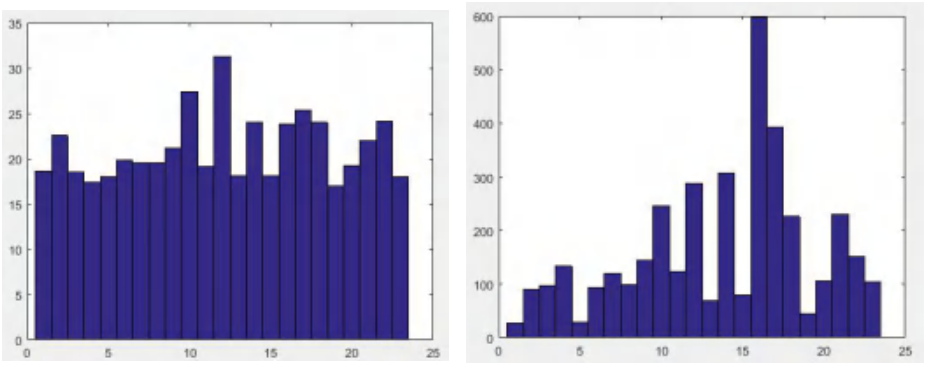
Traditionally, precision, accuracy, and recall are calculated to judge a system. However, for unsupervised learning, there are not specific value for calculating. Therefore, how to assess a unsupervised system is hard.

$$\text{pr}ecision = \frac{TP}{TP + FP}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

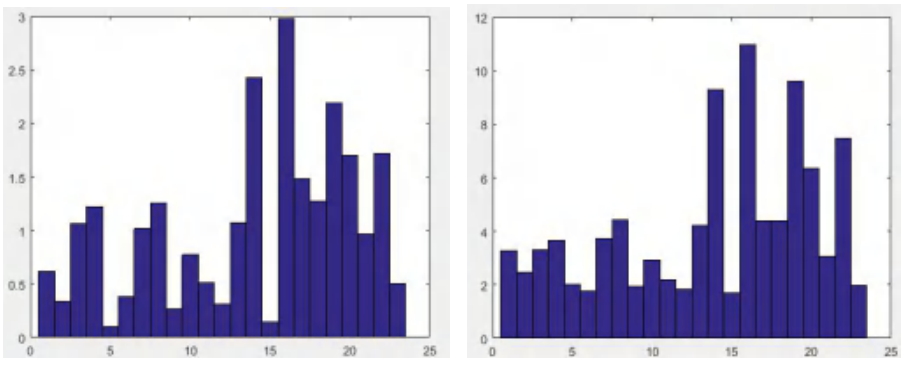
$$recall = \frac{TP}{TP + FN}$$

In order to analysis the results of this experiment, for the study used clustering method which belongs to unsupervised learning, the statistical data are calculated by the functions, and the visual results are presented. The visualization results of the extracted graphical features are shown in Fig. 11-14. It can be seen from the chart that in the same histopathological image of cervical cancer, the information of high density adherent nuclear tissue is more obvious, and its number is less than that of medium density nucleus.



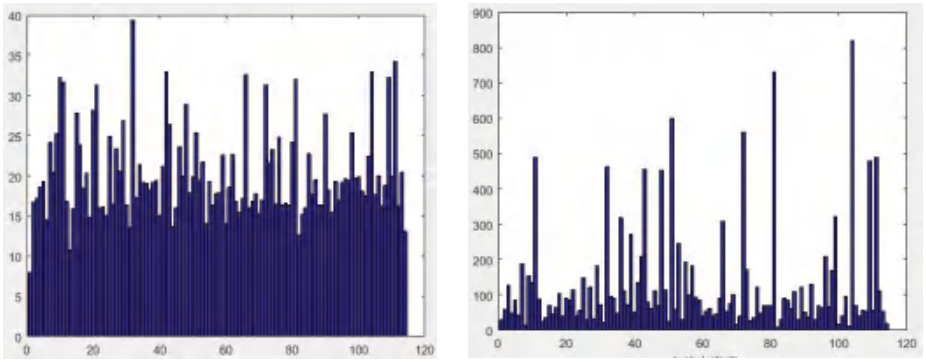
(a) (b)

Figure 11 Mean and variance of high density tissue



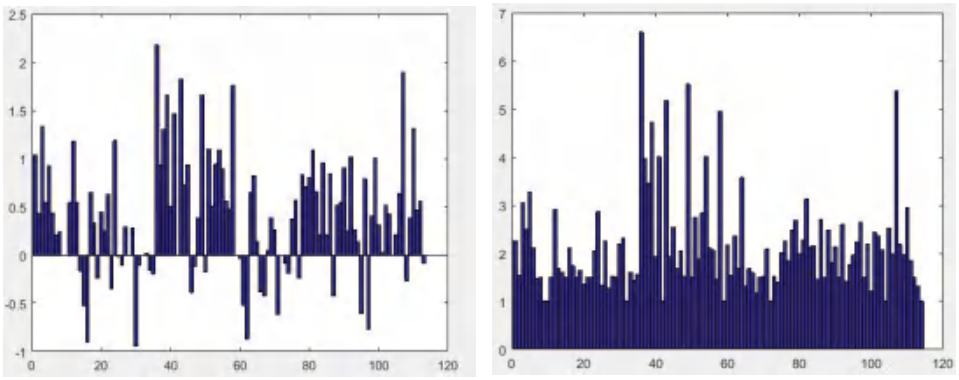
(c) (d)

Figure 12 Skewness and kurtosis of high density tissue



(e) (f)

Figure 13 Mean and variance of medium density tissue



(g) (h)

Figure 14 Skewness and kurtosis of medium density tissue

6.2 Independent test

Although figures 11-14 have shown specific result, compared with precision, accuracy, and recall used in supervised learning, histogram evaluation does not quantify the experiment result. Therefore, the independence test is considered to judge the accuracy. According to the form ,the question of accuracy of computer recognition turn out to the correlation between doctor judgment and computer decision so that I can assume a null hypothesis: there is no correlation between computer decision and doctor judgment. And what is supposed to do is prove the computers decisions are more like humans, that means the correlation larger, the accuracy higher. And then, by recording the results of each graph and listing the table, P value can be calculated for assessing

this study.

doctors / computers	positive	negative
positive	ture positive H_0	false positive $H_0(\beta)$
negative	false negative $H_0(\alpha)$	true negative H_0

The following are some final results after running. The software used in this study is MATLAB, and the variation environment is Windows 10. By recording the results of doctor and computer identification, I analyzed them with SPSS software. From the results shown in fig.19, the correlation between them is very high, especially in the low density region,which can acquire a considerable result.

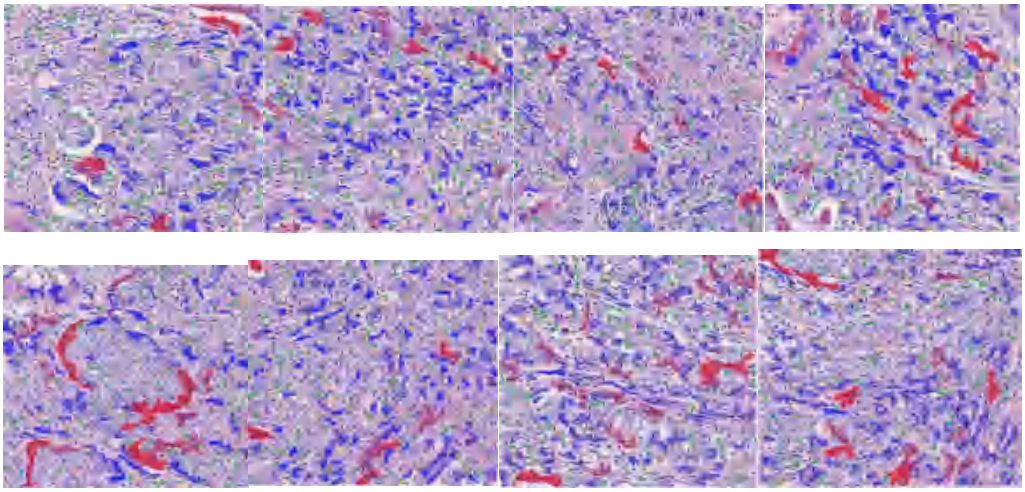


Figure 15 Final results

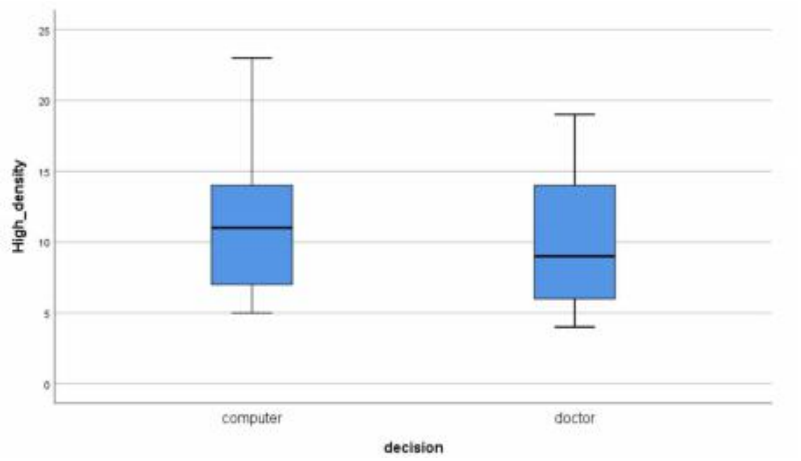


Figure 16 Box Whisker of High density

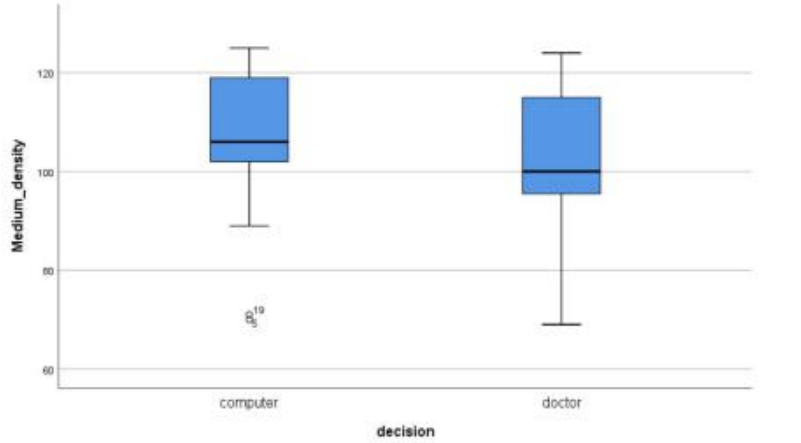


Figure 17 Box Whisker of Medium density

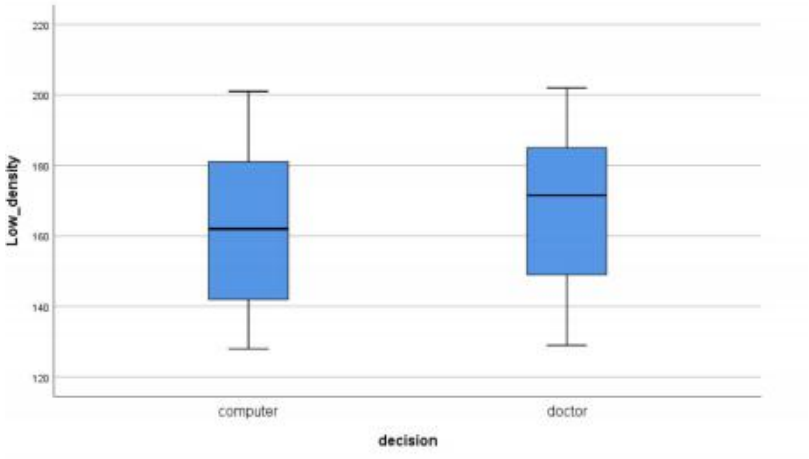


Figure 18 Box Whisker of Low density

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
High_density	Equal variances assumed	.291	.593	.920	37	.364	1.534	1.668	-1.845	4.913
	Equal variances not assumed			.916	35.470	.366	1.534	1.675	-1.864	4.932
Medium_density	Equal variances assumed	.001	.976	.677	37	.502	3.492	5.155	-6.953	13.938
	Equal variances not assumed			.677	36.858	.502	3.492	5.157	-6.957	13.942
Low_density	Equal variances assumed	.109	.743	-.151	35	.881	-1.115	7.383	-16.102	13.873
	Equal variances not assumed			-.151	34.394	.881	-1.115	7.361	-16.069	13.839

Figure 19 result of independent test

7. Reference

[1] Zhou Hui, Wang Dongyan, Luo Ming & Lin Zhongqiu. (2019). "FIGO 2018 gynecological cancer report" - Interpretation of cervical cancer guidelines. Chinese Journal of Practical Gynecology and obstetrics, 035 (001), 95-103

周晖, 王东雁, 罗铭 & 林仲秋. (2019). 《figo 2018 妇癌报告》——子宫颈癌指南解读. 中国实用妇科与产科杂志, 035(001), 95-103.

[2]Siegel, R. L. , Miller, K. D. , Fedewa, S. A. , Ahnen, D. J. , & Jemal, A. . (2017). Colorectal cancer statistics, 2017. CA A Cancer Journal for Clinicians, 67(3), 104-17.

[3]Gong jiaomei. (2013). Research on new screening techniques for cervical cancer

巩姣梅. (2013). 宫颈癌筛查新技术研究. (Doctoral dissertation, 郑州大学).

[4]Feng Ling & Yang Huzhen. (2013). Application of TCT, HPV DNA typing and colposcopy in cervical cancer screening. Contemporary medicine, 000 (025), 19-21.

冯玲, & 杨湖珍. (2013). Tct、hpv-dna 分型及阴道镜在宫颈癌筛查中的应用. 当代医学, 000(025), 19-21.

[5] Ding Haiyan, sun Yungao, & Ye Datian. (2000). Automatic identification of cervical cell smears by computer. Foreign medicine: Biomedical Engineering (2).

丁海艳, 孙允高, & 叶大田. (2000). 计算机自动识别宫颈细胞涂片技术. 国外医学:生物医学工程分册(2).

[6] P., Sukumar, R., K., & Gnanamurthy. (2016). Computer aided detection of cervical cancer using pap smear images based on adaptive neuro fuzzy inference system classifier. Journal of Medical Imaging and Health Informatics, 6(2), 312-319.

[7] Yang Yutao, Zhang Fan &Zeng Li (2013). A retrospective study on quantitative analysis of cell DNA in cervical cancer screening. China maternal and child health, 28 (14), 2304-2306.

杨玉涛,张帆&曾莉(2013).细胞 DNA 定量分析技术在宫颈癌筛查中的回顾性研究.中国妇幼保健,28（14）,2304-2306.

[8]Hoogendam, J. P. , Kalleveen, I. M. L. , De Castro, C. S. A. , Raaijmakers, A. J. E. , Verheijen, René H. M., & van den Bosch, Maurice A. A. J., et al. (2017). High-resolution t2-weighted cervical cancer imaging: a feasibility study on ultra-high-field 7.0-t mri with an endorectal monopole antenna. European Radiology, 27(3), 938-945.

[9] Liu Xinmei. (2016). Study on the application value of thin layer staining of capillary liquid-based cytology in cervical cancer screening. Chinese and foreign medicine, 035 (001), 54-55.

刘新梅. (2016). 探讨毛细式液基细胞学薄层染色技术对宫颈癌筛查应用价值研究. 中外医疗, 035(001), 54-55.

[10] Lin zhuandi, Peng Yiqiong, Liu Yongzhu, Xu Yin, & Qin Bifang. (2017). Comparison of folate receptor mediated cervical special staining and liquid based cytology in cervical cancer screening. Journal of chronic diseases (04), 466-468.

林钻娣, 彭奕琼, 刘永珠, 许茵, & 覃碧芳. (2017). 叶酸受体介导的宫颈特殊染色法和液基细胞学在宫颈癌筛查中的应用比较. 慢性病学杂志(04), 466-468.

[11] Li Xiaolin, Shi Junmei, Liu Tao, Li Guohui, Zhang Qingwen, &Feng Ni. (2015). Comparative analysis of cell DNA quantitative analysis and Pap staining cytology in the diagnosis of cervical cancer and precancerous lesions. Guangxi Medical (04), 67-69.

李晓琳, 师俊梅, 刘涛, 李国晖, 张清文, & 冯妮. (2015). 细胞 dna 定量分析法与巴氏染色细胞学诊断宫颈癌及癌前病变的对比分析. 广西医学(04), 67-69.

[12] Zheng Xiaojuan, Hu Xinrong & Tang Zeli. (2011). Preliminary study on gene changes of p16 and eIF4E in cervical cancer. Journal of practical cancer (1), 13-15.

郑晓娟,胡新荣&唐泽立.(2011). p16 与 eIF4E 在宫颈癌中基因改变情况的初步研究. 实用癌症杂志(1),13-15.

[13] Chizenga, E. P. , Chandran, R. , & Abrahamse, H. . (2019). Photodynamic therapy of cervical cancer by eradication of cervical cancer cells and cervical cancer stem cells. Oncotarget, 10(43).

[14] Qiang J., L. I. , Guru, K. , Lugade, A. , Brese, E. , & Chatta, G. . (2020). Abstract A05: Expansion of tumor-infiltrating lymphocytes (TIL) using Iovance's Gen 2 process from advanced bladder cancer for adoptive immunotherapy. Abstracts: AACR Special Conference on Bladder Cancer: Transforming the Field May 18-21, 2019 Denver, CO.

[15] Anousouya Devi, M & Subban, Ravi&Vaishnavi, J & Punitha, S. (2016). Classification of Cervical Cancer Using Artificial Neural Networks [J]. Procedia Computer Science, 89. 465- 472.

[16] Mustafa N,Isa N A M& Mashor M Y.(2007).Colour Contrast Enhancement on Preselected Cervical Cell for ThinPrep® Images, International Conference on International Information Hiding and Multimedia Signal Processing. IEEE Computer Society, 209- 212.

[17] Dong Jianjun (2006). Cervical cancer cell recognition based on artificial neural network (Doctoral dispersion, Shenyang University of Technology).

董建军. (2006). 基于人工神经网络的子宫颈癌细胞识别. (Doctoral dissertation, 沈阳工业大学).

[18] Ma Jin, & Cao Yang (2013). Research on the detection of cervical cancer cells by artificial neural network model. Electronic world, 000 (008), 214-214

马瑾, & 曹阳. (2013). 人工神经网络模型在子宫颈癌细胞检测方面的研究. 电子世界, 000(008), 214-214.

[19] Li Wenjie. (2016). A multi classifier fusion method for single cervical cell image segmentation, feature extraction and classification recognition

李文杰. (2016). 一种多分类器融合的单个宫颈细胞图像分割、特征提取和分类识别方法研究. (Doctoral dissertation).

[20] Rahmadwati, R. , Naghdy, G. , Ros, M. , Todd, C. , & Tescher, A. G. . (2012). Computer aided decision support system for cervical cancer classification. , 8499, 849919.

[21]A, J. L. , A, E. S. , C, A. G. B. , & B, M. A. . (2020). Machine learning for assisting cervical cancer diagnosis: an ensemble approach. Future Generation Computer Systems, 106, 199-205.

[22] Bondy, J. A. , & Murty, U. S. R. . (1978). Graph theory with application. The Mathematical Gazette, 62(419).

[23] Shij& Maljkj.(2000). Normalized cut and image segmentation. IEEE Trans on Pattern Anlysis and Machine Intelligence,22(8),26-3.

[24] Felzen Szwakb P F&Huttenlocher D P. (2004).Efficiet graph-based image segmentation.

Intemational Joural of Computer Vision,59(2),167-181.

[25] Pavan, M. , & Pelillo, M. . (2003). A new graph-theoretic approach to clustering and segmentation. IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE.

[26] Zhao Yinghong, Hong Yaling, & Sun cunjie. (2014). Cervical cancer cell segmentation method based on K-means clustering algorithm. Clinical medical engineering (9), 1089-1090.

赵英红, 洪雅玲, & 孙存杰. (2014). 基于 k 均值聚类算法的宫颈癌细胞分割方法. 临床医学工程(9), 1089-1090.

[27] Guan Tao, Zhou Dongxiang, Liu Yunhui, & Cai Xuanping. (2012). Cervical cell image classification algorithm based on adaptive threshold segmentation. Signal processing, 28 (009), 1262-1270.

关涛, 周东翔, 刘云辉, & 蔡宣平. (2012). 基于自适应阈值分割的宫颈细胞图像分类算法. 信号处理, 28(009), 1262-1270.

[28] Liu Jun, Yu Tingting, Shi Huijuan, & Lu Han. (2018). Cervical fluorescence polygenic dark region segmentation method based on registration image and level set algorithm. Chinese Journal of medical imaging, 26 (09), 61-68.

刘君, 余婷婷, 石慧娟, & 陆晗. (2018). 基于配准图像与水平集算法的宫颈荧光多生暗区分割方法. 中国医学影像学杂志, 26(09), 61-68.

[29] Bai Bing,Liu PeiZhong,Zhao Yong,Du Yan& Luo Ming.(2018). Automatic segmentation of cervical region in colposcopic images using K-means.Australasian Physical & Engineering Sciences in Medicine,41,1077–1085.

[30] Wei Ce & Hou Xiangping. (2015). Application value of two different staining methods in liquid based cytology (TCT). Chinese and foreign women's health research, 13.

魏策& 侯向萍.(2015). 两种不同染色方法在液基细胞学(TCT)检查中的应用价值.中外女性健康研究, 13.

[31] Yin Ruigang, Wei Shuai, Li Han, Yu Hong. (2016). An overview of unsupervised learning methods in deep learning. Computer system applications (8), 1-7.

殷瑞刚,魏帅,李晗,于洪. (2016). 深度学习中的无监督学习方法综述. 计算机系统应用(8), 1-7.

[32] Kaufman, L. and Rousseeuw, P.J. (1987), Clustering by means of Medoids, in Statistical Data

Analysis Based on the Norm and Related Methods, edited by Y. Dodge, North-Holland, 405–416.

[33] Ng, R. T. , & Han, J. . (2002). Clarans: a method for clustering objects for spatial data mining. IEEE Transactions on Knowledge & Data Engineering, 14(5), 1003-1016.

[34] Guha, S. , Rastogi, R. , & Shim, K. . (1998). Cure : an efficient clustering algorithm for large databases. Information Systems, 26(1), 35-58.

[35] Peng,Y.,Park,M.& Xu,M.(2010).Clustering Nuclei Using Machine Learning Techniques. IEEE/ICME International Conference, 52–57.

[36] Lee, H. B. , & Macqueen, J. B. . (1980). A k-means cluster analysis computer program with cross-tabulations and next-nearest-neighbor analysis. Educational & Psychological Measurement, 40(1), 133-138.

[37] Yin Jianhong & Wu Kaiya. (2003). Graph theory and its algorithm. China University of science and Technology Press.

殷剑宏, & 吴开亚. (2003). 图论及其算法. 中国科学技术大学出版社.

[38] Chen Xing & Li Jun. (2016). A survey of graph based segmentation algorithms. Computer and digital engineering, 44 (010), 2043-2047.

陈杏, & 李军. (2016). 基于图论的分割算法研究综述. 计算机与数字工程, 44(010), 2043-2047.

[39] Van den Heuvel M, Mandl R, Hulshoff Pol H .(2008). Normalized Cut Group Clustering of Resting-State fMRI Data. PLoS ONE 3(4): e2001.

[40] Miranda, G. H. B. , Barrera, J. , Soares, E. G. , & Felipe, J. C. . (2012). Structural analysis of histological images to aid diagnosis of cervical cancer.

[41] Park M , Jin J S &Xu M. (2009).Microscopic image segmentation based on color pixels classification. Proceedings of the First International Conference on Internet Multimedia Computing and Service. ACM, 2009.

[42] GRAHAM, RL, & HELL. (1985). On the history of the minimum spanning tree problem. ANNALS HIST COMPUT.

[43] Primm R C .(1957).Shortest connection networks and some generalizations.Bell systems Technology Journal,36,1389-1401.

[44] Kruskal J B .(1956).On the shortest spanning tree of a graph and the traveling salesman problem.Proceedings of the American Mathematical Society,7,48-50.

[45] Cruz-Roa A , Xu J &Madabhushi A . (2015).A note on the stability and discriminability of graph-based features for classification problems in digital pathology. 10th International Symposium on Medical Information Processing and Analysis. International Society for Optics and Photonics, 2015.

[46] Haralick, R. M. , Shanmugam, K. , & Dinstein, I. . (1973). Textural features for image classification. Studies in Media and Communication, SMC-3(6), 610-621.

[47]Leavers, V. F. . (1993). Survey: which hough transform. CVGIP Image Understanding, 58(2), 250-264.

[48] Zahn, C. T. , & Roskies, R. Z. . (1972). Fourier Descriptors for Plane Closed Curves. IEEE Computer Society.

[49]Zhang Yue (2012). Research on building materials image retrieval technology based on color and texture features.

张跃. (2012). 融合颜色和纹理特征的建材图像检索技术研究. (Doctoral dissertation, 武汉理工大学).

[50] Stricker, M. A. , & Orengo, M. . (1995). Similarity of color images. Proceedings of SPIE - The International Society for Optical Engineering, 2420, 381--392.