Journal of Artificial Intelligence Practice (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2025.080206
ISSN 2371-8412 Vol. 8 Num. 2

A two-stage cervical pathology cell detection model based on YOLOv7x and K-means

Siyu Chen, Jingren Wang

School of Information and Intelligent Engineering, Sanya College, Sanya, China

Keywords: Deep learning; Cervical pathological cell detection; YOLOv7x; K-means clustering algorithm

Abstract: As a malignant tumour that seriously threatens women's health, early diagnosis of cervical cancer is crucial to improve the cure rate. Cervical pathological cell detection is a key link in the early diagnosis of cervical cancer, but the traditional detection methods have the problems of low efficiency and accuracy relying on manual work. In this study, we combined the YOLOv7x target detection model with K-means clustering algorithm to construct a two-stage model YOLOv7x-CH for cervical pathological cell detection. To address the problems in the detection of cervical pathological cells of High-grade Squamous Intraepithelial Lesion (HSIL) category, we firstly constructed the To address the problems in the detection of HSIL, we first constructed the YOLOv7x baseline model, extracted the nucleoplasmic ratio features based on fine segmentation and grey scale symbiosis matrix texture features from the HSIL samples, and then constructed the two-stage model YOLOv7x-CH through K-means clustering. The experimental results show that compared with the baseline model, the two-stage model YOLOv7x-CH can improve the accuracy of cervical pathological cell detection from 0.56 to 0.813, and the average accuracy AP from 0.556 to 0.591, which effectively improves the accuracy of cervical pathological cell detection, and provides a more reliable technical support for the early diagnosis of cervical cancer.

1. Introduction

Cervical cancer is a major disease that seriously threatens women's health, and its morbidity and mortality rates are among the highest among female malignant tumours. According to authoritative statistics, a large number of women are diagnosed with cervical cancer every year, and early cervical cancer has a high cure rate if it can be detected and treated in time [1]. Cervical cytology screening is an important means of early diagnosis of cervical cancer, but currently the test mainly relies on the observation and diagnosis of cervical cell images under the microscope by clinicians or pathologists. Since cervical cell images contain a large number of cells, the detection process is extremely energy-consuming. Therefore, it is of great practical significance to develop an efficient and accurate cervical pathology cell detection algorithm to assist pathologists in diagnosis.

This study focuses on the problem of cervical pathological cell detection, with emphasis on HSIL (High-grade Squamous Intraepithelial Lesion) cervical pathological cell detection. The main contents include: firstly, a baseline model was constructed based on YOLOv7x[2] to analyse its

performance and problems in cervical pathology cell detection. According to the prediction results of the baseline model, it was found that a large number of background regions were predicted to be HSIL-type targets, which led to a low detection accuracy P and average accuracy AP of HSIL-type detection samples. The design scheme of this study to address this problem is as follows: the HSIL class detection frames output from the trained baseline model YOLOv7x are filtered - the original image regions corresponding to the HSIL class detection frames are extracted as 21-dimensional feature vectors, which include nucleoplasmic ratio features [6] based on fine segmentation [3-5] and 20-dimensional grey scale covariance matrix texture features [7-9] - these features are cellular image features that have been proved to be effective by many works; the above set of feature vectors is clustered into two classes using the K-means [10] algorithm [11], and the distance between the clustering centres of the two classes and the representative feature vectors of the labelled samples of the HSIL class (i.e., the clustering centre) [12], the set of samples corresponding to the cluster centre with small distance is determined as True HSIL class, and the rest as False HSIL class; the True HSIL class detection samples are used as the HSIL class detection frames in the final output of the model. After verifying that the above operations are effective, this study encapsulates the above operations into an independent sub-module, the filtering sub-network Cluster-HSIL, and cascades it over the YOLOv7x model [13] to construct the two-stage model YOLOv7x-Cluster-HSIL (abbreviated as YOLOv7x- CH). After systematic experimental analysis and optimisation of key parameters in the two-stage model - such as nucleoplasmic ratio features, grey scale covariance matrix eigenvector dimensions and attributes, and the number of clustering centres of HSIL labels - the two-stage model designed in this study YOLOv7x-CH effectively improves the model's detection performance for HSIL class samples.

In conclusion, this study combines YOLOv7x and K-means algorithm to propose a two-stage model YOLOv7x-CH for HSIL sample detection, and performs feature extraction and K-means clustering on HSIL samples output from the baseline model, which effectively solves the problem of low detection accuracy and low average accuracy of HSIL samples from the baseline model constructed by using the YOLOv7x architecture. The problem of low detection accuracy and average accuracy of HSIL samples using the baseline model built by YOLOv7x architecture is effectively solved.

2. Theoretical and technical basis

2.1 Analysis of cervical pathology cell dataset

In this study, we used a real cervical pathology cell dataset from a maternal and child health centre in a city. This dataset contains 7409 cervical microscope images cropped from whole-slide images (WSI), with image sizes ranging from 975×581 to 2100×1253. There are a total of 51,500 annotation frames in the dataset, covering the following categories: high-grade squamous intraepithelial lesions with atypical squamous cells (ASC-H), low-grade squamous intraepithelial lesions (LSIL), high-grade squamous intraepithelial lesions (HSIL), high-grade squamous intraepithelial lesions (HSIL), high-grade squamous intraepithelial lesions (HSIL), high-grade squamous intraepithelial lesions (HSIL), and high-grade squamous intraepithelial lesions (HSIL). intraepithelial lesions (HSIL), squamous cell carcinoma (SCC), atypical glandular cells (AGC), and trichomonas. In terms of data division, 6665 images were randomly selected as the training set and 744 as the test set in this paper. There is a difference in the distribution of the number of samples in each category, in which the number of samples in the HSIL category accounts for about 60% of the total number of samples. This distribution may have an impact on model training and performance evaluation, e.g., during model training, categories with a larger number of samples may be more

easily learned and recognised by the model, while categories with a smaller number of samples may face the problem of insufficient learning.

2.2 Performance evaluation index system

In this thesis, the evaluation metrics of intersection and integration ratio (IOU), precision (P), recall (R), accuracy, P-R curve, average precision (AP), and mean average precision (mAP) are used in the performance evaluation index system. The intersection and integration ratio (IOU) visualises the degree of overlap between the detection frame and the real label frame by calculating the ratio of the intersection and integration area of the two, and the closer the value is to 1, the higher the accuracy of the model in locating the target. Precision rate (P) is used to measure the accuracy of the model's detection results, and refers to the ratio of the number of correctly detected positive samples to the number of all detected positive samples. Recall (R) reflects the model's ability to detect actual positive samples, i.e., the number of correctly detected positive samples as a proportion of the number of all actual positive samples. The precision rate is the proportion of correctly predicted samples in the total number of samples. The P-R curve takes the recall rate as the horizontal axis and the precision rate as the vertical axis, and updates the precision rate and recall rate by sorting the detected samples in descending order of categorical confidence and calculating the IOU values of the detection frames and the Ground-truth, and the area surrounded by the curves and the axes is the average precision AP value. The AP value integrally reflects the performance of the model under different recall rates, and the AP value comprehensively reflects the model's performance in the precision rate under different recall rates. The higher the AP value, the better the model performance. The mean average precision (mAP) is used for the multi-category detection task and is the average of the AP values for each category. These evaluation metrics quantitatively assess the model performance from different dimensions, providing a strong basis for model optimisation and comparison.

3. Two-stage model design

This study builds the model using Python 3.7 and PyTorch 1.12 deep-learning framework. The experimental hardware is an NVIDIA GeForce RTX 2080 Ti GPU, an Intel (R) Xeon (R) Silver 4214 CPU with 80G RAM and 10G video memory. The software environment uses the Ubuntu 20.04.2 operating system, CUDA 11.0, and cuDNN 8.1.0 to ensure the efficiency and stability of model training and running.

3.1 Baseline Model

When constructing the YOLOv7x baseline model, the input image resolution is set to 640×640×3, the batch size is 8, and the number of iterations is 50. The SGD optimiser is combined with the OneCycle learning rate adjustment strategy, with an initial learning rate of 0.01, a cycle learning rate of 0.1, a learning rate momentum of 0.937, and a weight decay coefficient of 0.0005. Mosaic and Mixup data are used to augment the model, with the weights of 1 and 0.3, respectively, and the model uses the binary cross-entropy (BCE) loss function to calculate the category probability and the target confidence. The BCE loss function is used to calculate the category probability and target confidence score loss, and the Softmax function is replaced by multiple independent logistic classifiers for the classification loss.

Training the YOLOv7x baseline model based on the above parameters and analysing the prediction confusion matrix (refer to Figure 1) revealed that a large number of backgrounds were misclassified as HSIL foreground classes, with 1,816 False Positive (FP) detections, which is

similar to the number of True Positive (TP) samples correctly detected by HSIL, which is 1957. This is due to the different criteria for discriminating abnormal cells between medical labels and the labels of the target detection task, and the large proportion of HSIL-type samples and complex cell morphology in this experimental dataset, which led to the problem of difficult samples, which in turn led to the lower P-value of HSIL-type detection accuracy. Trying to use focus loss to replace BCE loss, the results showed that focus loss could not significantly improve the model performance, so it was decided to carry out post-processing operations for HSIL-type detection samples.

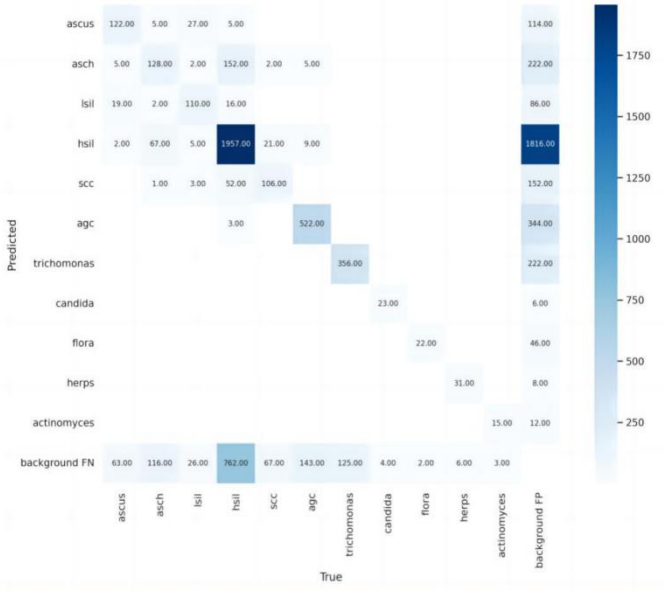

Figure 1 Yolov7x baseline model confusion matrix3.2

3.2 Feature vector clustering

In order to solve the HSIL class sample detection problem, we extract HSIL class detection samples from the baseline model prediction results to construct the clustering dataset under the conditions of confidence threshold 0.1 and IOU threshold 0.65. Based on the detection frame and label frame IOU values to determine the sample labels, the generated clustered dataset has 2080 True HSIL (i.e., HSIL-like TP detection samples) samples and 4537 False HSIL (i.e., HISL-like FP detection samples) samples.

Based on the medical characteristics of HSIL cells, combined with the YOLOv7x model characteristics, we decided to focus on the nuclear-to-plasmic ratio of HSIL class and the morphological characteristics of cells. For HSIL class test samples, we chose to calculate the nucleoplasmic ratio and extract texture features to design more representative feature vectors.

3.2.1 Nucleoplasmic Ratio Features

Nucleoplasmic ratio is the ratio of nucleus to cytoplasm area. In order to calculate accurately, the HSIL detection samples are firstly separated by colour channels, and the background interference is excluded by filtering and colour space conversion, and the threshold is set to establish a mask. Then the improved thresholding method was used to obtain the nucleus and cytoplasm thresholds after grey-scale map conversion, denoising, histogram calculation and smoothing, to complete the fine segmentation, Refer to Figure 2 for the effect diagram.

The calculated nucleoplasmic ratio index shows that the mean value of nucleoplasmic ratio of True HSIL samples is 054, and the variance is 0.72; the mean value of False HSIL is 0.73, and the

variance is 1.32. The difference between the mean value and the variance of nucleoplasmic ratio of the two types of samples is obvious, so we believe that the nucleoplasmic ratio can be used as a valid feature indicator.

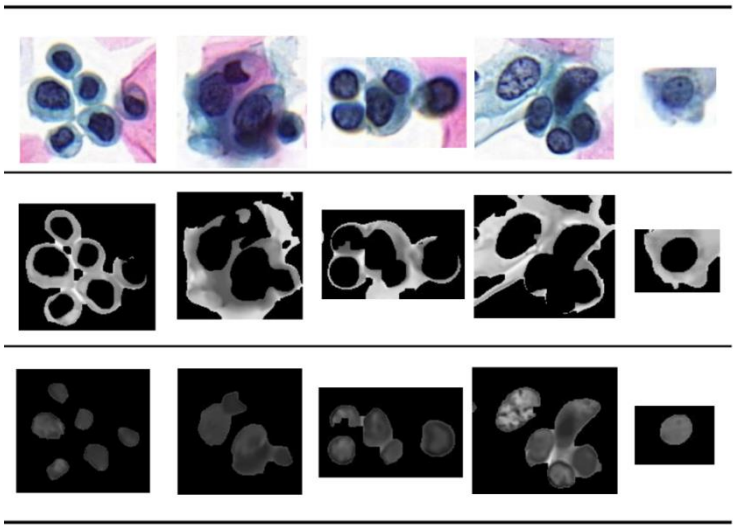

Figure 2 The effect diagram of accurate segmentation (top, middle and bottom are HSIL detection samples, cytoplasmic mask and cytosolic mask respectively)

3.2.2 grey-level co-occurrence matrix features

The greyscale covariance matrix (GLCM) is used to extract cellular texture features, and the matrix is constructed by calculating the probability of different pixel pairs and generating feature attributes such as contrast, dissimilarity, homogeneity, energy, correlation, and so on, based on this. After several sets of parameter comparison experiments, we found that calculating the above five features based on two distances of d=[20,30] and two angles of theta=[0,90] achieved the best results. So we generate a 20-dimensional texture feature vector for each HSIL detection sample, and combine it with the nucleoplasmic ratio feature to form a 21-dimensional feature vector.

3.2.3 K-means Clustering

After obtaining the 21-dimensional feature vectors, we use the K-means clustering algorithm for clustering. The reason for using the clustering algorithm with unsupervised learning is that the post-processing operation of the prediction results is essentially a secondary learning process away from the labels, and the unsupervised learning satisfies this characteristic. The clustering dataset we constructed - the set of feature vectors of all HSIL detection samples - contains low noise content due to the kernel-to-quality ratio metrics where we eliminated outliers and the grey scale covariance matrix approach where the extracted feature vectors are computed based on the grey scale information of the images and a fixed formula. Thus, the accuracy of the K-means algorithm is maximised.

Since the K-means algorithm can only divide the data into two classes, but can not determine which class is more suitable for the HSIL class Ground Truth, so it is necessary to use the labels of the HSIL class: first, the samples corresponding to the labels will be extracted according to the above method, and the same 21-dimensional feature vectors will be generated, and then these feature vectors are subjected to K-means clustering, and the representative feature vectors are taken out -Cluster centres of each cluster. The K-means algorithm needs to specify the total number of clusters , which depends on the data distribution of the HSIL labelled samples, so the HSIL labelled

dataset is downscaled to two dimensions and visualised by Principal Component Analysis (PCA) (e.g., Fig. 3).

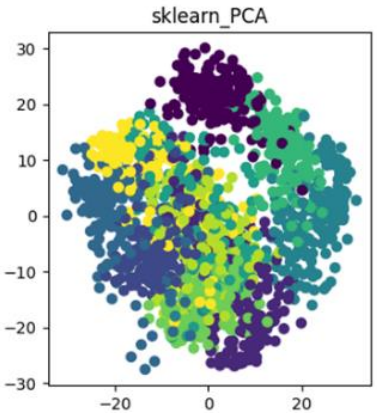

Fig. 3 PCA dimensionality reduction (2D) results for HSIL tag feature vector set

Observing the data distribution, we set the number of clustering centres to 10 and calculate the sum of the distances between the two clustering centres of HSIL detection samples and the 10 clustering centres of HSIL labeled samples, respectively, and the set of clusters of the detection samples corresponding to the smallest value can be considered to be more compatible with the HSIL class Ground Truth, i.e., the True HSIL class.

3.3 Filtering the Subnet

We encapsulate the work of Section 3.2 into a solo sub-module, naming it the Cluster-HSIL filtering sub-network. The structure of this network is shown in Figure 4:

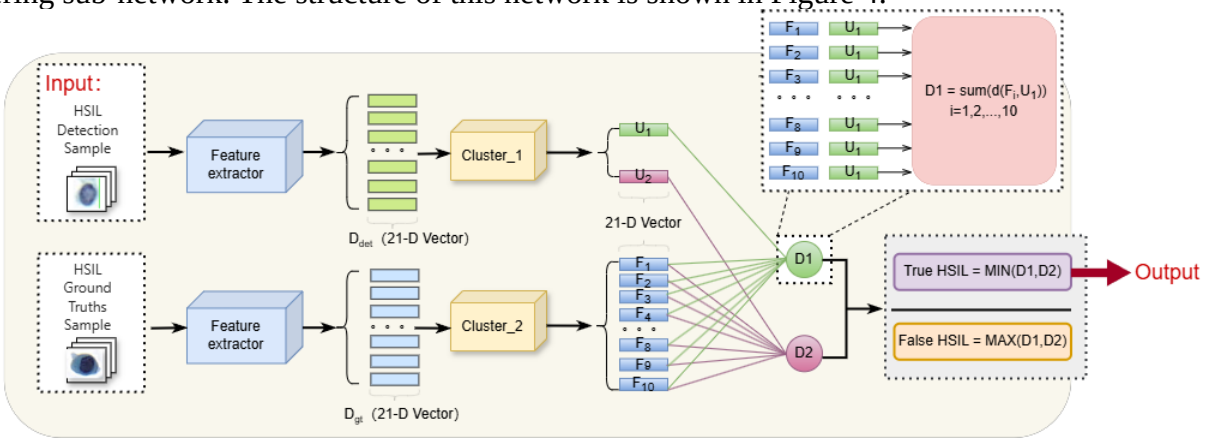

Fig. 4 Structure of the filtering subnetwork Cluster-HSIL

The input of the lower branch is the HSIL class label samples obtained by cropping from the test dataset, and the input of the encapsulated feature extractor (same as above) is the HSIL class label samples obtained by cropping from the test dataset. the cluster centres {U1,U2} of the two classes; the input of the lower branch is the HSIL class label samples obtained by cropping from the test dataset, which are fed into the encapsulated feature extractor (same as above) to obtain the 21-dimensional set of feature vectors, and the feature vector set is clustered into 10 classes using K-means clustering apparatus (i.e., Cluster_2 in Fig.), and the cluster centre of each class is returned as the representative feature vectors. Then the two branches intersect, the two cluster centres {U1,U2} of the clustered dataset and the 10 representative feature vectors are calculated and summed to obtain the sum of distances [D1,D2], [D1,D2] and {U1,U2} correspond to each other,

and the set of clustered samples corresponding to min([D1,D2]) serves as the True HSIL class of the prediction result, i.e., the filtered HSIL detection samples, and the remaining classes as the predicted False HSIL classes. Finally the detection frame output filter corresponding to the True HSIL class is taken as the final HSIL class detection frame output of the model. After Cluster-HSIL, the clustering results are shown in Table 1:

Table 1 Clustering results based on filter subnetwork

	Prediction			
Label		True HSIL	False HSIL	Total
	True HSIL	1942(TP)	138(FN)	2080
	False HSIL	458(FP)	4079(TN)	4537
	Total	2400	4217	6617

Before the Cluster-HSIL operation, there were 2080 TP samples and 4537 FP samples in the HSIL class. After processing, the number of FP samples decreased to 679, and the number of TP samples was 1906. From the classification indexes, the clustering post-processing operation significantly improved the accuracy and precision rate, but the recall rate slightly decreased (see Table 2 for the comparison of specific indexes). This indicates that this operation can effectively improve the HSIL class detection precision rate P, but has a certain impact on the recall rate R, which needs to be comprehensively evaluated.

Table 2 Comparison of classification evaluation indexes before and after Cluster-HSIL processing operation

	Accuracy	Precision	Recall
Before	0.3143	0.3143	1
After	0.8711	0.7373	0.9163

3.4 The two-stage model YOLOv7x-CH

Although Cluster-HSIL improves the HSIL class detection accuracy, it needs further validation. To this end, Cluster-HSIL is cascaded to the output of the YOLOv7x baseline model to construct the two-stage model YOLOv7x-CH (refer to Figure 5 for its structure). We first classify the baseline model model detection samples, except for the HSIL class the rest of the categories of detection samples do not go through Cluster-HSIL, but are predicted by the baseline model and directly output the detection frame. the HSIL class samples are filtered by Cluster-HSIL, and the HSIL samples that pass the filtering screening are True HSIL, which means that the YOLOv7x-CH model's final HSIL class detection frame output, and the model's evaluation metrics P, R and AP values are calculated (refer to Table 3).

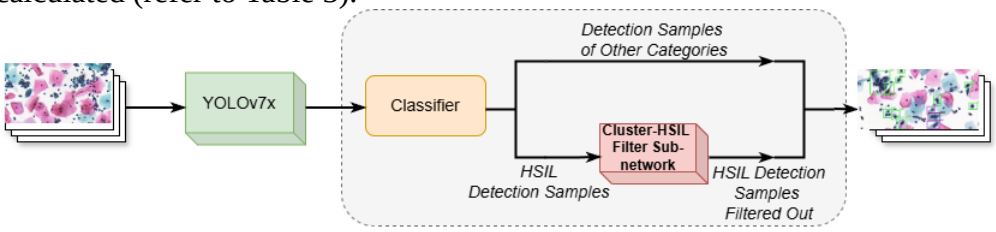

Figure 5 YOLOv7x-CH structural unfolding diagram

From the values of evaluation indexes, comparing with the baseline model YOLOv7x, the P-value of the second-stage model YOLOv7x-CH for the HSIL class was significantly improved from 0.560 to 0.813; the R-value was decreased from 0.604 to 0.526; however, from the value of AP

of the integrated index, the second-stage model YOLOv7x-CH had a more substantial improvement, the AP@.5value increased from 0.556 to 0.591, an increase of 3.5 percentage points, AP@.5:.95value increased from 0.275 to 0.330, an increase of 5.5 percentage points. With the values of the other categories of evaluation metrics remaining unchanged, mAP@.5value up from 0.612 to 0.616, an improvement of 0.4 percentage points.

Table 3 Comparison of the values of the baseline model and Cluster-HSIL on the evaluation indicators of the HSIL category

	P	R	AP@.5	AP@.5:.95	mAP@.5
YOLOv7x	0.560	0.604	0.556	0.275	0.612
YOLOv7x-CH	0.813	0.526	0.591	0.330	0.616

Taken together, the two-stage model YOLOv7x-CH successfully addressed the issue of low detection accuracy for HSIL class samples. By effectively filtering out FP (false positive) samples using the Cluster-HSIL filtering sub-network, it substantially improved the detection accuracy for HSIL class samples. Additionally, the model's AP (average precision) value for HSIL class samples and the overall mAP (mean average precision) value were significantly enhanced.

4. Conclusion

Focusing on the field of cervical pathology cell detection, this study combines the YOLOv7x target detection model with the K-means clustering algorithm to construct and optimize the two-stage detection model YOLOv7x-CH for the difficulties in HSIL-like cervical pathology cell detection, and achieves a series of research results of significant value.

In terms of model construction, through in-depth analysis of the performance of the YOLOv7x baseline model in cervical pathology cell detection, it was found that there were a large number of FP samples in the detection of HSIL-like samples, which led to a low detection accuracy P value and average accuracy AP value. To solve this problem, a two-stage model, YOLOv7x-CH, was innovatively designed in this study. Specifically, the baseline model was first trained based on YOLOv7x, and 21-dimensional feature vectors were extracted based on the grey-scale covariance matrix and fine segmentation of the HSIL-like detection samples outputted from YOLOv7x, which contained 1-dimensional nucleoplasmic ratio features and 20-dimensional grey-scale covariance matrix texture features, and then these feature vectors were extracted based on the k-means algorithm. We use the K-means algorithm to cluster these feature vectors to achieve the screening of HSIL class detection samples. As a result, the detection accuracy P-value is improved from 0.56 to 0.813, and the average accuracy AP-value is improved from 0.556 to 0.591. These improvements ensure the overall performance of the model while effectively screening out a large number of FP samples, and provide more reliable technical support for the early diagnosis of cervical cancer.

In future work, we hope to further investigate the difficult sample problems in other categories in the dataset, such as the large similarity between the targets of the ASC-H and HSIL categories, and the large number of misdetected samples in the AGC category, etc. We also hope to explore other difficult sample problems in the dataset, such as the large similarity between the targets of the ASC-H and HSIL categories. Meanwhile, we also consider exploring other feature extraction methods and considering other effective morphological features other than nucleoplasmic ratio to improve the clustering effect of the filtered sub-network and to improve the R-value of the recall of the HSIL class.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel R L, Torre L A, Jemal A. Global cancer statistics 2018: GLOBOCAN

estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: a cancer journal for clinicians, 2018, 68(6): 394-424.
[2] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv:2207.02696, 2022.
[3] Isa N A M. Automated edge detection technique for Pap smear images using moving K-means clustering and modified seed based region growing algorithm[J]. International Journal of The Computer, the Internet and Management, 2005, 13(3): 45-59.
[4] Bergmeir C, Silvente M G, Ben ́ez J M. Segmentation of cervical cell nuclei in high-resolution microscopic images: a new algorithm and a web-based software framework[J]. Computer methods and programs in biomedicine, 2012, 107(3): 497-512.
[5] Arya M, Mittal N, Singh G. Cervical cancer detection using segmentation on pap smear images[C]. Proceedings of the International Conference on Informatics and Analytics. 2016: 1-5.
[6] Sajeena T A, Jereesh A S. Automated cervical cancer detection through RGVF segmentation and SVM classification[C]. 2015 International Conference on Computing and Network Communications (CoCoNet). IEEE, 2015: 663-669.
[7] Partio M, Cramariuc B, Gabbouj M, Visa A. Rock texture retrieval using gray level co-occurrence matrix[C]. Proc. of 5th Nordic Signal Processing Symposium. 2002, 75.
[8] Jia A D, Li B Z, Zhang C C. Detection of cervical cancer cells based on strong feature CNN-SVM network[J]. Neurocomputing, 2020, 411: 112-127.
[9] Sukumar P, Gnanamurthy R K. Computer aided detection of cervical cancer using pap smear images based on hybrid classifier[J]. International Journal of Applied Engineering Research, Research India Publications, 2015, 10(8): 21021-21032.
[10] Jain A K, Dubes R C. Algorithms for clustering data[J]. Technometrics, 1988, 32(2):227-229.
[11] Dubey A K, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset[J]. International journal of computer assisted radiology and surgery, 2016, 11: 2033-2047.
[12] Bae J H, Kim M, Lim J S, Geem Z W. Feature selection for colon cancer detection using k-means clustering and modified harmony search algorithm[J]. Mathematics, 2021, 9(5): 570.
[13] Xiang Y, Sun W, Pan C, Yan M, Yin Z, Liang Y. A novel automation-assisted cervical cancer reading method based on convolutional neural network[J]. Biocybernetics and Biomedical Engineering, 2020, 40(2): 611-623.