Scientific
Research
Publishing

Detection and Classification of Lung Cancer Cells Using Swin Transformer

Yuru Chen[1,2], Jing Feng[1,2]*, Juan Liu[1,2], Baochuan Pang[2], Dehua Cao[2], Cheng Li[2]

[1]Institute of Artificial Intelligence, School of Computer Science, Wuhan University, Wuhan, China
[2]Landing Artificial Intelligence Center for Pathological Diagnosis, Wuhan University, Wuhan, China
Email: *gfeng@whu.edu.cn

Abstract

Lung cancer is one of the greatest threats to human health. It is a very effective way to detect lung cancer by pathological pictures of lung cancer cells. Therefore, improving the accuracy and stability of diagnosis is very important. In this study, we develop an automatic detection scheme for lung cancer cells based on convolutional neural networks and Swin Transformer. Microscopic images of patients' lung cells are first segmented using a Mask R-CNN-based network, resulting in a separate image for each cell. Part of the background information is preserved by Gaussian blurring of surrounding cells, while the target cells are highlighted. The classification model based on Swin Transformer not only reduces the computation but also achieves better results than the classical CNN model, ResNet50. The final results show that the accuracy of the method proposed in this paper reaches 96.16%. Therefore, this method is helpful for the detection and classification of lung cancer cells.

Keywords

Lung Cancer, Classification, Swin Transformer

1. Introduction

Lung cancer is one of the leading culprits that threaten human health. According to the 2022 World Health Organization Cancer Report, 609,360 people will die from cancer in the United States in 2022, which is equivalent to nearly 1700 deaths per day, including approximately 350 deaths per day from lung cancer, which is the leading cause of death from all cancers [1]. Lung cancer has an extremely high mortality rate, and early diagnosis and treatment can dramatically improve patients' chances of survival [2]. Previous methods of lung cancer di-

agnosis include computed tomography (CT), chest X-ray, and cytopathological identification. Screening and detecting lung cancer cells are crucial in cancer prevention and control efforts [3]. Lung cancer diagnosis and ancillary tests rely on cytology and small biopsy specimens obtained by minimally invasive means [4]. Specimens of lung cancer cells are usually obtained from patients' sputum exfoliated cells, alveolar lavage fluid, bronchial secretions, or pleural effusions. Compared with other screening methods, this method is convenient, quick, and basically non-invasive, which is very suitable for the initial screening.

Traditionally, lung cancer cytopathological images are used by pathologists or physicians to analyze cell morphology, number, differentiation, and other characteristics to reach a diagnosis. In recent years, as the number of patients with the disease has increased, the large number of lung cancer patients has brought thousands of data to be analyzed, and processing these data requires a large number of professionals. With the shortage of pathologists in some areas, it is unreasonable to use manual review data to cause a waste of human resources. The long-term repetitive and boring work also increases the possibility of misjudgment by professionals. Therefore, the research on cytopathological image-assisted diagnosis systems for lung cancer is of great practical significance. Combining advanced computer technology and the diagnostic experience of cytology experts can, to a certain extent, solve the current medical troubles of cancer cell diagnosis and reduce the workload and artificial influence of pathologists. This work can largely improve the efficiency of early lung cancer screening and reduce the mortality rate of lung cancer patients [5].

In the last decade, with the development of computer hardware and deep learning algorithms, artificial intelligence has been used to process the stream of data generated throughout the clinical pathway [6]. Computer-aided medical analysis techniques have also been rapidly developed with advances in image analysis algorithms and the rise of big data algorithms [7]. Using machine learning algorithms to identify and detect cancer has been shown to be feasible [8] [9] [10]. Today, many cytopathological recognition methods have been proposed as the techniques for image classification are becoming mature. However, since cells in different organs and tissues have different characteristics, the guidelines for physicians to determine whether a cell is diseased or not may change accordingly. There is no universal cytopathological image recognition method. Current methods for lung cancer cell detection suffer from low prediction accuracy, high resource consumption, and poor real-time performance. In this paper, we propose a transformer-based lung cancer cell detection network, which solves the above problems to some extent.

2. Related Work

The identification and detection of lung cancer cells consist of two main steps: cell nucleus segmentation and cell image classification. Segmentation of lung cancer cells involves segmenting one or more lung cells in an image to facilitate

pathologists to clearly observe their morphology, color, and other features. The cell nucleus segmentation is to prepare for the classification afterward.

There are many traditional image segmentation methods, which are widely used in cell nucleus segmentation. Threshold segmentation [11] [12] is the simplest method to distinguish foreground objects from the background. The basic idea of clustering segmentation [13] [14] is to calculate the similarity between each pixel point and group the pixels with high similarity into one class, so as to segment the image. There are some other traditional methods for segmenting cell nuclei, such as the watershed algorithm [15] [16] and the active contour method [17]. All these traditional methods have obvious advantages and disadvantages, and individual methods are only applicable in some specific scene conditions. And they often have limitations for the complex environments that occur in reality. Therefore, a combination of multiple methods is often used in practice, which also brings new problems such as great computational effort and complex computational principles.

The success of deep learning has brought new life to medical image segmentation. In 2019, Yiming Liu *et al.* [18] used a combination of coarse and fine segmentation methods, and first trained Mask-RCNN to obtain coarse segmentation results. The local fully connected conditional random field is used in fine segmentation, and finally the two are fused. In 2020, Cai *et al.* [19] proposed a Dense U-Net structure based on the U-Net model, which uses a dense cascade form to segment the body skin cell images. In 2021, Liu Z [20] proposed an improved backpropagation (BP) neural network model for color fundus image segmentation. It can be seen that deep learning techniques are widely used in medical images and can be used to solve the segmentation problems of the skin, MRI, retinal images, and cell images to achieve automatic segmentation of targets.

The traditional method of lung cancer cell classification requires manual extraction of cell features. The advent of deep learning has simplified this step [21]. The global linking and weight sharing features of convolutional neural networks make them well suited for processing images, which has led to the derivation of many classical CNN models. In 2014, Simonyan and Zisserman proposed the VGG model [22] with a deeper network structure. Compared to other neural networks, it uses a smaller size convolutional kernel, which increases the nonlinear representation of the network while reducing the parameters. In 2015, ResNet was introduced to solve the gradient vanishing problem, which is common in neural networks [23]. It introduced residual blocks, and the network performed Identity Mapping through shortcuts with good results. In 2016, Huang G, *et al.* [24] effectively alleviated the gradient disappearance problem by reusing the feature map in the network while enhancing the transfer of features in the network. In 2017, Teramoto *et al.* [25] developed an automatic classification scheme for lung cancer based on microscopic images using deep convolutional neural networks (DCNN). And their classification accuracy was evaluated using triple cross-validation. In the obtained results, about 71% of the images

were correctly classified. In 2020, Daniel Gonzalez *et al.* [26] used three different convolutional neural networks for small cell lung carcinoma (SCLC), large cell neuroendocrine carcinoma (LCNEC), and mixed/unclassifiable three categories for classification and diagnosis, which eventually achieved good results on a limited dataset.

In 2017, the transformer framework proposed by Google [27] attracted a lot of attention. It not only became a mainstream model in the field of natural language processing, but also started to expand to the field of computer vision. In 2020, Google proposed Vision Transformer (ViT) [28]. The direct use of transformers for image classification in the article broke the reliance on CNN—the method used in most image processing work—in the CV field. In 2021, Liu Z *et al.* proposed Swin Transformer [29], which surpasses backbone networks such as EfficientNet in terms of performance. It introduced a sliding window mechanism and a hierarchical structure, making the Swin Transformer the new Backbone of machine vision. It reached the SOTA level in a variety of machine vision tasks such as image classification, target detection, and semantic segmentation. ViT has also been used in medical image processing. For example, the staff of [30] [31] used transformers to distinguish COVID-19 from other types of pneumonia by computed tomography (CT) or X-ray images, meeting the urgent need for fast and effective treatment of COVID-19 patients.

As can be seen from the references, the current lung cancer cytopathology image detection technology is not mature enough, and the accuracy of detection is low. CNN can only extract local features through a convolution kernel, while the ViT model can learn the features of the whole image through an attention mechanism, which can better analyze the image. Therefore, the work in this paper is very meaningful in the field of early diagnosis of lung cancer.

3. Materials and Methods

3.1. Image Data Set

A total of 347 images of lung washout cells were collected from 10 patients by exfoliation or interventional cytology under bronchoscopic guidance. The pixel size of the images was all 512 × 512. Each image contained one or more HE-stained lung cells. The nuclei of the lung cells were dark blue in color. After labeling and counting by professional pathologists, there were finally 2473 lung cells, which contained 143 cancerous cells, 724 normal cells and 1606 noisy blocks containing impurities and incomplete cells.

3.2. Cell Segmentation and Data Enhancement

In this study, NucleAIzer [32], a deep learning framework for cell nucleus segmentation, was used to segment each lung cancer cell image, and 2473 images of individual cell nuclei were obtained. The paper provides a deep learning framework for cell nucleus segmentation called nucleAIzer, in which the paper first uses mask-RCNN for initial training. Then the training images are clustered into

134 classes based on image styles. Next, new image/mask pairs are generated for each style of images in the training set using cycleGAN. The mask-RCNN model is updated with these enhanced data. Finally, u-Net is used to calibrate the edges. The framework is able to segment the images with different styles and types of cell nuclei more accurately. The final segmentation of the test set data in the 2018 DSB competition is higher than the first place. NucleAIzer also shows very good segmentation results on our dataset.

Like the training of CNNs, the training of transformers requires a sufficient amount of data. To maximize the use of each data, we treat each individual segmented cell as separate data in the experiment. In this case, the amount of data increased from 347 to 2473 sheets. Considering the surrounding environment, when a doctor or pathologist determines whether a cell is cancerous or not, the information is added to the reference. Two conditions were combined—to retain the environmental information of the target cells and to highlight the target to be analyzed to distinguish other cells in the same picture. We adopted a compromise approach: Gaussian blurring of the region outside the target. The results of several experiments show that the best results are achieved when the size of the Gaussian kernel is set to (65, 65), the cells before and after splitting and Gaussian blurring are shown in Figure 1.

It is noted that the number of the three categories in the data is unevenly distributed. Also, in order to increase the training data and avoid overfitting the model, we increased the number of cancerous and normal cells in the dataset. The different image orientations of the microscope during data acquisition can lead to differences in the position and angle of individual cells in the plane. Therefore, the experiments are performed by rotating, inverting and adding noise operations for data enhancement while enhancing the robustness of the model. In the rotation operation, the images are rotated clockwise by a random number of degrees in the range of 0 to 180. For the add noise operation, two types of noise, pretzel noise and Gaussian noise, are added to the images. The final enhanced data totaled 3106, which contained 500 cancerous cells, 1000 normal cells and 1606 noisy data.

3.3. Swin Transformer Structure

The transformer structure used for lung cancer cytopathology image classification is shown in Figure 2(a). Each stage in the figure consists of Patch Merging and Swin Transformer Block. The input H * W three-channel image first passes through a Patch Partition module. The image is divided into multiple patches of size 4×4, and the dimensionality becomes $\frac{H}{4} * \frac{W}{4} * 48$. Then it passes through a Linear Embedding layer, which can embed features into any dimension. At this point, the dimension is recorded as C. After that, it passes through the core module Swin Transformer Block, and the number of tokens remains the same. All these are the work of stage 1.
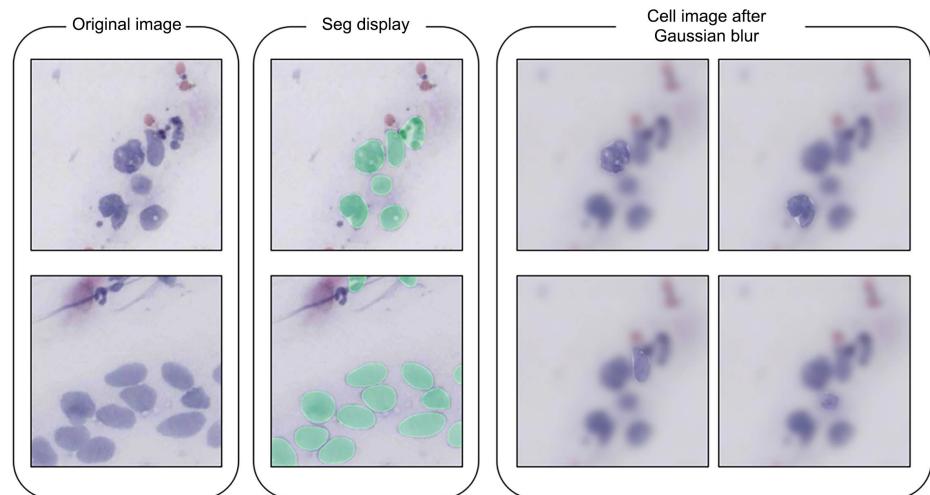
Figure 1. Sample images of lung cancer cells before segmentation, after segmentation and after Gaussian blur.

In stage 2, a Patch Merging module is first employed. The purpose is to fuse patches in the 2 × 2 region to produce a hierarchical feature representation. The feature dimension of each new patch is thus changed to 4C. The dimensionality is then reduced to 2C in order to reduce the subsequent computation. The dimensionality is then kept constant after the Swin Transformer Block. The subsequent stage3-stage4 repeats the previous operations, continuously fusing the adjacent patches. The size of the feature map output from the stage4 is $\frac{H}{32} * \frac{W}{32} * 8C$.

The structure of one of the core modules, the Swin Transformer Block, is shown in **Figure 2(b)**. Compared with the traditional transformer, Swin Transformer uses W-MSA (Window MSA) and SW-MSA (Shifted Window MSA) instead of MSA (Multi-head self-attention module). Traditional transformers compute attention based on the global picture, so the computational complexity is very high. The Swin Transformer reduces the computation by limiting the attention computation to each window. In order not to lose global information, Shifted Window is added to better interact with other windows. This makes the hierarchical feature and linear time complexity possible.

4. Experiment

4.1. Experimental Environment

The experiments in this paper were conducted on Ubuntu 18.04.5 LTS operating system. The pytorch = 1.10.1 deep learning framework is used. Python language version is 3.6. The GPU used in the experiments is NVIDIA GeForce RTX 2080Ti. In training, Adam is used as the optimizer, and the batch size is set to 12. The initial learning rate is set to 0.0001. Every 30 epochs are trained. The initial learning rate is set to 0.0001. Every 30 epochs are trained, the learning rate becomes 10% of the original. The ratio of the training set to test set was 7:3, and the epoch is 100.
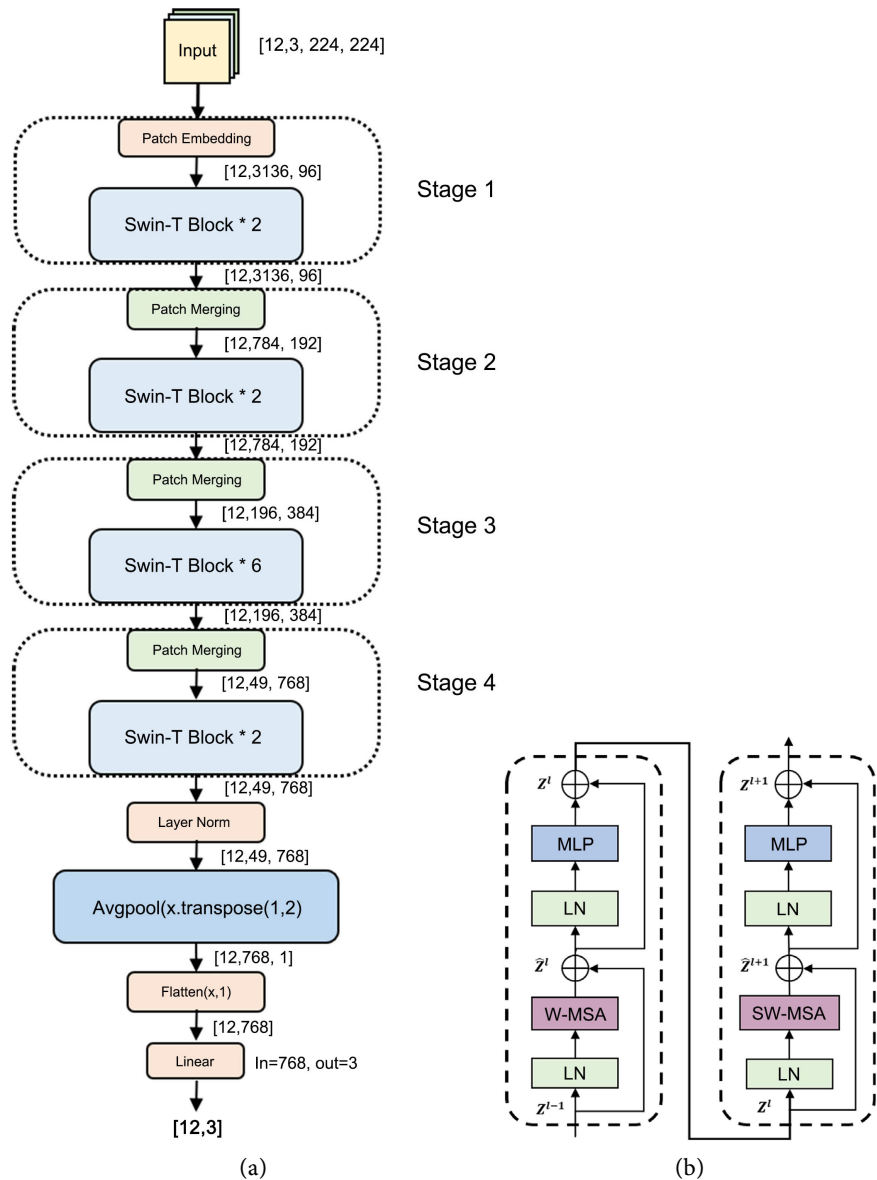
Figure 2. (a) Architecture of the Swin Transformer used for lung cancer cell classification; (b) Two successive Swin Transformer Blocks.

4.2. Evaluation Method

Accuracy, precision, recall and specificity were used in the experiments to evaluate the performance of the lung cancer cell classification model. The number of positive samples with accurate prediction was recorded as true positive (TP). The number of negative samples with accurate prediction was recorded as true negative (TN). The number of negative samples predicted to be positive was recorded as false positive (FP). The number of positive samples predicted to be negative is a false negative (FN). Precision denotes the probability of the number of correctly predicted positive samples to the number of all samples predicted to be positive. Recall denotes the ratio of the number of correctly predicted positive samples to the number of actual positive samples. Accuracy is the ratio of the

number of correctly predicted to the total number of samples. Specificity represents the probability of being correctly judged as negative among the actual negative samples. The calculation formulas of the above four evaluation indicators are as follows.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

4.3. Experimental Results and Analysis

The performance of the Swin Transformer model on the test set is shown in **Table 1**. The average precision, recall and specificity of lung cancer cell detection on the test set were calculated to be 95.20%, 92.60% and 98.17%, respectively. The accuracy of model classification was 96.14%. It proved that the detection of lung cancer cells using Swin Transformer is feasible. The confusion matrix of the test set is shown in **Figure 3**. In the confusion matrix, we can see that all the images in the noise category are classified correctly, and only a small number of errors are generated in the abnormal and normal categories.

Table 1. Precision, recall and specificity of lung cancer cell classification.

	Precision	Recall	Specificity
Abnormal	0.944	0.801	0.991
Noise	1.0	1.0	1.0
Normal	0.912	0.977	0.954
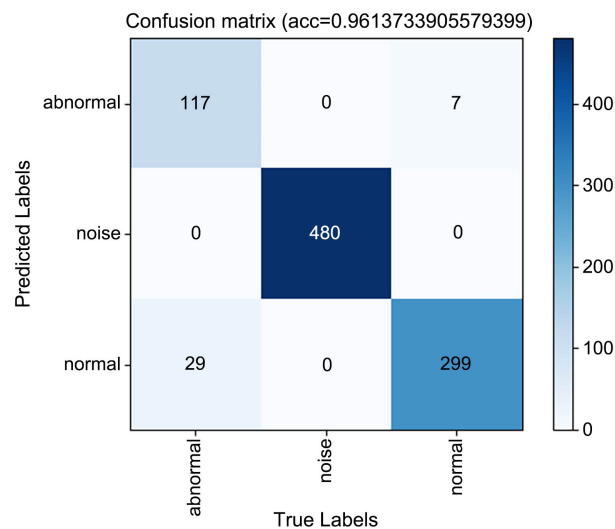
Figure 3. Confusion matrix of classification result.

Table 2. Accuracy, average precision, average recall and average specificity of different models on different datasets.

Models (dataset)	Acc	Avg.P	Avg.R	Avg.S (%)
Resnet50 (lung cells)	94.21	92.14	92.31	97.18
Resnet50 + FPN (lung cells)	93.56	91.70	92.40	96.63
Swin Transformer (lung cells)	**96.14**	**95.20**	**92.60**	**98.17**
Resnet50 (Herlev)	66.30	70.33	68.39	94.24
Resnet50 + FPN (Herlev)	46.74	54.64	49.30	90.87
Swin Transformer (Herlev)	**70.29**	**75.06**	**73.57**	**94.79**
Resnet50 (SIPaKMeD)	98.02	98.00	98.02	99.52
Resnet50+FPN (SIPaKMeD)	90.37	90.42	90.46	97.60
Swin Transformer (SIPaKMeD)	96.95	97.00	97.02	99.24

4.4. Extended Experiments

In order to observe the performance of the Swin Transformer model, the resNet50 and resNet50 + FPN models, which perform very well in the field of image classification, are selected for comparison with the current model, where FPN stands for Feature Pyramid Network. This is a feature fusion technique. The basic idea of FPN is to improve the effectiveness of the network by fusing the features of higher and lower layers together, *i.e.*, multi-scale feature fusion, so as to fully utilize the features of each stage of the network.

Then experiments were conducted using three different models in two publicly available cervical cell datasets—Herlev and SIPaKMeD—as a way to demonstrate the generalization performance of the model. The SIPaKMeD dataset is a five-category labeled cervical cell dataset with a total of 4049 cervical cells. The overall precision, recall, and specificity of the model were calculated by calculating the mean of the different categories. The experimental setup and dataset division was the same as before.

Table 2 shows the results of the different data sets on the different models. The results show that Swin Transformer performs slightly worse than ResNet50 on the SIPaKMeD dataset, except on the other two datasets, where the results are significantly better than all other classification models. The accuracy of the lung cancer cell dataset, which is our main focus, reached 96.14, which is nearly two percentage points higher than the resNet50 model. This demonstrates the effectiveness of the Swin Transformer for lung cancer cell image classification and that it can perform well on other cell image datasets as well.

5. Conclusion

In this paper, a Swin Transformer-based lung cancer cell classification model is proposed. The experiments firstly segmented the lung cancer cell images to separate each cell, then defocused the background of the target cells using Gaussian blur, and finally put them into the Swin Transformer model for classification.

The experimental results showed that the accuracy of classification reached 96.16%. Therefore, it can be proved that using Swing Transformer to detect lung cancer cells is effective.

Acknowledgements

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Siegel, R.L., Miller, K.D., Fuchs, H.E., *et al.* (2022) Cancer Statistics, 2022. *CA: A Cancer Journal for Clinicians*, **72**, 7-33. https://doi.org/10.3322/caac.21708

[2] Asuntha, A. and Srinivasan, A. (2020) Deep Learning for Lung Cancer Detection and Classification. *Multimedia Tools and Applications*, **79**, 7731-7762. https://doi.org/10.1007/s11042-019-08394-3

[3] da Silva, G.L.F., de Carvalho Filho, A.O., Silva, A.C., de Paiva, A.C. and Gattass, M. (2016) Taxonomic Indexes for Differentiating Malignancy of Lung Nodules on CT Images. *Research on Biomedical Engineering*, **32**, 263-272. https://doi.org/10.1590/2446-4740.04615

[4] VanderLaan, P.A. (2018) Updates in Lung Cancer Cytopathology. *Surgical Pathology Clinics*, **11**, 515-522. https://doi.org/10.1016/j.path.2018.04.004

[5] Fernandes, K., Chicco, D., Cardoso, J.S., *et al.* (2018) Supervised Deep Learning Embeddings for the Prediction of Cervical Cancer Diagnosis. *PeerJ Computer Science*, **4**, e154. https://doi.org/10.7717/peerj-cs.154

[6] Kann, B.H., Hosny, A. and Aerts, H.J.W.L. (2021) Artificial Intelligence for Clinical Oncology. *Cancer Cell*, **39**, 916-927. https://doi.org/10.1016/j.ccell.2021.04.002

[7] Han, F.J., Yu, L. and Jiang, Y. (2020) Computer-Aided Diagnosis System of Lung Carcinoma Using Convolutional Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, 14-19 June 2020, 690-691. https://doi.org/10.1109/CVPRW50498.2020.00353

[8] Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V. and Madabhushi, A. (2019) Artificial Intelligence in Digital Pathology—New Tools for Diagnosis and Precision Oncology. *Nat Rev Clin Oncol*, **16**, 703-715. https://doi.org/10.1038/s41571-019-0252-y

[9] Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S. and Fuchs, T.J. (2019) Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images. *Nature Medicine*, **25**, 1301-1309. https://doi.org/10.1038/s41591-019-0508-1

[10] Fourcade, A. and Khonsari, R.H. (2019) Deep Learning in Medical Image Analysis:

A Third Eye for Doctors. *Journal of Stomatology, Oral and Maxillofacial Surgery*, **120**, 279-288. https://doi.org/10.1016/j.jormas.2019.06.002

[11] Zhao, J., Liang, L.K., He, Y.J., *et al.* (2019) Cervical Nucleus Segmentation Method in Complex Background. *Journal of Harbin University of Science and Technology*, **24**, 25-31.

[12] Sun, H.F., Yang, J.H., Fan, R.B., *et al.* (2020) Stepwise Local Stitching Ultrasound Image Algorithms Based on Adaptive Iterative Threshold Harris Corner Features. *Medicine*, **99**, e22189. https://doi.org/10.1097/MD.0000000000022189

[13] Feng, F., Liu, P.X., Li, L., *et al.* (2018) Research on GSA Algorithm Improved by FCM Fusion in Medical Image Segmentation. *Computer Science*, **45**, 252-254.

[14] Bai, X.Z., Sun, C.X. and Sun, C.M. (2019) Cell Segmentation Based on FOPSO Combined with Shape Information Improved Intuitionistic FCM. *IEEE Journal of Biomedical and Health Informatics*, **23**, 449-459. https://doi.org/10.1109/JBHI.2018.2803020

[15] Gamarra, M., Zurek, E., Escalante, H.J., *et al.* (2019) Split and Merge Watershed: A Two-Step Method for Cell Segmentation in Fluorescence Microscopy Images. *Biomedical Signal Processing and Control*, **53**, Article ID: 101575. https://doi.org/10.1016/j.bspc.2019.101575

[16] He, A.L., Cheng, X.B., Liao, L.C., *et al.* (2020) A Watershed Remote Sensing Image Segmentation Method Coupled with H-Minima and Mathematical Morphology. *Journal of East China University of Science and Technology* (*Natural Science Edition*), **43**, 396-400.

[17] Hsu, W.Y., Lu, C.C. and Hsu, Y.Y. (2020) Improving Segmentation Accuracy of CT Kidney Cancer Images Using Adaptive Active Contour Model. *Medicine*, **99**, e23083. https://doi.org/10.1097/MD.0000000000023083

[18] Liu, Y.M. (2019) Nucleus Segmentation of Cervical Cancer Images Based on Deep Learning and Conditional Random Fields. North Central University, Minneapolis.

[19] Cai, S., Tian, Y., Lu, I.H., *et al.* (2020) Dense-Unet: A Novel Multiphoton *in Vivo* Cellular Image Segmentation Model Based on a Convolutional Neural Network. *Quantitative Imaging in Medicine and Surgery*, **10**, 1275-1285. https://doi.org/10.21037/qims-19-1090

[20] Liu, Z. (2021) Construction and Verification of Color Fundus Image Retinal Vessels Segmentation Algorithm under BP Neural Network. *The Journal of Supercomputing*, **77**, 7171-7183. https://doi.org/10.1007/s11227-020-03551-0

[21] Martínez-Mása, J., Bueno-Crespob, A. and Martínez-España, R. (2020) Classifying Papanicolaou Cervical Smears through a Cell Merger Approach by Deep Learning Technique. *Expert Systems with Applications*, **160**, Article ID: 113707. https://doi.org/10.1016/j.eswa.2020.113707

[22] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

[23] He, K., Zhang, X., Ren, S., *et al.* (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[24] Huang, G., Liu, Z., Van Der Maaten, L., *et al.* (2017) Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 4700-4708. https://doi.org/10.1109/CVPR.2017.243

[25] Teramoto, A., Tsukamoto, T., Kiriyama, Y., *et al.* (2017) Automated Classification

of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks. *BioMed Research International*, **2017**, Article ID: 4067832. https://doi.org/10.1155/2017/4067832

[26] Gonzalez, D., Dietz, R.L. and Pantanowitz, L. (2020) Feasibility of a Deep Learning Algorithm to Distinguish Large Cell Neuroendocrine from Small Cell Lung Carcinoma in Cytology Specimens. *Cytopathology*, **31**, 426-431. https://doi.org/10.1111/cyt.12829

[27] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 5998-6008.

[28] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al.* (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

[29] Liu, Z., Lin, Y., Cao, Y., *et al.* (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

[30] Costa, G.S.S., Paiva, A.C., Junior, G.B. and Ferreira, M.M. (2021) Covid-19 Automatic Diagnosis with CT Images Using the Novel Transformer Architecture. *Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*, Virtual Event, 15-18 June 2021, 293-301.

[31] van Tulder, G., Tong, Y. and Marchiori, E. (2021) Multi-View Analysis of Unregistered Medical Images Using Cross-View Transformers. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Virtual Event, 27 September-1 October 2021, 104-113. https://doi.org/10.1007/978-3-030-87199-4_10

[32] Hollandi, R., Szkalisity, A., Toth, T., *et al.* (2020) NucleAIzer: A Parameter-Free Deep Learning Framework for Nucleus Segmentation Using Image Style Transfer. *Cell Systems*, **10**, 453-458.E6. https://doi.org/10.1016/j.cels.2020.04.003