

Classification of Lung Cancer Using Nuclear Chromatin Characteristics

Migyung Cho*

Professor, Department of Computer & Media Engineering, Tongmyong University, Busan 48520, Korea
mgcho@tu.ac.kr

Abstract

Background/Objectives: Studies on automatic diagnosis to assist pathologists have been accelerating due to artificial intelligence technology. We investigated whether two types of lung cancers, squamous cell carcinoma (Sca) and neuroendocrine tumor (NET), can be automatically classified with only the features of nuclear chromatin, similar to the method for diagnosing cancer by a pathologist.

Methods/Statistical analysis: In order to classify lung cancer with only nuclear chromatin features, it is necessary to extract texture features of the nuclear chromatin area or train a deep learning model with only the nuclear chromatin area as input. We developed a segmentation algorithm that automatically extracts nuclear areas based on the k-means algorithm to automatically generate training and testing datasets for deep learning models. And using four types of deep learning models, we performed lung cancer classification using only nuclear chromatin features.

Findings: The pathologist judges the presence and type of cancer by looking at the size and placement of the nuclei in the cell, and the texture characteristics. In this paper, a data set for a deep learning model was created by extracting only the nuclear region from a pathological image with a 40x magnification of WSI using an automatic segmentation algorithm. For four classification models, an average of 84% classification accuracy was obtained. Since only the nuclear region was used for training and testing, the deep learning model extracted and classified the size, shape, and texture features of the nuclei. Therefore, the classification of lung cancer performed in this study is similar to the method that the pathologist classifies cancer based on the features of nuclear chromatin. Since it showed 84% classification accuracy for the four types of models, it was proved that the automatic classification system using deep learning technology can be used to assist pathologists' diagnosis.

Improvements/Applications: The accuracy of a cancer diagnosis system that applies only deep learning models has limitations. To improve diagnostic accuracy, we are going to study on a model that combines a deep learning model and machine learning technology that extracts and uses biomarker features for analyzing nuclei texture patterns.

Keywords: Pathological imaging, lung cancer classification, nuclear chromatin analysis, deep learning, segmentation, WSI (Whole Slide Image)

1. Introduction

Studies on the medical image diagnosis system based on deep learning technology for various types of medical images are ongoing, and some commercialized products are emerging. During the past 10 years, automatic diagnosis system for the field of pathology for diagnosing cancer has been actively developed [1-4]. The era of diagnosing through glass slides by magnifying them with a microscope has passed. Now is the era of diagnosis by viewing the scanned image on a monitor [2-6]. A WSI (Whole Slide Image), a scanned image from its glass slide, is a huge image ranging from 2G to 70G, so a high-performance computer and parallel technology are required to process WSI. However, as recent computer technologies support such big capacity image processing, more research is being conducted on an automatic diagnosis system for pathological images.

The need for an automated diagnosis system for pathological imaging is due to difficulties in reproducing pathologists' diagnosis for disease [7-8]. There may be inconsistent diagnosis among pathologists for the same pathological image, and reproducibility cannot be guaranteed even if one person makes a diagnosis. In addition, it is a time consuming task because pathologists have to repeat the same diagnosis everyday through a microscope to diagnose a large-volume pathological image at the

nuclear level. Therefore, if an automatic cancer diagnosis system with guaranteed reproducibility can be developed to assist pathologists in diagnosis, the accuracy of diagnosis can be improved and the diagnosis time for doctors can be reduced.

As artificial intelligence technology for visual recognition surpasses human capabilities, the development of artificial intelligence-based medical imaging diagnostic systems is accelerating, and commercialized products are increasingly used in hospitals. However, the commercialized system is limited to some fields, and it is expected that the development of diagnostic systems for various types of medical images will be actively progressed. The problem to be solved in the development of an artificial intelligence-based medical diagnosis system is to secure a large amount of high-quality data, which is difficult because it requires a lot of cost.

In this study, an artificial intelligence-based experiment was conducted to classify two types of lung cancer with only nuclear chromatin, and the data set required for the experiment was automatically extracted from the pathology image by applying our segmentation algorithm. We developed a new segmentation algorithm for obtaining nuclei chromatin areas from a pathological image. The prepared data set was applied to four types of CNN models DenseNet[9], MobileNet[10], InceptionNet[11], and ResNet[12] and experiments were conducted while tuning hyper-parameters. As a result of the experiment, it was confirmed that the type of lung cancer can be classified only by the characteristics of nuclear chromatin by obtaining an average of 84% classification accuracy.

This paper is organized as follows. In Chapter 2, we will explain in detail the segmentation algorithm used to obtain the nuclei areas, the method for generating the dataset to be used for deep learning, and the experimental method for classification. And in Chapter 3, we compare and analyze the experimental results obtained by applying to the four CNN models, and conclude in Chapter 4.

2. Materials and Methods

The two types of lung cancer used for classification were squamous cell carcinoma (Sca) and neuroendocrine tumor (NET). WSIs data used in the experiment were provided excluding all personal information including personal information, from a university hospital. With the help of a pathologist, 19 WSIs containing Sca and 14 WSIs containing Net were used. And datasets necessary for training and testing were automatically generated from these WSIs through our proposed segmentation algorithm.

2.1. Overview of Our System

Figure 1 shows an overview of our system. WSI is a huge image, and a WSI we used has a capacity of 3G, so it is difficult to process it by putting it in memory at once. Therefore, images to be used in the experiment were prepared by clipping the WSI to a size of 2000 pixels horizontally and vertically with a 40 times magnification. The prepared original images are made into grayscale images through pre-processing such as color normalization and Gaussian blurring, and then our nuclear segmentation is applied. After nuclear segmentation, a post-processing is required to remove very small nuclei or false positive segmented areas.

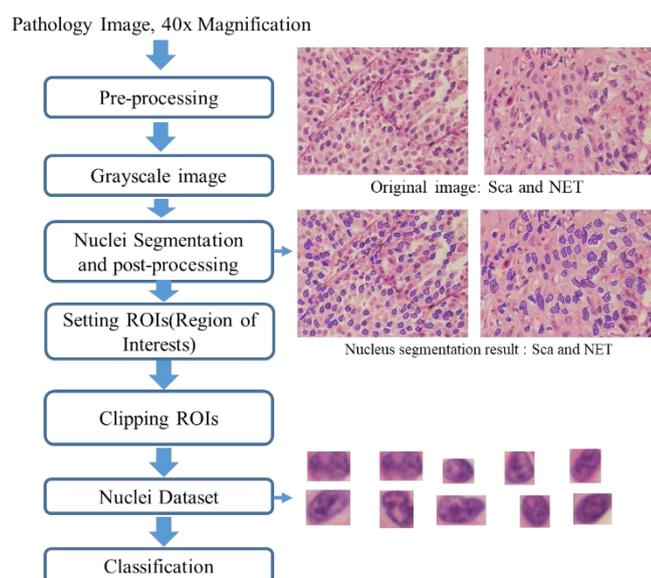


Figure 1. Overview of our system

When the post-processing was completed, the nuclei region was designated as an ROI, and all subdivided nuclear regions were saved as image files to create a data set for the deep learning model. For the generated dataset, 70% was divided into the training set and 30% into the test set. After training with four types of CNN models on the training data set, classification accuracy was measured by performing a classification with a test set.

2.2 Segmentation of Nuclei

Our segmentation algorithm to obtain the nuclear area consists of four steps. First, we converted the RGB color image to the HSV image in order to segment the nuclear region easily. Since the nuclear region has a deep magenta color compared to the background, only the nuclear area can be extracted by setting hue, saturation, and intensity in the HSV image. Second, in HSV image, K-means algorithm was performed to designate the hue boundary value for the nuclear region. Fig. 2(a) is the image with squamous cell carcinoma (Sca) we used, and the image can be divided into three areas: the nucleus, the background, and the cytoplasm. We applied the k-means algorithm for K=3 to obtain the minimum and maximum hue values for the nuclei. For saturation and intensity, we obtained the boundary values through experiments.

Third, the nuclei area was extracted by specifying the minimum and maximum of hue, saturation, and intensity values for the HSV image. Fig. 2 (b) shows the result of extracting pixels in the range between the minimum and maximum values of hue, saturation and intensity from HSV image. And Fig. 2(c) shows the boundaries of the extracted regions in shown Fig. 2(b). We used the findContours function of OpenCV library to get boundaries. Fourth, we measured the area and roundness of the nuclear area using boundary information. And we removed too small nuclear in blue circle and low roundness in yellow circle. We calculate the roundness of a segmented nuclei area by equation (1).

$$Roundness = \frac{area(P(G_i))}{area(convex\ hull(P(G_i)))} \quad (1)$$

In equation (1) a polygon $P(G_i) = (v_1, v_2, v_3 \dots v_n, v_1)$ means i-th segmented nuclei G_i and n is the total number of vertices. The convex hull ($P(G_i)$) means a convex polygon of minimum area which contains all the points of a polygon $P(G_i)$. Roundness is the ratio of the segmented $P(G_i)$ to area of its convex hull($P(G_i)$). In general, the nuclear region has a shape close to an ellipse. Therefore, in this study, only segmented polygons with a value of 0.85 or more of roundness were treated as nuclei .In Fig. 2(c), the

segmented regions of the yellow ellipse have circularity values of 0.85 or less, so it is highly likely that they are false positives, not nuclei. Therefore, as a result of the final segmentation, that is, it was excluded from the area of nuclear chromatin. The ones marked in blue in 2(d) show the finally segmented nuclei. The all segmented nuclei are numbered for identification.

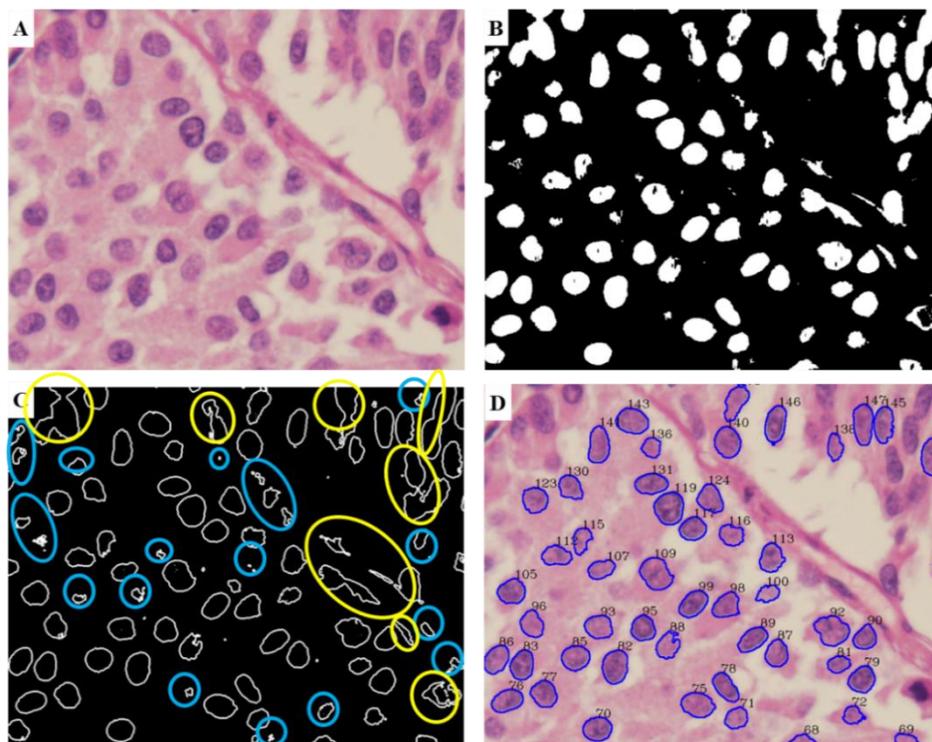


Figure 2. Segmentation Process: A. Original image, B. Extracted areas by specifying hue, saturation, and intensity values for the HSV color model, C. Boundaries of B, D. Final segmented nucleus marked in blue

2.2. Preparing dataset

A total of 3,125 images were generated for the two types of lung cancer squamous cell carcinoma (Sca) and neuroendocrine tumor (NET) from 33 WSIs containing Sca and containing Net, respectively. Figure 3 shows some of the two types of dataset obtained from two segmented nuclei images. All segmented nuclear chromatin regions marked in blue were designated as ROIs and saved as image files, respectively. The size of the ROIs was designated as the size of the smallest square containing the subdivided nuclear chromatin. Therefore, the size of the cropped images differs according to the size of nuclear chromatin. In order to use the stored image file as an input to the deep learning model, the size of the images was resized to the same size, 64x64 pixels. The number of saved image file is 3,125. The reason for saving it as an image file is to use it as an input for a deep learning model. 2,188 data, corresponding to 70% of the data set, were used for training, and the remaining 937, corresponding to 30%, were used as test data sets to measure performance.

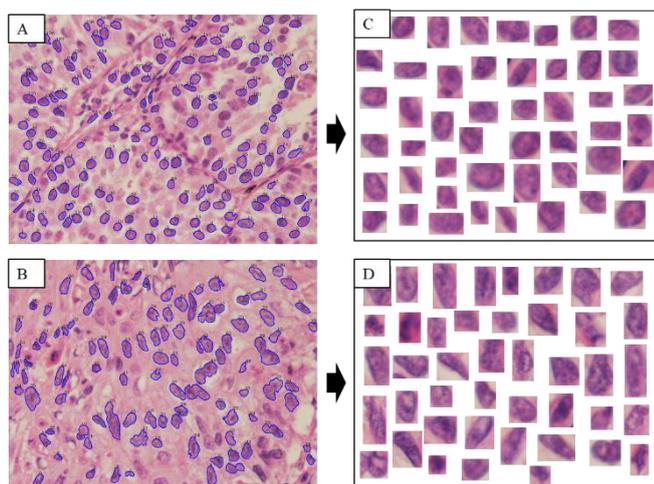


Figure 3. Dataset of two types of lung cancer: A. Segmented nucleus of Squamous cell carcinoma (Sca) marked in blue, B. Segmented nucleus of NeuroEndocrine Tumor (NET) marked in blue, C. Saved images from ROIs of A, D. Saved images from ROIs of B.

2.3 Experiments

Two types of lung cancer nuclear chromatin datasets were tested with four types of models, ResNet, InceptionNet, MobileNet, and DenseNet which are known to have the best classification ability for artificial intelligence-based classification. For the four state-of-the-art models, it was applied a learning rate of 0.00005, 20 epochs, a batch size of 16, and a dropout ratio of 0.5, and used the transfer learning with imageNet weights initialized. Since the number of data sets is small, data augmentation techniques such as 60 degree rotation, horizontal and vertical flip, and random zoom were applied to the dataset.

3. Results and Discussion

To compare and analyze the performance of our system, the accuracy, precision, sensitivity and f1 score were measured. Table 1 shows the classification results of two type lung cancers using nuclear chromatin with four CNN models. The best performance was Densenet, which gave the best results in metric of accuracy, precision, sensitivity, and f1score. Next, Resnet gave the second performance, and the worst result was InceptionNet, with an accuracy of only 81%. In InceptionNet, the sensitivity to NET was the lowest at 75%.

Table 1: Classification results for four models

		Accuracy	Precision	Sensitivity	F1 score
DenseNet	NET	0.86	0.86	0.88	0.87
	Sca		0.87	0.85	0.86
InceptionNet	NET	0.81	0.87	0.75	0.80
	Sca		0.76	0.88	0.82
MobileNet	NET	0.82	0.87	0.77	0.82
	Sca		0.78	0.88	0.83
ResNet	NET	0.86	0.87	0.87	0.87

	Sca		0.86	0.85	0.86
--	-----	--	------	------	------

Figure 4 shows the confusion matrix for the classification results of the four models. With the confusion matrix, we can see why MobileNet and InceptionNet are less accurate than the other two models. Of the two types of data, the classification accuracy for NET was poor. In particular, for the NET data among lung cancer, it can be visually confirmed that InceptionNet and MobileNet have lower classification accuracy than ResNet and DenseNet.

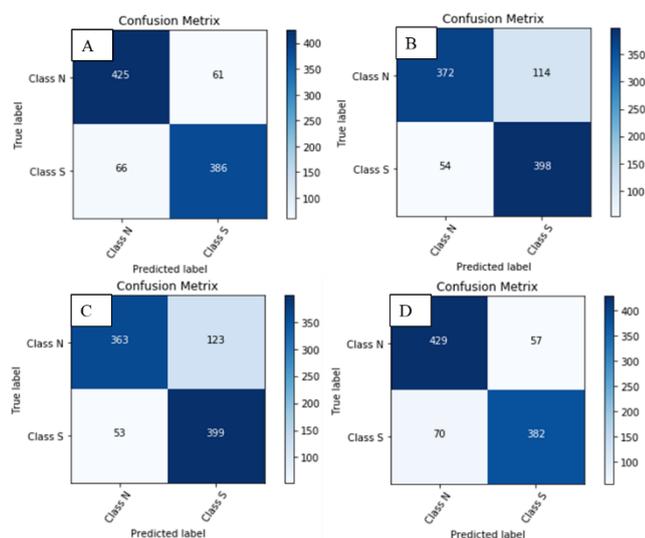


Figure 4. Confusion Matrices: A. ResNet, B. MobileNet, C. InceptionNet, D. DenseNet

Pathologists closely examine the texture features of nuclear chromatin to determine whether it is cancer or not. This is because nuclear chromatin contains a lot of information. In this study, an experiment was conducted to classify two types of lung cancer with only nuclear chromatin data with four types of CNN models that are known to show the best performance in classification. The classification accuracy of the top two models is 86% and the average accuracy of the four models is 84%, which proves that an automatic diagnosis system can be made that can support pathologists with only nuclear chromatin characteristics.

4. Conclusion

In this paper, an experiment was performed using an artificial intelligence model to see if the presence and type of cancer can be classified only by the features of nuclear chromatin. Two types of lung cancers, squamous cell carcinoma (Sca) and neuroendocrine tumor (NET) were used. Also a segmentation algorithm that automatically extracts nuclear areas from WSI based on the k-means algorithm was developed to automatically generate training and testing nucleus datasets. And lung cancer classification was performed using a dataset consisting only of nuclear chromatin regions with 4 types of deep learning models. As a result of the experiment, an average of 84% classification accuracy was obtained. The classification of lung cancer performed in this study is similar to the method that the pathologist classifies cancer based on the features of nuclear chromatin. The pathologist judges the presence and type of cancer by looking at the size and placement of the nuclei in the cell, and the texture characteristics. Since it showed

84% classification accuracy for the four types of models, it was proved that the automatic classification system using deep learning technology can be used to assist pathologists' diagnosis. In the future, to improve the accuracy, we are going to study on a model that combines a deep learning model and machine learning technology that uses biomarker features for analyzing nuclei texture patterns.

5. Acknowledgment

This research was supported by the Tongmyong University Research Grants 2019(2019A025-1).

6. References

1. Sheikh TS, Lee Y, Cho M. Histopathological Classification of Breast Cancer Images Using a Multi-Scale Input and Multi-Feature Network. *Cancers (Basel)*. 2020 Jul;12(8):2031.doi: 10.3390/cancers12082031. PMID: 32722111; PMCID: PMC7465368.
2. Abels E, Pantanowitz L. Current state of the regulatory trajectory for whole slide imaging devices in the USA. *J Pathol Inform*. 2017;8:23.
3. Ho J, Parwani AV, Jukic DM, Yagi Y, Anthony L, Gilbertson JR. Use of whole slide imaging in surgical pathology quality assurance: design and pilot validation studies. *Hum Pathol*. 2006;37(3):322–331.Pantanowitz L. Digital images and the future of digital pathology. *J Pathol Inform*. 2010;1:15.
4. Pantanowitz L, Sinard JH, Henricks WH, et al. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med*. 2013;137(12):1710–1722.
5. Bueno G, Deniz O, Fernandez-Carrobles MM, Vallez N, Salido J. An automated system for whole microscopic image acquisition and analysis. *Microsc Res Tech*. 2014;77(9):697–713.
6. Farahani N, Parwani A, Pantanowitz L. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathol Lab Med Int*. 2015(7):23–33.
7. Veta M., Pluim J.P., Van Diest P.J., Viergever M.A. Breast cancer histopathology image analysis: A review. *IEEE Trans. Biomed. Eng*. 2014;61:1400–1411. doi: 10.1109/TBME.2014.2303852.
8. Igloukov V.I., Rakhlin A., Kalinin A.A., Shvets A.A. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer; Cham, Switzerland: 2018. Paediatric bone age assessment using deep convolutional neural networks; pp. 300–308.
9. G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.
10. Howard, Andrew G., Zhu, Menglong, Chen, Bo, Kalenichenko, Dmitry, Wang, Weijun, Weyand, Tobias, Andreetto, Marco and Adam, Hartwig. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. (2017).arxiv:1704.04861.
11. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.
12. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.