An algorithm for classification of ovarian cancer histopathology images and prediction of genetic variants

Seo Jeong Shin[1], MS, Jin Roh[2], MD, Ph.D, Seng Chan You[3], MD, MS, Ho kyun Jeon[1], Kwang Soo Jeong[1], Suk-Joon Chang[4], MD, Hee-Sug Ryu[4], MD, Jang-Hee Kim[2], MD, Rae Woong Park[1, 3], MD, Ph.D
[1]Department of Biomedical Science, Ajou University Graduate School of Medicine, Republic of Korea; [2]Department of Pathology, Ajou University Hospital, Republic of Korea; [3]Department of Biomedical Informatics, Ajou University School of Medicine, Republic of Korea; [4]Department of Obstetrics and Gynecology, Ajou University School of Medicine, Republic of Korea

Is this the first time you have submitted your work to be displayed at any OHDSI Symposium?

Yes _____ No__√__

Abstract

Diagnosis of ovarian cancer is confirmed by microscopic visual analysis of histopathology slides. The development of a system that automatically performs classification of cancer tissue may reduce the burden of pathologists' workload. In addition, algorithms that predict BRCA1/2 mutation can also reduce cost burden of performing high-cost next-generation sequencing (NGS) test. In this study, a convolution neural network (CNN) algorithm was developed using histopathology images of patients with ovarian cancer and DNA sequence data from The Cancer Genome Atlas and Ajou University Hospital based on genomic CDM. The results showed that the algorithm could distinguish cancer with 81.22% accuracy and BRCA1/2 mutation with 90.37% accuracy. Additionally, by HeatMap visualization, our model localized the lesion well out of the whole-slide image. This suggested that the CNN model can be used as a clinical decision supporting system and as the primary screening system for ovarian cancer and additional NGS test.

Introduction

Ovarian cancer has the highest mortality rate among women, and 70% is found to be in the third, or higher, stage owing to the absence of early symptoms[1]. Visual analysis of histopathology slides by experienced pathologists is one of the main methods to evaluate the stage and subtypes of ovarian cancers. Recently, additional genetic screening tests such as NGS have been actively performed because a targeted therapy (poly-ADP ribose polymerase (PARP) inhibitors) has been approved for patients with deleterious or suspected deleterious germline or somatic *BRCA1/2*-mutated advanced ovarian cancer[2, 3]. The clinical decision supporting system, as the primary screening method, can be used to reduce costs and time spent on labor-consuming pathologic diagnosis workflow. In this study, we aimed to develop a deep learning algorithm for automatic classification of cancer tissues and prediction of genetic variants using histopathology images of patients with ovarian cancer and their DNA mutation data based on genomic CDM[4].

Methods

Overall study process is composed of data processing and model building (Figure 1).

1. Histopathology image and DNA sequence data

Hematoxylin and eosin-stained whole-slide images of 590 normal individuals or patients with high-grade serous carcinoma were downloaded from The Cancer Imaging Archive (TCIA). Among the patients, 282 available DNA sequence data acquired from the next-generation sequencing (NGS) tests were also downloaded from The Cancer Genome Atlas (TCGA). Whole-slide images for the same subtype of ovary cancer and NGS data of 34 patients were obtained from Ajou University Medical Center (AUMC).

2. Mutation status and labeling of tissue images

NGS data obtained from TCGA and AUMC were converted to Genomic-CDM, proposed as an extension of OMOP-CDM, and then compared with the state of changes using the data visualization and analysis tool "GeneProfiler"[4]. In particular, variants labeled as having pathogenic effects on genes such as *TP53*, *BRCA1*, and *BRCA2*, which are clinically important in ovarian cancer, were compared. Using the pathologist's review, images were divided into two groups (1,458 cancer and 1,525 non-cancer). According to the *BRCA1/2* mutation status in cancer groups, images were labeled as two groups (1,381 negative and 75 positive). Three regions of interest (ROI) were extracted from each slide and were used to train algorithms.

3. Convolution neural network (CNN) algorithm and HeatMap visualization

CNN was used to develop algorithms for cancer classification and *BRCA1/2* mutation prediction. The algorithms were learned using the ROI area, which is a part of the overall tissue image. HeatMap visualization was conducted to see if the algorithm accurately finds lesions in the entire image in the absence of the ROI assignment by pathologists. The whole-slide image was cut to the ROI size and then put into the model to predict whether cancer was present. The corresponding predicted values were used to represent colors at the position of the ROI.

Result

Mutational frequency analysis was conducted using the open source tool "GeneProfiler." Results showed that the frequencies of pathogenic mutations in AUMC were similar in TCGA, except for the *BRCA1* (Figure 2). Of these, the frequency of the *BRCA1/2* to be used to predict mutations was 8.15%–35.30%, which was consistent with the results from previous studies. Based on the presence/absence of cancer and *BRCA1/2* mutation, images were categorized as cancer/non-cancer group and positive/negative group, respectively; tissue images were labeled using CNN algorithms, and new image samples were used to predict the accuracy of the algorithms. The CNN algorithm developed using the labeled ROI was tested with new samples. The results showed that the accuracy in classifying a cancer tissue was 81.22% (AUC 0.89), and the accuracy in predicting *BRCA1/2* mutation in cancer tissues was 90.37% (AUC 0.90; Figure 3a). The results of the HeatMap visualization in the whole image indicated that cancer region tiles showed high cancer prediction scores compared with the overall region as we expected (Figure 3b).

Conclusion

In this study, using TCIA/TCGA data, which were public, and AUMC local hospital data, the CNN algorithm was developed to distinguish cancer and predict *BRCA1/2* mutation. The algorithm can be used successfully to support the pathologists' decision-making process, allowing to provide a higher level of care to patients by reducing the time spent in identifying images. Further studies using decoding and clustering methods to identify histologic properties contributing to cancer and *BRCA1/2* mutations predictions can be the basis for research on the pathogenesis of ovary cancer.

Acknowledgement

This work was supported by the Bio Industrial Strategic Technology Development Program (20001234) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [grant number: HI16C0992].

References

1. Lim MC, Won YJ, Ko MJ, Kim M, Shim SH, Suh DH, Kim JW. Incidence of cervical, endometrial, and ovarian cancer in Korea during 1999–2015. J Gynecol Oncol. 2019 Jan;30(1):e38.
2. Antoniou A, Pharoah PD, Narod S, et al. Average risks of breast and ovarian cancer associated with

BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet. 2003;72:1117-1130.
3. NCCN Guidelines for Patients: Ovarian Cancer, Version 1.2017
4. Shin SJ, You SC, Park YR, Roh J, Kim JH, Haam S, Reich CG, Blacketer C, Son DS, Oh S, Park RW. Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. J Med Internet Res 2019;21(3):e13249

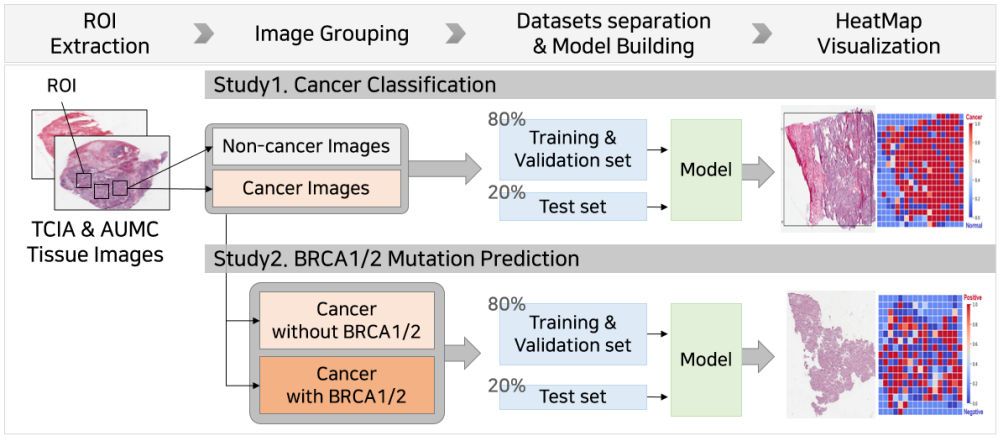

Figure 1. Overall study process. TCIA: The Cancer Image Archive; AUMC: Ajou University Medical Center; ROI: Region of Interest

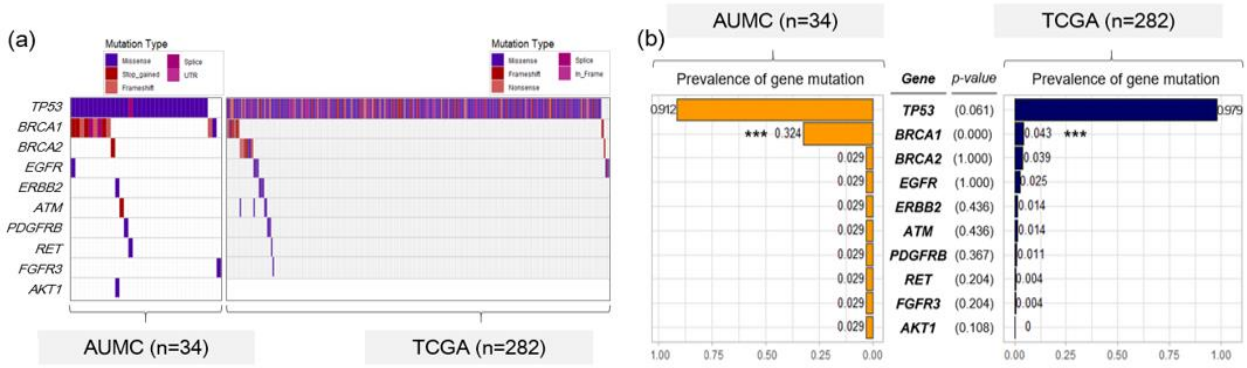

Figure 2. Sequence variance: (a) profile and (b) comparison between databases. The DNA change was occurred most frequently in *TP53* at both databases and only *BRCA1* has a different frequency between databases. AUMC: Ajou University Medical Center; TCGA: The Cancer Genome Atlas

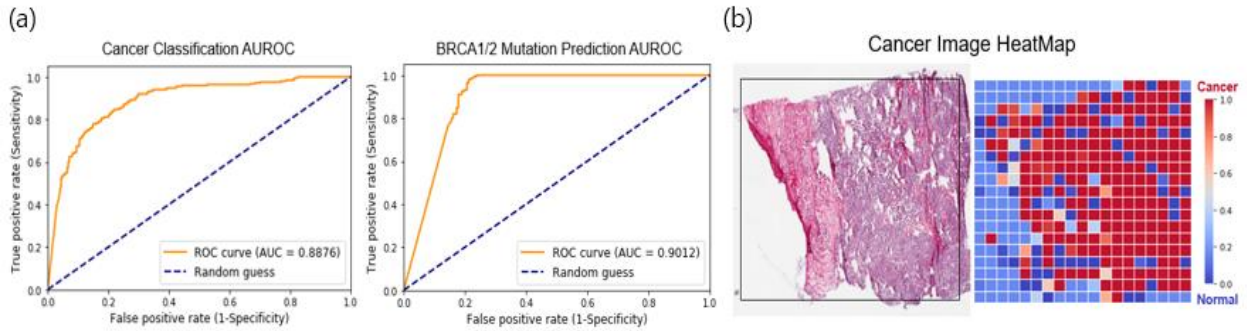

Figure 3. (a) Accuracy of the cancer classification model and the *BRCA1/2* mutation prediction model. (b) HeatMap visualization of the predictive value at the level of the whole slide image.