creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

**저작자표시**. 귀하는 원저작자를 표시하여야 합니다.

**비영리**. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

**변경금지**. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

**저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.**

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

[UCI]I804:48009-200000288670

공학박사 학위논문

# 신장이식 거부반응과 유방암 전이여부 예측을 위한 합성곱 신경망 기반 병리영상 분석

A study on digital pathology analysis using convolutional neural network for prediction of renal allograft rejection and breast metastasis

울 산 대 학 교 대 학 원
의 학 과
김 영 곤

# 신장이식 거부반응과 유방암 전이여부 예측을 위한 합성곱 신경망 기반 병리영상 분석

지 도 교 수　김 남 국　고 현 정

이 논문을 공학박사 학위 논문으로 제출함

2020 년 2 월

울 산 대 학 교 대 학 원
의 학 과
김 영 곤

김영곤의 공학박사 학위 논문을 인준함

심사위원 박 종 헌 (인)

심사위원 고 현 정 (인)

심사위원 김 남 국 (인)

심사위원 이 경 분 (인)

심사위원 이 준 구 (인)

울 산 대 학 교 대 학 원

2020 년 2 월

# 감사의 글

새로운 삶에 대한 희망찬 패기와 도전으로 시작했던 박사 학위, 비로소 그 과정을 이제 마치며 후회 없던 지난 시간들을 되돌아봅니다. 박사 학위를 시작한 2017 년부터 오늘까지 그 시간들은 저에게 있어서 중요한 성장의 시간이었고, 너무나도 감사했던 시간의 연속이었습니다. 그동안 옆에서 지켜봐 주시고 응원해 주신 많은 분들이 있었기에 버틸 수 있었으며 끝까지 완주할 수 있었다고 생각합니다.

결실을 맺기까지 부족한 저를 지도해 주시고, 그 어떤 학생도 받아볼 수 없는 값진 경험들을 쌓을 수 있도록 많은 기회를 주신 김남국 교수님, 서준범 교수님께 진심으로 감사의 말씀을 전합니다. 문제 해결을 위한 기술들의 폭넓은 이해와 접근 방법에 대한 생각을 발전시켜 주셨을 뿐만 아니라, 융합연구의 중요한 소통과 협업에 대한 올바른 방향성을 몸소 보여주심으로써 말로 다 표현하지 못할 많은 것들을 배울 수 있는 시간이었다고 생각합니다.

박사 생활에 있어서 많은 조언을 해 주신 주신 고현정 교수님, 이준구 교수님, 황정은 박사님, 이현나 박사님, 윤지혜 박사님, 배현진 박사님, 조용원 팀장님, 그리고 서로의 지식을 활발히 공유함으로써 공학에 재미를 더해 준 박범희, 우일상, Chenjiang Wu, 장령우, 병리 영상 연구팀에게도 감사의 말을 전하며, 연구실 생활에 있어서 좋은 추억을 가질 수 있도록 함께 생활했던 연구진들 모두에게 감사의 말씀을 전합니다.

마지막으로 항상 곁에서 힘이 되어주셨던 부모님께 이 논문을 바칩니다. 이 외에 미처 언급하지 못한 고마운 분들이 너무나 많습니다. 그분들의 이름을 모두 새기지 못함을 죄송하게 생각하며, 대신 제 깊은 감사의 말로 이 글을 마칠까 합니다.

Abstract

Pathology is the study of the causes and effects of diseases and the branch of medicine that aims to the final diagnosis. In order to diagnose the diseases, samples of tissues should be closely examined through the microscope. During this examination process, even experienced pathologists need to zoom-in and out several times while moving for all over tissue regions. Moreover, pathologic diagnoses mainly depend on visual scoring by pathologists, a process that can be highly time-consuming, laborious, and susceptible to inter- and/or intra-observer variations. Additionally, in the case of intraoperative situations, immediate decisions based on the frozen section examination are required. To overcome these issues, this paper provides some studies with deep learning algorithms for accurate and fast pathological diagnoses on two tasks; 1) *A fully automated system for prediction of renal allograft rejection,* 2) *Metastases classification in sentinel lymph nodes on frozen tissue section.*

For the first task, we propose a fully automated system using two different CNN methods. The fully automated system consists of two parts; Classification of regions of interest and detection of C4d positive and negative peritubular capillaries (PTCs) in giga-pixel immunostained slides. The performance of the detection method was evaluated using optimal parameters of the novel method to enlarge the size of labeled masks. Fifty and forty pixels of the enlarged size images showed the best performance in detecting the C4d positive and negative PTCs, respectively. Additionally, the feasibility of deep-learning-assisted labeling as an independent dataset to enhance detection in this model was evaluated.

For the second task, we analyzed two different approaches (classification/segmentation methods) for metastases classification in sentinel lymph nodes on frozen tissue section by comparing results of ours with that of others from the digital pathology challenge held by HeLP (2018 HEalthcare ai Learning Platform). Comparisons results showed that the convolutional neural network (CNN)-based classification method showed higher area under the receiver operating characteristic than that of segmentation method while the classification method took five times more than that of segmentation method. With ImageNet and

CAMLEYON16 data, we evaluated feasibility of using pre-trained models. The CAMELYON16 pre-trained model-based CNN classification model trained with equal or less than 80 frozen tissue section slides showed higher AUC than that of models with none of pre-trained model or ImageNet based pre-trained model while models trained with any of or none of pre-trained model with more than 80 frozen tissue section slides showed comparable AUCs to each other. When only less dataset is available, CAMELYON16 pre-train model enhanced the CNN based classification model.

The fully automated systems using deep learning on two tasks including prediction of renal allograft rejection and metastases classification in sentinel lymph nodes on frozen tissue section were developed and evaluated. The system of the first task is highly reliable, efficient, and effective, making it applicable to real clinical workflow. The investigation of the system of the second task might be helpful for efficient training, and fast and accurate diagnosis in the frozen diagnosis in intraoperative biopsy.

Key words: Deep learning, digital pathology, renal allograft, image classification, object detection, semantic segmentation, metastasis.

Contents

List of Tables

List of Figures

List of Abbreviations

| | |
|---|---|
| ACDC-LungHP | Automatic Cancer Detection and Classification in whole slide LUNG HistoPathology |
| AMC | Asan Medical Center |
| ASAP | Automated Slide Analysis Platform |
| AUC | Area Under the Curve |
| CAD | Computer-Aided Diagnosis |
| CAM | Class Activation Map |
| CAMELYON | CAncer MEtastases in LYmph nOdes challeNge |
| CNN | Convolutional Neural Network |
| CTVIE 19 | Computer Tomography Ventilation Imaging Evaluation 2019 |
| FFPE | Formalin Fixed Paraffin Embedded |
| FNR | False Negative Rate |
| H&E | Hematoxylin And Eosin |
| HeLP | HEalthcare ai Learning Platform |
| IDC | Invasive Ductal Carcinoma |
| ILC | Invasive Lobular Carcinoma |
| ISBI | International Symposium on Biomedical Imaging |
| JAMA | Journal of the America Medical Association |
| LYSTO | LYmphocyte aSsessmenT hackathOn |
| MC | Mucinous Carcinoma. |
| MICCAI | Medical Image Computing & Computer Assisted Intervention |
| OCT | Optimum Cutting Temperature |
| PAIP | Pathology AI Platform |
| PTC | PeriTubular Capillary |
| RFC | Random Forest Classifier |

| | |
|---|---|
| RGBA | Red, Green, Blue, Alpha |
| ROI | Region Of Interest |
| RSNA | Radiological Society of North America |
| SGD | Stochastic Gradient Descent |
| SVM | Support Vector Machine |
| TNR | True Negative Rate |
| TPR | True Positive Rate |
| VerSe'19 | large scale VERtebrae SEgmentation 19 |
| WSI | Whole Slide Image |

## 1. Introduction

For decades, research and development for computer-aided diagnosis (CAD) system using the conventional machine learning methods in radiology were developed by using texture[1] based SVM or Bayesian for segmentation of pulmonary embolism[2,3] and brain tumor[4], detection of polyp[5,6], breast cancer[7-11], and cognitive state of brain[12-14]. The conventional machine learning method based CAD systems in pathology have been developed as well. For instance, quantitative pathology using Bayesian[15], grading of lymphocytic infiltration[16], classification of cancer[17] using texture[18,19] or morphological feature[20,21], and segmentation of nuclei[22,23] and cancer[24] have been developed with the conventional machine learning methods. However, the conventional method based CAD systems produced more false positives than human readers[25], which resulted in additional biopsies[26]. Since a deep learning method that is based on artificial neural network with deep layers achieved the best performance[27] in the Large Scale Visual Recognition Challenge 2012 using ImageNet dataset[28], deep learning methods have been recently adopted to develop applications in many fields such as image recognition[27,29-31], speech recognition[32-34], gene expression[35,36], question answering[37], and language translation[38,39].

Deep learning performance of convolutional neural networks (CNNs) may be enhanced by using massive datasets. Although the use of massive manual labeled datasets is highly time-consuming, these datasets showed comparable performance to expert clinicians. For example, the diagnostic performance of a CNN model, trained using 130K fundus images, was comparable to that of expert ophthalmologists in diagnosing diabetic retinopathy[40]. Moreover, a CNN model trained using 130K dermoscopy images and patients' skin images was as accurate as dermatologists in distinguishing skin carcinoma from benign lesions[41].

Recently, implementation of digital pathology has been rising because of workforce crisis and increased need of consultation and collaboration. Digital pathology has many advantages in terms of time saving, slide storage, remote working, and second-opinion practice, and is becoming a part of routine procedure in diverse areas such as primary diagnosis,

multidisciplinary clinic, and frozen section diagnosis[42]. CNNs can be used to develop fully automatic pathologic diagnosis systems. Although it is ideal to examine the entire area of a specimen with a light microscope, it is impossible to closely examine all specific regions of each specimen in real clinical settings. Thus, to reach a final diagnosis, pathologists alter the magnification. Moreover, pathologic images are very complex, with eye fatigue reducing diagnostic accuracy over time. In addition, subjective evaluations might be susceptible to inter- and/or intra-observer variations. These drawbacks may be overcome by CNNs. For example, CNNs have been applied to giga-pixel immunostained images to detect breast cancer metastases to sentinel lymph nodes[43-45] and prostate cancer in biopsy specimens[43,46]. CNN models have also been applied to immunostained images to detect brain and colon cancers[47,48]. The CNN techniques using digital histopathological images have been investigated and showed satisfactory results in the detection of tumor areas and lymph node metastases in prostate, lung, and breast cancers[44,49,50].

In this paper, we provide some study designs using various with deep learning algorithms for accurate and fast pathological diagnoses on two tasks; 1) Metastases classification in sentinel lymph nodes on frozen tissue section. 2) A fully automated system for prediction of renal allograft rejection.

1-1. A fully automated system for prediction of renal allograft rejection

The demand for kidney transplants is increasing worldwide. Renal biopsy is the gold standard for the evaluation of allograft rejection. Deposition of C4d in peritubular capillaries (PTCs), the tiny blood vessels surrounding renal tubules, is an established marker of antibody-mediated allograft rejection[51]. C4d score, defined as the proportion of C4d positive PTCs on immunostaining[52], is one of the most important factors in the diagnosis of antibody-mediated rejection. Ideally, C4d score should be determined by counting all C4d positive and negative PTCs. However, it is practically impossible for pathologists to quantify all PTC samples, as the microscopic evaluation of PTCs is too time consuming, poor reproducible, and labor-intensive. Pathologists must therefore visually estimate the proportion of C4d positive PTCs.

However, they may overlook some microscopic foci or inaccurately estimate the proportion of C4d positive PTCs, and their estimates may be susceptible to inter- and/or intra-observer variations[53-55]. Because automated PTC counting may result in a more accurate diagnosis, deep learning studies using CNN models are required to diagnose allograft rejection in kidney transplant recipients.

To develop clinically applicable system to identify regions of interest and to detect C4d positive and negative peritubular capillaries in giga-pixel immunostained slides, we proposed and evaluated deep-learning-assisted labeling with more efficiency, enhancing the detection model with pathologists' insights with enlarged masks, and a fully automated system with combining CNN based classification and detection as routine pathologists' workflow to predict renal allograft rejection. The overall procedure is described in Figure 1-2. The system scans digital images from immunostained pathologic slides and removes background areas by Otsu's thresholding[56]. Histogram equalization is processed to reduce variations such as illumination or degree of staining. After selecting all candidate regions of interest (ROIs) with sufficient tissue in each slide, the CNN model classifies ROIs as feasible or non-feasible and detects all C4d positive and negative PTCs in feasible ROIs to determine C4d scores. To train the CNN detection model, enlarged masks with certain sized margins are used as input, so that each enlarged mask includes neighborhood information, such as renal tubules, present near the PTCs. Deep-learning-assisted labeling from independent dataset are used for results determined by the detection model which was trained using dataset by manual labeling. The effectiveness of the enlarged mask and deep-learning-assisted labeling was assessed by comparison with FROC. The CNN detection model using enlarged masks trained with margin sizes of 50 and 40 pixels performed better than those without enlarged masks for the detection of C4d positive and negative PTCs, respectively. In comparisons of deep-learning-assisted labeling, the CNN detection model trained with either or both data by deep-learning-assisted labeling performed better than the model trained with data by manual labeling.

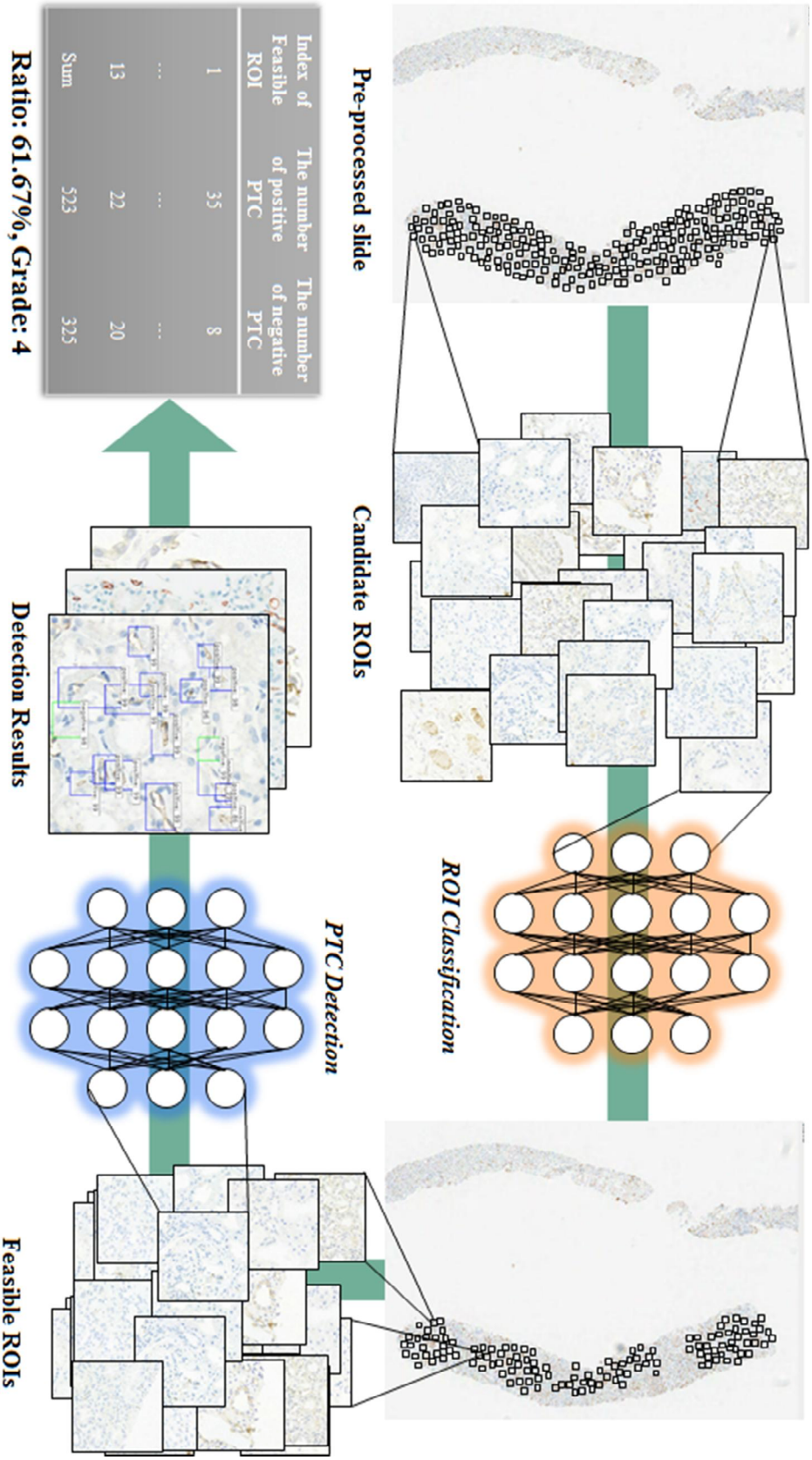



*Figure 1-1. Overall procedure for prediction of renal allograft rejection.*

1-2. Metastases classification in sentinel lymph nodes on frozen tissue section

Breast cancer is the most common cancer in women, accounting for approximately one-third of all cancers in women globally. For patients with localized breast cancer, the treatment of choice is surgical removal of the primary tumor[57]. In order to reduce disease recurrence or metastasis, lymph node sampling or dissection should be performed during surgery. Because axillary lymph node dissection may cause morbidity, such as arm-lymphedema and nerve injury, sentinel lymph node sampling is recommended in order to determine the nodal metastases status and if extensive lymph node dissection is required[58-61]. Although some recent studies suggested that the role of sentinel lymph node biopsy has been diminished in early breast cancer patients[62-65], sentinel lymph node sampling is still considered important due to its cost– and time– effectiveness and usually performed intraoperatively using the frozen section technique and which allows surgeons to make immediate decisions during surgery[66]. However, pathologists frequently experience problems while making diagnoses of frozen sections.

First, frozen section diagnosis should be made as quickly as possible in order to minimize the waiting time for surgeons which can cause surgical and anesthetic complications. The turnaround time of the frozen section diagnosis is usually kept less than 20 to 30 minutes, including the gross examination, tissue cutting and staining, and the microscopic examination[67]. Second, microscopic examination of a frozen section is more difficult than that of a conventional section because of inferior quality of the sections due to the frozen artifact. There are also components, such as capillaries, histiocytes, and germinal centers, in lymph nodes and which can be mistaken for metastatic carcinoma. Furthermore, frozen section diagnosis is extremely difficult in some patients who have underwent neoadjuvant systemic therapy before surgery. In order to overcome such difficulties, the deep learning algorithm might be helpful. For example, the ‘CAncer MEtastases in LYmph nOdes challeNge’ (CAMELYON16 and CAMELYON17) competitions disclosed that some deep learning algorithms achieved better diagnostic performance than a panel of 11 pathologists participating in a simulation exercise designed to mimic routine pathology workflow[44,68]. However, digital

slides which were used in most of those previous studies had not been created from frozen tissue sections, but from formalin fixed paraffin embedded (FFPE) tissue sections. To our best knowledge, there has not been any reported study using frozen tissue section of SLNs until the present time. In addition, the previous studies did not include post–neoadjuvant cases, which has been increasing but difficult to histologically examine[69].

For classification of digital slide showing cancer, CNN based classification model are used[70-73] to generate heat map that is used to determine if the cancer regions are detected or not in the digital slide. Figure 3-1 shows an example of the frozen tissue and the corresponding heat map. The frozen tissue with labeled marking (tumor region colored at green line) and the heat map of the tissue are shown in Figure 1-1 (a) and (b). The red color of the heat map shows high confidence of exist of the metastasis while the blue color of the heat map means no confidence of exist of the cancer. In case of CNN based segmentation that is faster than that of CNN based classification, that has been used for segmentation of nuclei[73,74] not the classification task.

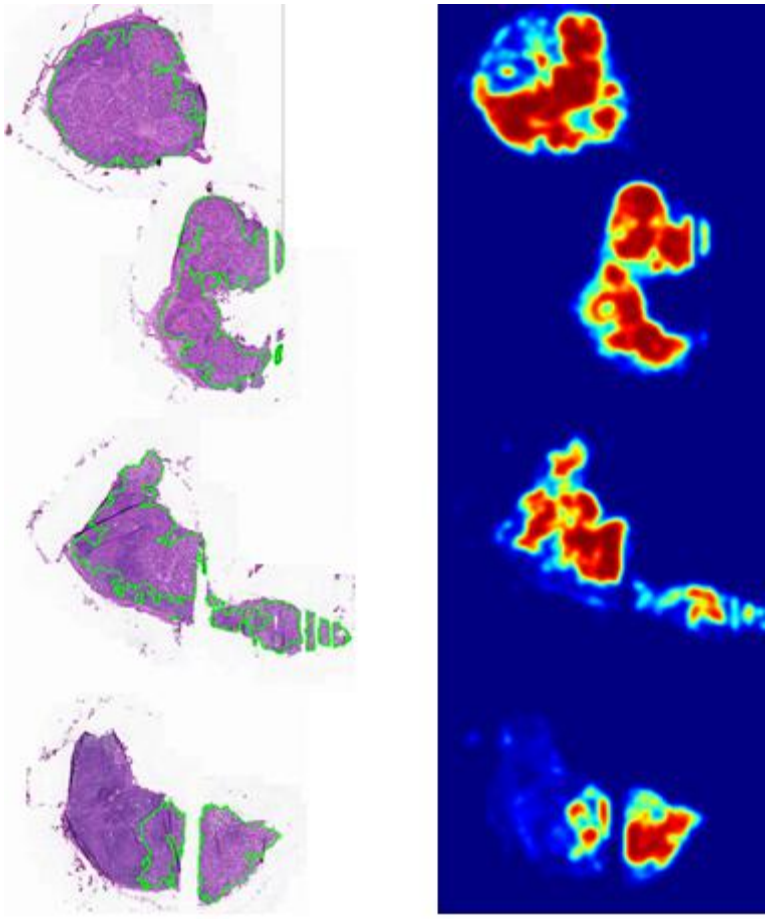



*Figure 1-2. An example of the frozen tissues and the corresponding heat maps. (a) WSI with metastasis regions annotated by boundaries colored at green. (b) A heat map where red color shows high confidence of exist of the metastasis.*

Considering the intraoperative biopsy, the inference time is another main factor to be applied to routine clinical practice. Thus, we investigated two different ways (CNN based classification/segmentation methods) for the same purpose. The results of our methods were compared results from one of the challenges originating from the HeLP (HEalthcare ai Learning Platform) contents that aimed to measure the model performances for classification of metastases per slide in hematoxylin and eosin–stained frozen tissue sections of SLNs in breast cancer patients.

## 1.3. Outline

The remainder of this thesis is organized as follows. In Chapter 2, backgrounds of machine and deep learning, and computer vision with CNN are described. In Chapter 3, a fully automated system for prediction of renal allograft rejection is described. In Chapter 4, metastases classification in sentinel lymph nodes on frozen tissue section is described. Finally, the discussion and conclusions of this thesis are described in Chapters 5.

## 2. Backgrounds

### 2.1. Machine learning and deep learning

Machine learning is a set of methods that automatically extract features from pattern in data[75,76]. The machine learning is a subtype of artificial intelligence, which constructs data driven mathematical model to make decision without explicit information needed by domain expert. There are many types of the machine learning method such as decision tree[77,78] that is one of the predictive modeling approaches in statistics, support vector machine (SVM)[79,80] constructing hyperplanes in a high dimensions, Bayesian [81,82] using probability distribution. Artificial neural network[83,84] inspired by the biological neural networks in brain is a one of the conventional machine learning methods, which allows the artificial neural network model to be trained with data and discover features for the purpose of classification, detection, or segmentation. The conventional machine learning methods except for the artificial neural network have limitations in their ability to extract features from the raw data. In other words, the conventional machine learning methods generally depends on feature extractor defined by domain expertise.

Since a deep learning method that is based on artificial neural network with deep layers achieved the best performance[27] in the Large Scale Visual Recognition Challenge 2012 using ImageNet dataset[28], the deep learning methods have been recently adopted to develop applications in many fields such as image recognition[27,29-31], speech recognition[32-34], gene expression[35,36], question answering[37], and language translation[38,39]. Human error to ImageNet dataset is approximately 5% while deep learning algorithms began to show lower error rate than that of human since 2015[85]. Gulshan et al.[40] proposed a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Approximately 130K images were labeled and used to train and validate the deep learning model, which showed comparable performance with expert ophthalmologists. Bejnordi et al.[44] assessed deep learning algorithms submitted for CAMELYON16 challenge for detection of lymph node metastases in women with breast cancer. Some of deep learning algorithm showed higher performance than that of

a panel of 11 pathologists and algorithm performance was comparable with an expert pathologist without time limitation. A study of CAD system using deep learning for lung nodule detection outperformed physicians[86]. Dunnmon validate a feasibility of using deep learning based CAD system for classification of abnormalities with chest X-ray, which showed 0.98 AUC [87].

2.2. Deep learning for CAD

Since a team who used the deep learning won the first prize in the 2012 ImageNet competition, which showed the beyond performance[27], the deep learning methods have been recently studied to overcome the issues of the conventional machine learning CAD systems. The deep learning has been adjusted for development of CAD systems such as radiology and pathology. Many of modalities in radiology such as chest radiography, CT, MRI, and mammography were used for analysis using deep learning methods. For example, deep learning using chest radiography or CT for classification of abnormalities[87-90], detection of lung lesions such as nodule[86,90-92], cancer[93-96], tuberculosis[97,98], and segmentation of lung[99,100] or bone structure[101] have been studied. Analysis using the deep learning has been developed for segmentation of lung[102], brain tumor[103], brain classes[104] and cartilage[105] and detection of breast cancer[106-109] in for MRI image, as well. CAD systems using digital slide in pathology have been studied with deep learning methods, as well. For instance, classification of glioma[70], prostate cancer[50], mutation state in prostate cancer[110], breast cancer[44,111], and segmentation of stroma and epithelial[112].

2.3. Open challenge with medical image

More thousands of papers or algorithms to solve problems in many of medical fields have been proposed every year, but few papers published in journal might be fairly validated and reproduced[113]. To compare each algorithm fairly, global open challenges in MICCAI (Medical Image Computing & Computer Assisted Intervention), ISBI (International Symposium on Biomedical Imaging), RSNA (Radiological Society of North America) have been organized

using medical data that data holder provides. Looking at the medical data available during the open challenge, it is possible to analyze what problems the hospitals or medical facilities want to resolve.

For instance of the pathology, LYSTO (LYmphocyte aSsessmenT hackathOn) challenge has been organized in 2019 MICCAI, which is for development of automatic assessment of lymphocytes in tissue sections of several types of cancer in human specimens stained with CD3 and CD8 immunohistochemistry. All participants were given 20,000 training patches with the corresponding number of lymphocytes. PAIP (Pathology AI Platform) organized a challenge for segmentation of liver cancer and estimation of viable tumor burden in MICCAI 2019. All participants were given a total of 100 WSIs for training and test dataset. ACDC-LungHP (Automatic Cancer Detection and Classification in whole-slide LUNG HistoPathology) challenge was held in 2018 MICCAI, which is for fast automatic analysis and detection of lung cancer in WSIs (Whole Slide Images) in order to reduce the burden for pathologists. A total of 200 H&E (Hematoxylin and Eosin) stained biopsy samples with cancer were given to all participants. CAMLEYON 16 and 17[68,114] were hold in ISBI 2016 and 2017 to reduce tedious and time-consuming tasks of pathologists. Two challenges were for classification and detection of metastasis slides, metastasis region, and pN-stage. Approximately 400 WSIs of lymph node with the corresponding mask files (.xml) were given for two challenges. In case of CAMLEYON 16, a review paper on competition has been published in JAMA (Journal of the America Medical Association)[44]. VerSe'19 (large scale Vertebrae Segmentation) challenge was held in MICCAI 2019 by providing 120 (+40 hidden) CT data with voxel-level vertebral annotations, which is for segmentation of vertebral. CTVIE 19 (Computer Tomography Ventilation Imaging Evaluation 2019) challenge provided 50 CT data with ventilation annotation, which is for segmentation of ventilation. With MRI image, segmentation of liver, spleen, and kidneys was another challenge by CHAOS challenge in ISBI 2019. The RSNA began holding open challenges using radiography every year. Challenges were about prediction of pediatric bone age with X-ray of their hand, pneumonia detection with chest X-ray, intracranial hemorrhage detection with brain CT in 2017, 2018, and 2019,

respectively.

## 2.4. Convolutional neural network

CNN is a one of the artificial neural networks and it can be applied to analyzing two- or three-dimensional images by extracting features in computer vision. Advantage of CNN based deep learning method is to extract not only low-dimension features, but also mid- and high-dimension features by itself in different to achieve the highest performance. The CNN based deep learning architecture for classification task is generally made up of consecutive multiple layers including convolution operation, pooling operation and connected layers as shown in Figure 2-1. The convolution operations are placed between input image layer and fully connected layer. The pooling operation and additional operation such as normalization operation such as mini-batch normalization, drop out, etc., are normally followed to the convolution operation.

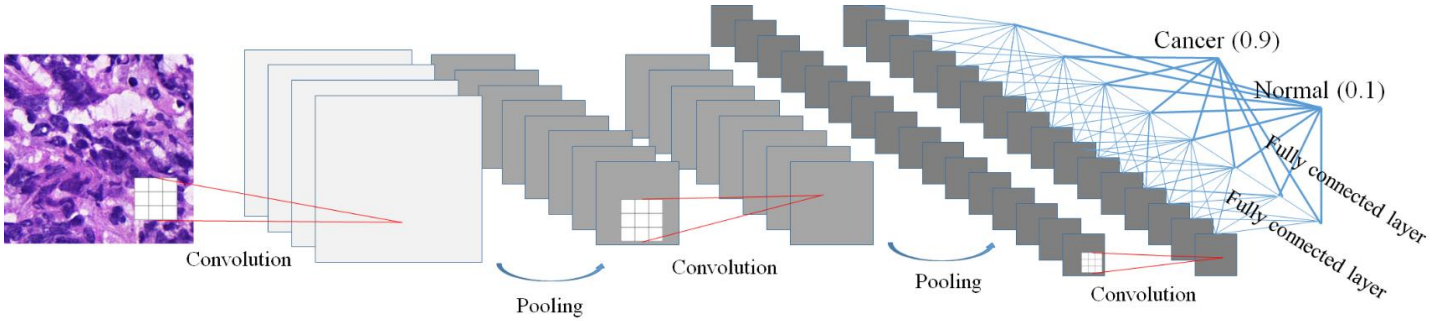



*Figure 2-1. Typical CNN architecture for classification.*

The convolution operation of CNN is a mathematical operation that multiply one function by another function within interval and then, integrate all results. Figure 2-2 shows an example of convolution operation with vertical Sobel filter[115]. The processing obtaining a feature value in a feature map was described by convolution operation. A feature map which resolution is almost same as the input image are obtained with stride 1 that moves the 3×3 filter one pixel at a time.

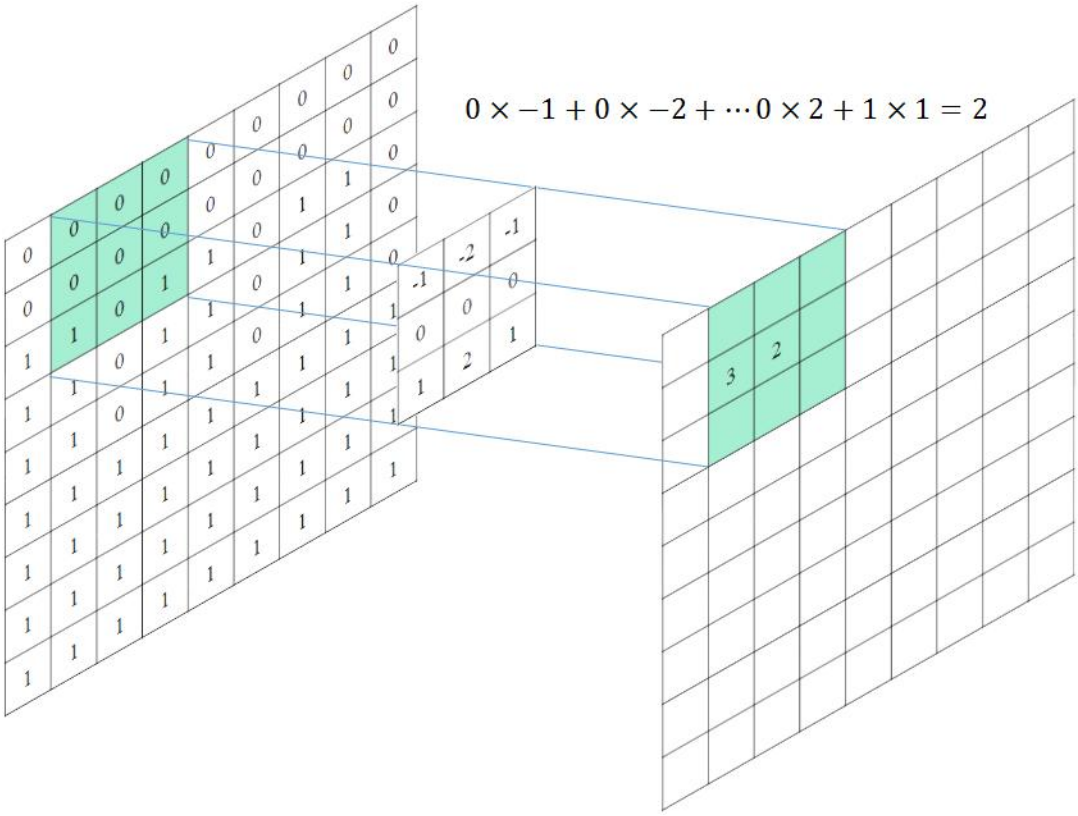



*Figure 2-2. An example of convolution operation using 3×3 filter with stride 2 in two-dimension image.*

Figure 2-3 shows an example of Sobel filter to extract vertical and horizontal features with convolution operation. Figure 2-3 (a) shows an input image and Figure 2-3 (b) and (c) shows 3x3 filters to extract vertical and horizontal edge features, respectively. Feature maps calculated by the vertical and horizontal filters are shown in Figure 2-3 (d) and (e), respectively. Diagonal or any features are extracted with filters defined by developer by modifying the values (i.e. weights) in the filter or enlarging the size of filter. The convolution operation is widely used in computer vision for not only edge extraction[116,117], but also noise reduction[118,119] and super resolution[120,121].

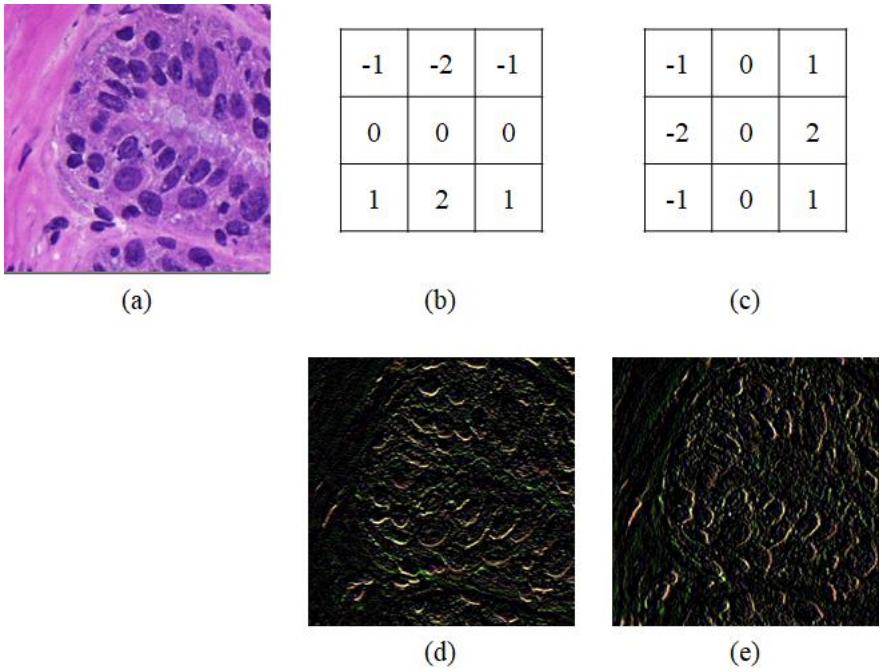



*Figure 2-3. An examples of Sobel filter for convolution operation. (a) Input image, (b) vertical filter for extracting vertical edge, (c) horizontal filter for extracting horizontal edge, (d) vertical feature map, (e) horizontal feature map.*

The pooling operation is for non-linear down sampling[122], which downsizes input image or feature map into a set of non-linear mapping as shown in Figure 2-4. There are several non-linear down sampling methods such as max pooling, min pooling, and average pooling. Figure 2-4 (a) shows a 4×4 input feature map and Figure 2-4 (b) and (c) shows result by max pooling and average pooling, respectively. Among them, the max pooling is the common non-linear down sampling method. The purpose of the pooling is for extracting dominant feature of the previous feature map by reducing the input dimension, but also controlling overfitting. In here, the stride parameter in convolution and pooling is used to control the output dimension. In the last two layers (i.e. fully connected layers), It connects all neurons of the first fully connected layer to that of next layer. The principle of the fully connected layer is same as the multi-layer perceptron [123] as shown in Figure 2-5.

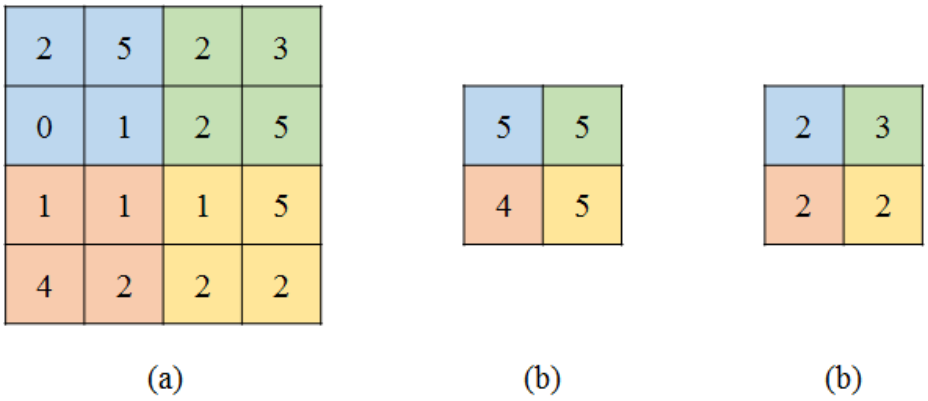



*Figure 2-4. Max pooling with a 2×2 filter with stride 2. (a) Input feature map, (b) result by max pooling, (c) result by average pooling.*

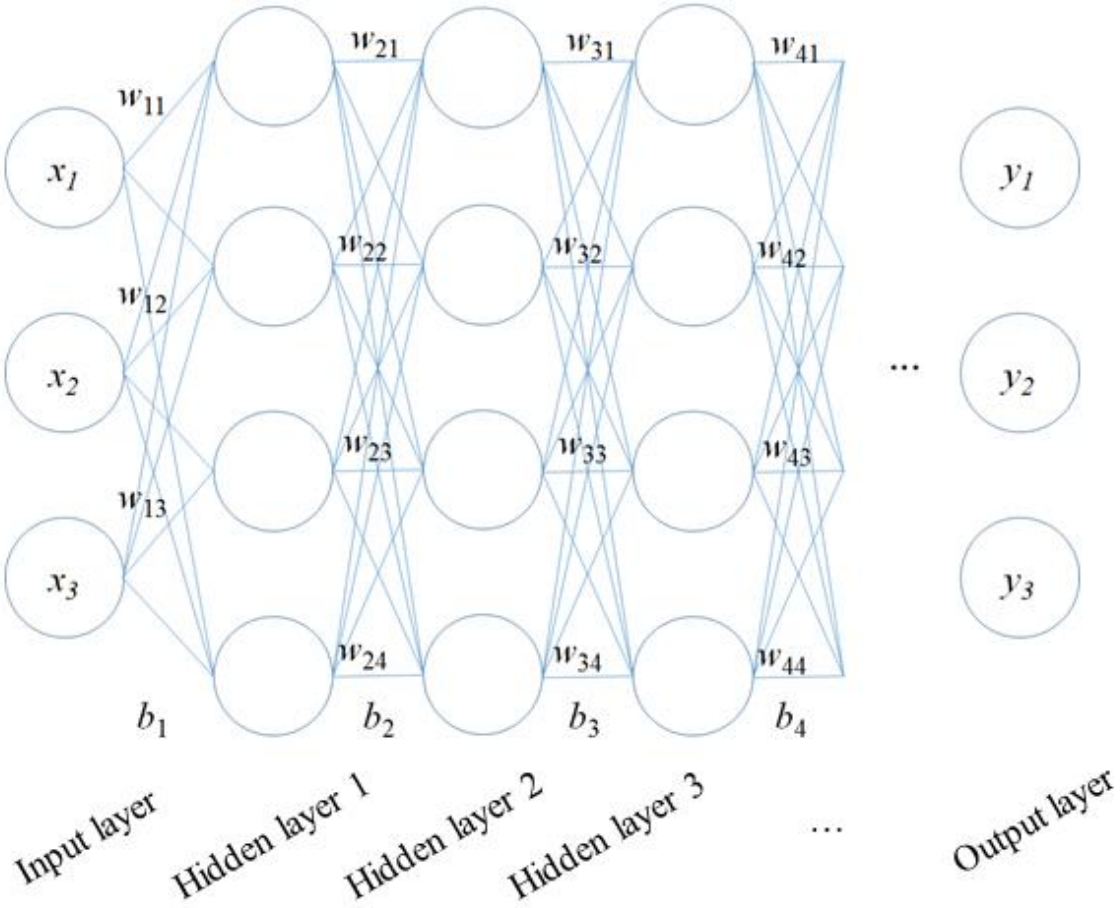



*Figure 2-5. An example of multi-layer perceptron.*

2.5. Whole slide image

The WSI, known as virtual microscopy, is a digitized pathology image scanned by imaging scanner. Jpeg 2000[124] are used to encode the WSI and decode a certain regions effectively. Jpeg 2000 is an image decoding technique using wavelet algorithm with multi-layer pyramid as shown in Figure 2-6, which is for offering accessibility to a certain regions without full decoding the WSI.

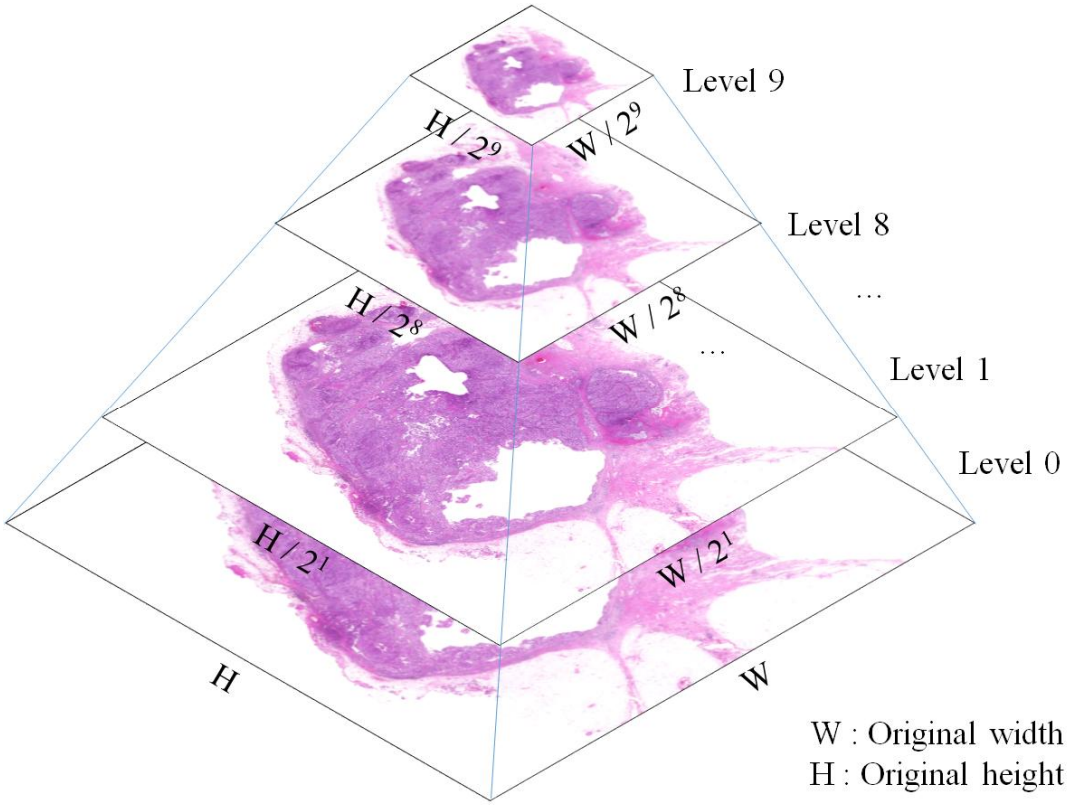



*Figure 2-6. Structure of multi-layer pyramid with 10 levels.*

The resolution of WSI is approximately more than 100K×100K, which needs tremendous memory allocation. It is impossible for developer to allocate the whole RGB values to a variable. To overcome this issue, OpenSlide library[125] that is a simple interface to read WSI efficiently was used to handle the WSI. The OpenSlide covers various types of file format

developed by vendors. Hamamtsu saves the WSI as various format such as “.vms", “.vmu”, and “.ndpi”. Leica saves the WSI as “.scn” and MIRAX saves the WSI as “.mrxs”, Philips saves the WSI as “.tiff”.

## 2.6. Pre-processing

*Stain normalization*

Stain normalization is a key factor to reduce color variations between stained tissue slides. That variations showing different color are caused due to different factors such as machine, stress of staining, manufacturer, technician, and hospital. Even though same hospital with same machine scanned the stained digital slide, the condition of staining can be scanned differently as shown in Figure 2-7. Patches of digital slides obtained from other patients in the same hospital with the same machine for the same purpose are shown in Figure 2-7 (a), which showed different color conditions. To reduce the color variations, stain normalization for reducing the color variations of colorectal cancer[126] was used as shown in Figure 2-7 (b).

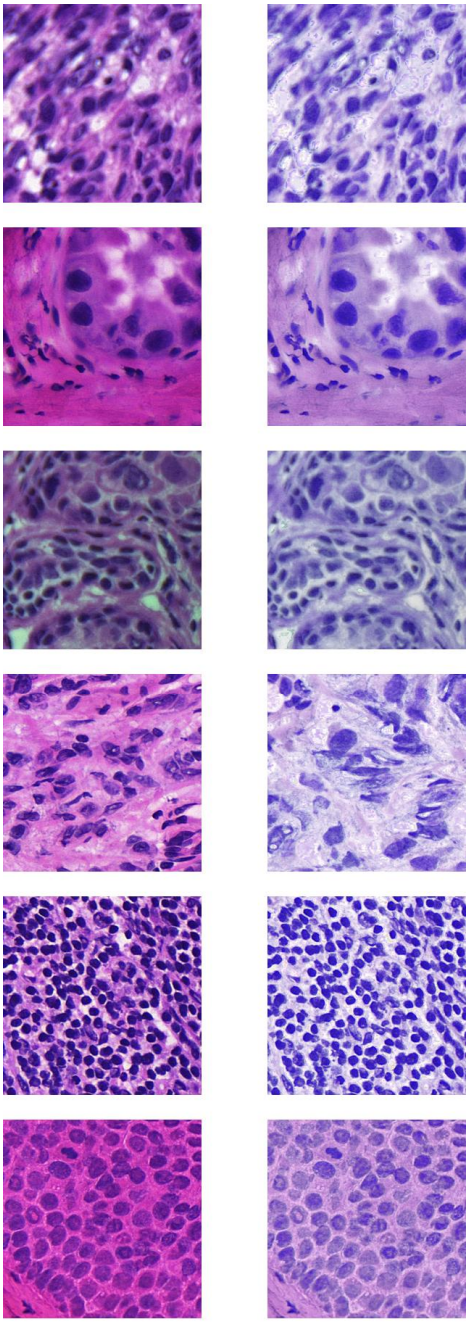
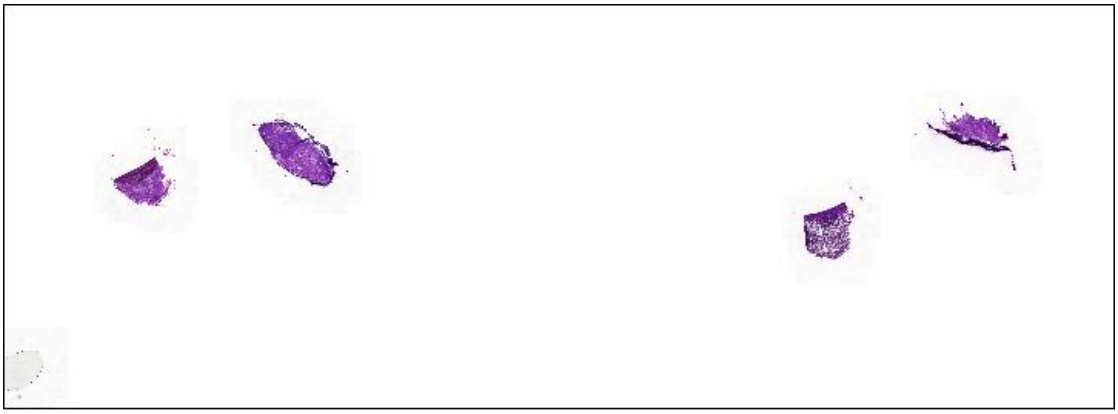



*Figure 2-7. An example of stain normalization. (a) Tumor and normal tissue patches showing color variations, (b) stain normalized patches.*

*Foreground segmentation*

Area of the tissue, foreground, obtained by biopsy is smaller than that of background regions. For efficient processing for extracting patches to be fed into CNN based model and inferencing heat maps, foreground segmentation is definitely required. Foreground of a total of 400 WSIs was account for approximately 22%. To segment foreground, Otsu's threshold[56] was used. Figure 2-8 shows an example of foreground segmentation. Figure 2-8 (a) shows an input WSI and (b) showed a foreground mask processed by Otsu's threshold.



(a)

(b)

*Figure 2-8. An example of foreground segmentation. (a) An input WSI and (b) a foreground mask by Otsu's threshold.*

*Patch-based approach*

Tissue section is scanned at from 20× to 40× magnification and it is digitized to the huge size of WSI which resolution is over 100K × 100K with RGBA (R, red; G, green; B, blue, A, Alpha). A total size of memory for allocating pixels is over 27G bytes (100K × 100K × 3 bytes), which is impossible to handle a variable on any of program language. Input size of 100K × 100K arrays is also impossible to be used as the input of CNN model. To overcome this, patch-based approach is used with sliding window with a certain size of stride as shown in Figure 2-9.

*Figure 2-9. An example of sliding window for patch extraction.*

When extracting patches with patch-based approach, appropriate level of the patch should be considered in order to train the CNN model efficiently. The level of the path highly depends on tasks. Figure 2-10 shows an example of showing patches with 448 × 448 size at different levels. Figure 2-10 (a)-(h) are patches at level 0 to 7. In case of classification of metastasis,

high level of patch such as 0 or 1 should be selected due to small size of cancer cells as shown in Figure 2-10 (a) and (b). Those kinds of objects which size is too small can be quite blurred and indistinguishable when low level such as 7 or 8 is selected as shown in Figure 2-10 (g) and (h). In case of classification of ductal carcinoma in situ, it is impossible to be classified when high levels such as 0 or 1 is used to extract patches, while middle levels such as 5 to 8 is appropriate to be distinguishable.

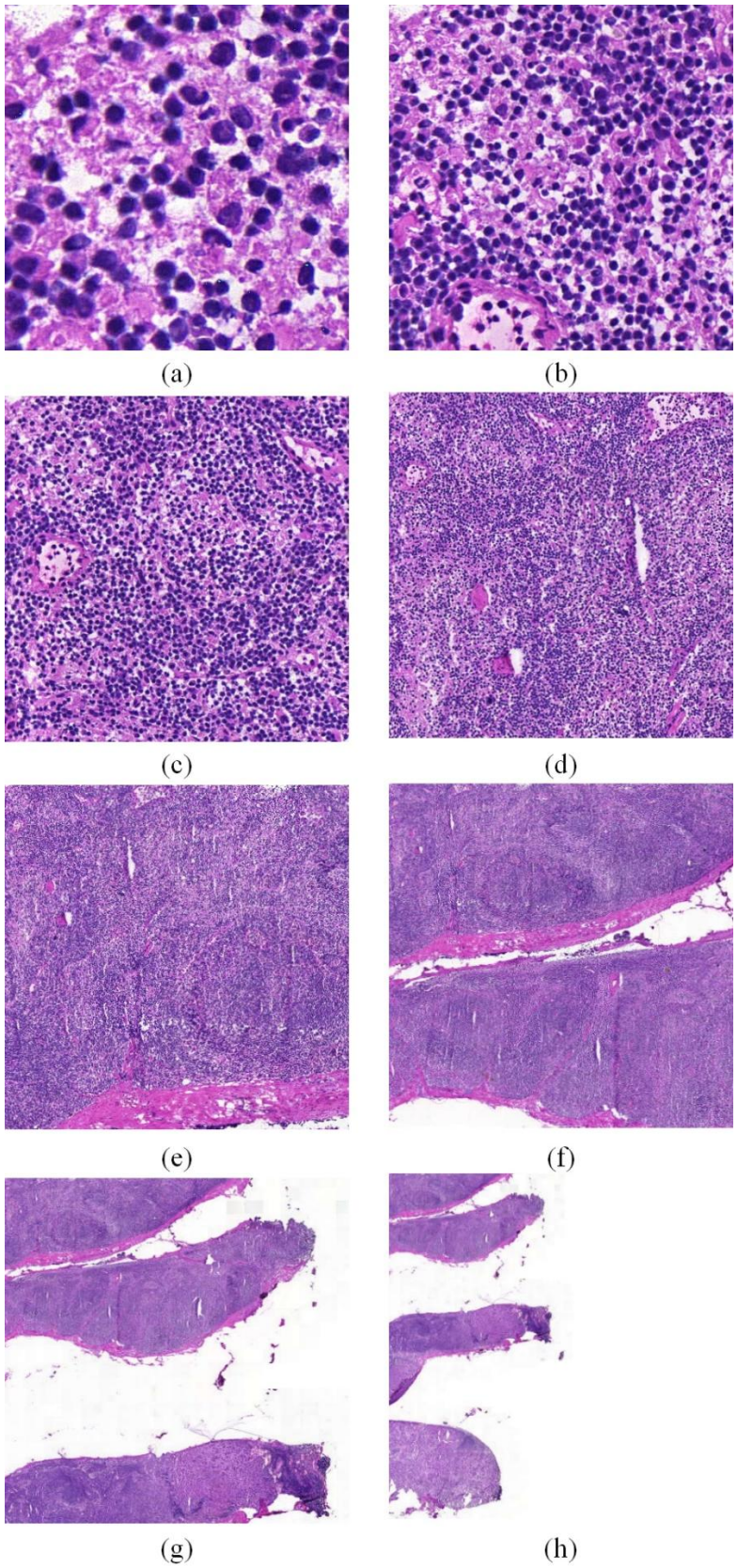



*Figure 2-10. An example of patches with 448×448 size at different levels from (a) 0 to (h) 7.*

2.7. Labeling tool

To mark the location of metastasis in WSI, the ASAP (Automated Slide Analysis Platform) software that is an open platform was used as labeling tool. The ASAP covers various types of format developed by vendors. It offers many of tools such as dots, polygons, and measurements to measure, mark, and draw the boundaries of objects. Figure 2-11 shows an example of drawing a region with the spline tool colored at green boundary.

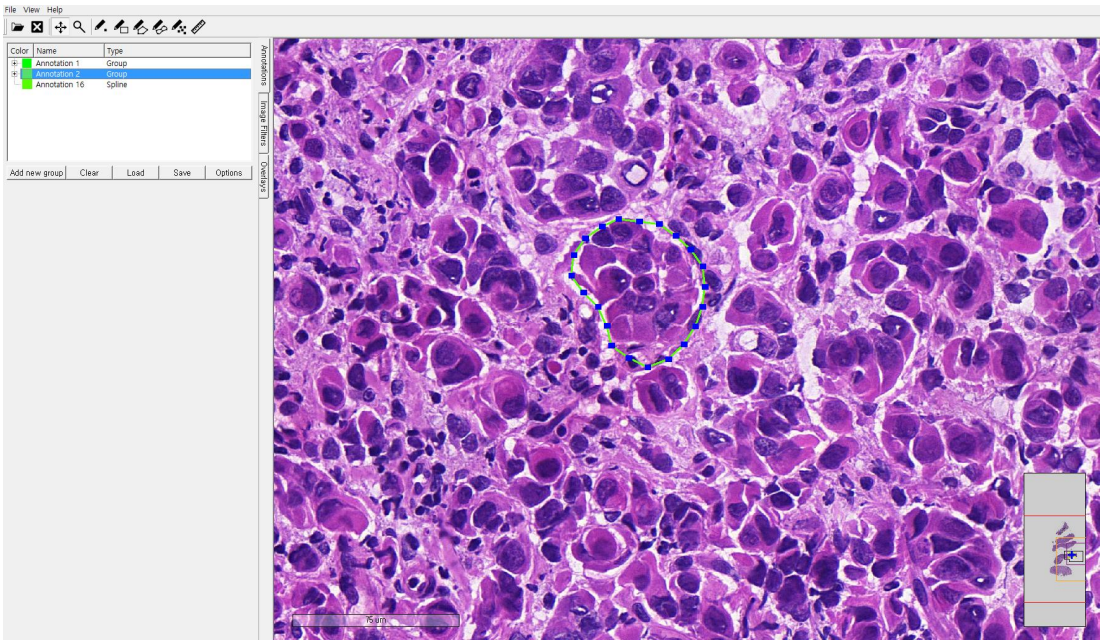



*Figure 2-11. An example of drawing a region with the spline tool colored at green boundary.*

The labeling tool, ASAP software, saves the labeling information as the ".xml", not the image file formats such as ".jpg", ".png", ",tiff" due to massive file size of the image. The labeling masks marked by pathologists were obtained by parsing the ".xml" files as shown in Figure 2-12.

```
import glob
import openslide
import xml.etree.ElementTree as ET

for file_mrxs in sorted((glob.glob(path_slide+'*.mrxs'))):

    file_name = str(file_mrxs.split('/')[-1])[0:-5]
    print ('file_name: ', file_name)


    if os.path.isfile(path_xml + file_name + '.xml'):

        ex_slide = openslide.OpenSlide(file_mrxs)
        size_level = ex_slide.level_dimensions[4]
        print ('size_level: ', size_level)
        temp_thumbnail = np.array(ex_slide.read_region ([0,0], 4, [size_level[0], size_level[1]]))

        tree = ET.parse(path_xml + file_name + '.xml')
        root = tree.getroot()

        mat_abnormal = np.zeros ((size_level[1],size_level[0]))

        ## --- Start Parsing XML ---
        ## Subset of annotation
        for reg in root.iter('Annotation'):
            points = []

            ## Subset of region
            for regs in root.iter('Region'):

                ## Subset of vertex
                for vertes in regs.iter('Vertex'):x
                    x = int(round(float(vertes.get('X'))))/int(16)
                    y = int(round(float(vertes.get('Y'))))/int(16)
                    points.append((x, y))
            cnt = np.array(points).reshape((-1, 1, 2)).astype(np.int32)

            mat_abnormal = cv2.fillPoly(mat_abnormal, [cnt], [255])

        mat_abnormal = mat_abnormal.astype(int)

    cv2.imwrite(path_out + file_name + '.png', mat_abnormal)
```

*Figure 2-12. Source codes to parse the ".xml" files to generate image files.*

## 3. A fully automated system for prediction of renal allograft rejection

### 3.1. Materials and methods

#### 3.1.1. Subjects

The institutional review board for human investigations at Asan Medical Center (AMC) approved the study protocol with removal of all patient identifiers from the images, and waived the requirement for informed consent, in accordance with the retrospective design of this study. A total of 380 needle biopsies of renal allografts were obtained from patients who underwent renal transplantation at AMC from 2009 to 2016; all samples had been stored in the Department of Pathology.

To obtain representative samples, cases in the period were randomly selected without consideration for specific pathologic diagnosis. Consequently, 108 C4d positive and 272 C4d negative cases were retrieved including 46 zero-day allograft biopsies. Two pathologists meticulously reviewed all slides and modified false negative results; finally, 189 cases were classified as C4d positive and 191 cases as C4d negative. C4d was assessed immunohistochemically using a Ventana BenchMark XT autostainer (Ventana Medical Systems, Tucson, AZ, USA), and 380 whole slides were imaged using a digital slide scanner (Pannoramic 250 Flash, 3DHISTECH, Budapest, Hungary) with a 20× objective lens (specimen-level pixel size, 0.221 × 0.221 μm). All samples were anonymized before analysis and labeling.

The 380 slides were divided into subsets 1 and 2, consisting of 200 and 180 slides, respectively, to validate the feasibility of using deep-learning-assisted labeling. Subset 1 was used to train the CNN model for classification and detection, and subset 2 was used to validate the model. The slides in the two subsets were randomized for training (60%), test (20%), and validation (20%). Computational complexity of feasible ROI classification and PTC detection takes constant time. The average time (and standard deviation) for automatic system per slide was 785.81 (176.97) sec. In this average time, the time for classification per slide and detection per ROI was 712.23 (±132.62) and 0.49 (±0.02) sec where the number of average feasible

ROIs was 147.61 (±105.02).

3.1.2. Methods

3.1.2.1. Feasible ROI classification

All randomly identified candidate ROIs were independently labeled by three pathologists as feasible or non-feasible criteria. Sensitivity was maximized by identifying as many feasible ROIs as possible within each slide. A ROI was classified as non-feasible when more than two-thirds of its image consisted of suboptimal areas, defined as 1) an artifact or poorly stained area that limited proper interpretation; 2) areas without PTCs, such as a large vessel, glomerulus, or vacant area; and 3) scarred or infracted areas[127]. Examples for criteria are shown in Figure 3-1.

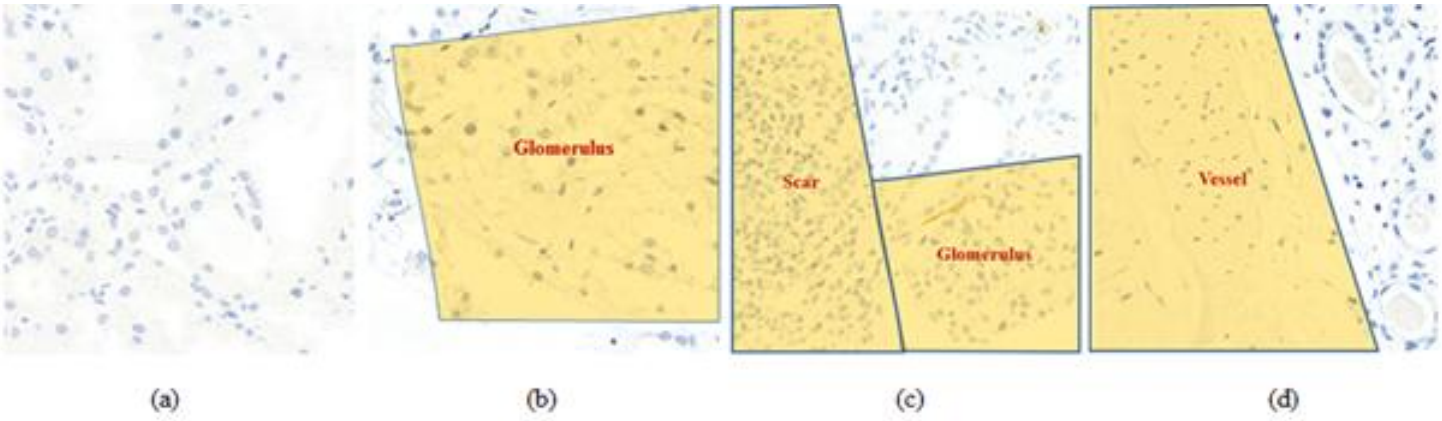



*Figure 3-1. Decision criteria to classify feasible and non-feasible ROIs. (a) Feasible ROI, (b)-(d) non-feasible ROIs from dominant ambiguous regions including scar, glomerulus, and vessels.*

A ROI size of 1024 × 1024 pixels was defined by the pathology team, as an image of this size provided a field of vision similar to a 400× optical microscopic view, the maximum magnification used in routine practice. The total number of ROIs in subset 1 was 2723, including 2134 feasible and 769 non-feasible ROIs.

The CNN classification model was trained to classify feasible and non-feasible ROIs

using an Inception V3 network[128] and an ImageNet pre-trained model[28]. To prevent model overfitting from unbalanced data, adjacent regions were assessed to equalize proportions between classes. Augmentation methods in real time included horizontal and vertical flipping, rotation (0–90°), and zooming in and out (0–10%). The model was implemented in Keras using the Tensorflow with an NVIDIA GTX 1080 Ti GPU, binary-cross entropy loss, stochastic gradient descent optimizer (SGD) with learning rates of 10−5, and dropout with probability of 0.5. The learning rate was reduced to one-tenth per one-third of total epochs 2000 and more detailed parameters are listed in Table 3-1. Training was terminated at the lowest loss of the test set. The performance of the CNN classification model was evaluated by determining its sensitivity and specificity.

*Table 3-1. Parameters used for training CNN classification model and CNN detection model.*

| Classification model | | Detection model | |
|---|---|---|---|
| Optimizer | SGD | Optimizer | Adam |
| Learning rate | 1e-5 | Learning rate | 1e-5 |
| Weight decay | 1e-6 | Weight decay | 0.0 |
| Epochs | 2000 | Epochs | 150 |
| Momentum | 0.9 | $\beta 1, \beta 2$ | 0.9, 0.999 |
| | | Epsilon | 1e-4 |

3.1.2.2. PTC detection

Three pathologists independently labeled C4d positive and C4d negative PTCs in feasible ROIs of subset 1 by hand drawing using in-house software. After completing these tasks by self, they had a meeting for the discussion of conflicted cases and made a consensus. In addition, pseudo negative PTCs, consisting of non-PTC regions, such as tubules and glomeruli that can be confused with PTCs, were drawn to train the model robustly.

Widths and heights of labeled PTC masks ranged from 25 to 392 pixels. A total of 1823

PTCs were identified by manual labeling, including 549 C4d positive and 1274 C4d negative PTCs, whereas a total of 3836 PTCs were identified from data by deep-learning-assisted labeling, including 1597 C4d positive and 2239 C4d negative PTCs. Examples are shown in Figure 3-2.

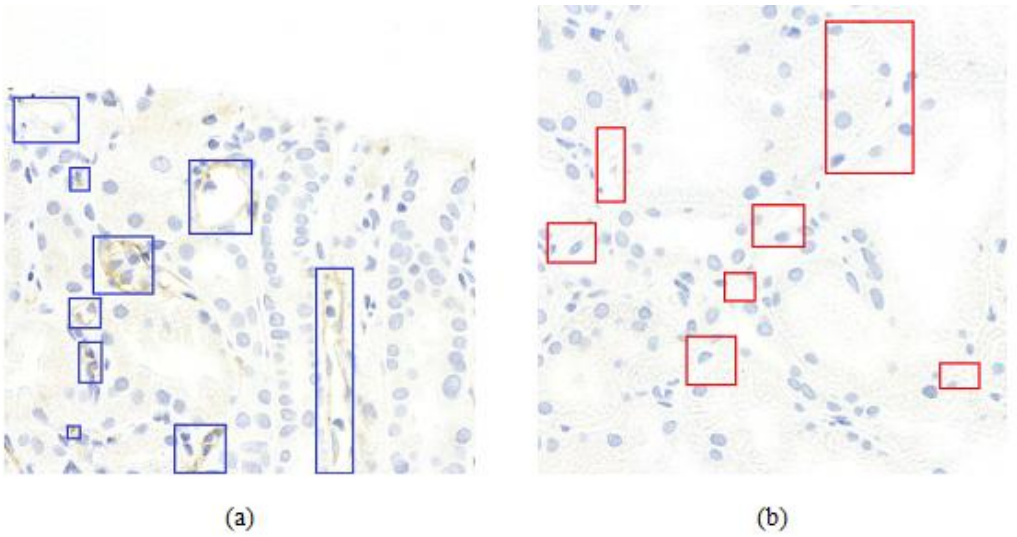



*Figure 3-2. Gold standard examples of C4d negative and positive in PTC. Blue and red rectangles show the positive and negative PTC in (a) and (b)..*

In general object detection tasks, the object is normally placed on a complex or object-independent background, with object boundaries determined using fitted coordinates. However, pathologic detection of PTCs is different. Because PTCs are capillaries located near tubules, their presence constitutes additional information during training using an enlarged rather than a fitted mask as shown in Figure 3-3. An enlarged mask was used because the boundaries of tubules near PTCs could help recognize PTCs on slides. To evaluate the optimal enlarged margin size around manual labeled data, various margin sizes (0–80 pixels at 10-pixel intervals) were adjusted when training the detection model.

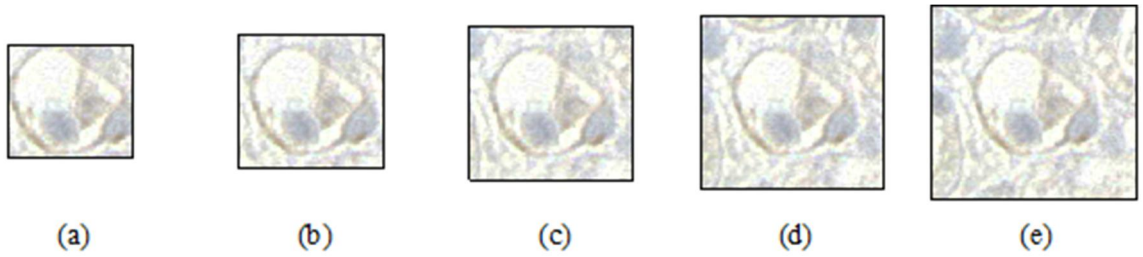



*Figure 3-3. Example of labeled C4d positive PTC with various margin sizes. Margin sizes of (a) 0, (b) 10, (c) 20, (d) 30, (e) 40 pixels.*

Labeling for massive amounts of data to enhance deep learning performance is time- and labor-intensive. These drawbacks may be overcome by using data by deep-learning-assisted labeling in place of or in addition to manual labeled data. A deep learning model trained for detection can be used to evaluate candidate objects first, followed by confirmation or modification involving little labor to acquire massive data, with the latter called data by deep-learning-assisted labeling. Deep-learning-assisted labeling can reduce the labor required. Figure 3-4 shows a process used to acquire deep-learning-assisted labeling for detection of C4d positive and negative PTCs. Firstly, two types of CNN model were trained from subset 1 with feasible ROIs and manual labeled mask data (Figure 3-4(a) and (b)). The CNN classification model trained from subset 1 was used to identify feasible ROIs in subset 2 (Figure 3-4(c)), and the CNN detection model trained from subset 1 was used to identify candidate C4d positive and negative PTCs in all feasible ROIs. Finally, data by deep-learning-assisted labeling were selected by confirming all candidate PTCs as being C4d positive or C4d negative using an in-house re-labeling tool (Figure 3-4(d)). In addition, this procedure was used to test false negative PTCs not detected by the model. If the center of the boundary box identified by the detection model did not deviate significantly from the center of the actual PTC, the PTC was confirmed as C4d positive or negative.

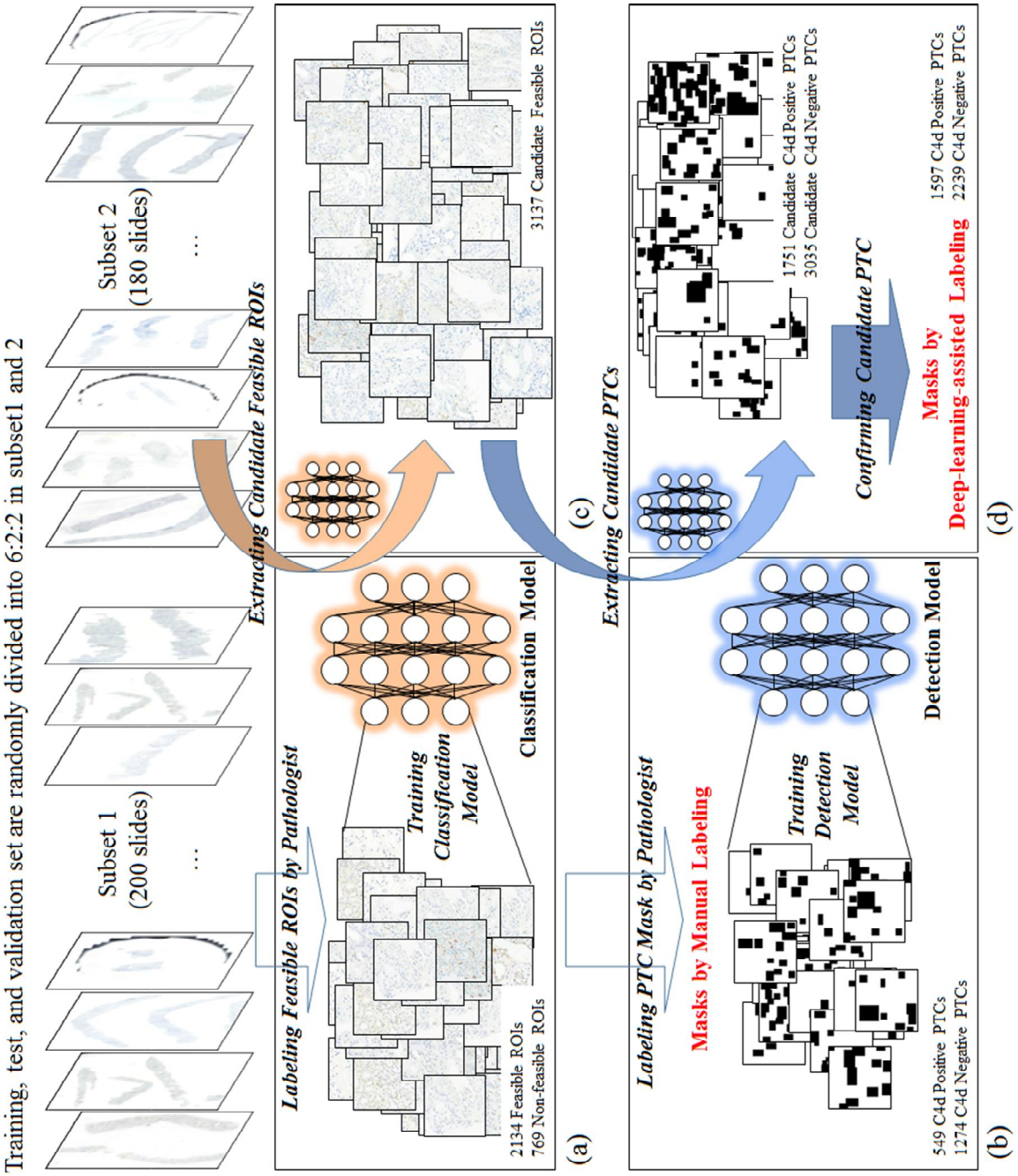



*Figure 3-4. Sequence for deep-learning-assisted labeling. All slides are randomly divided into 6:2:2 as training, test, and validation set in subset 1 and 2. (a) Training classification model with feasible ROIs in subset 1. (b) Training detection model with manual labeled masks in the feasible ROIs (c). Extracting candidate feasible ROIs in subset 2 by the classification model.*

*(c) Extracting candidate PTCs by the detection model and confirming results of (d) as deep-learning-assisted labeling.*

C4d negative and positive PTCs were detected using region proposal based Faster R-CNN detection algorithm4 with ImageNet pre-trained model based Resnet506. Augmentation methods in real time included horizontal and vertical flipping, rotation (0–90°), and zooming in and out (0–10%). The model was implemented in Keras using the Tensorflow with an NVIDIA GTX 1080 Ti GPU. Smooth L1 loss for bounding box regression and categorical-cross entropy loss for classification network in backbone were used. Adam optimizers with learning rates of 10−5 for region proposal and classification network. More detailed parameters are listed in Table 3-1. Training was terminated at the lowest loss of the test set. Training was terminated at the lowest loss of the test set.

To evaluate the effectiveness of enlarging margins with Faster R-CNN detection algorithm and of using data by deep-learning-assisted labeling with Faster R-CNN and one-shot based YOLO v2 detection algorithm [129], FROC scores, defined as the average sensitivity at seven predefined false positive rates (1/8, 1/4, 1/2, 1, 2, 4, and 8) per ROI, were calculated.

Stress test to see if the 380 slides datasets where the number of positive PTC and negative PTC masks were 2146 and 3513 are sufficient was conducted. To train the different detection CNN model performance with Faster RCNN for detecting C4d positive and negative PTCs with different amount of training data, all labeled data including subset 1 and subset 2 were used. All data were shuffled and divided into 80% and 20% as training and fixed validation set. Of training set, different training data were randomly selected to train each model at rates of 40%, 60%, 80%, and 100%. Test set for tuning each detection models were randomly selected at rages of 10% in each different training data. To measure performance for detecting C4d positive and negative PTCs, relative sensitivities at as sensitivity were calculated.

## 3.2. Results

3.2.1. Feasible ROI classification

The CNN classification model trained from subset 1 was tuned with high specificity to minimize false positives. The sensitivity and specificity of the CNN classification model were 0.7951 and 0.9941, respectively.

To validate the use of deep-learning-assisted labeling, the CNN classification model trained from subset 1 was used to determine candidate feasible ROIs in subset 2. This model was used to extract feasible ROIs from all tissue regions of subset 2 with high specificity. The mean ± SD number of ROIs per slide was 89.23 ± 34.22. An example of classification results for all regions of a slide is shown in Figure 3-5. Tissues containing feasible and non-feasible ROIs were colored red. ROIs containing tubules with PTCs were classified as feasible, whereas ROIs containing scars and glomeruli were classified as non-feasible. Figure 3-6 shows examples of CAM result for true positive cases in ROI classification to verify where regions are focused by the CNN classification model by using class activation map[130] (CAM).

*Figure 3-5. Feasible and non-feasible ROI classification results. Tissues including feasible ROIs are colored red.*

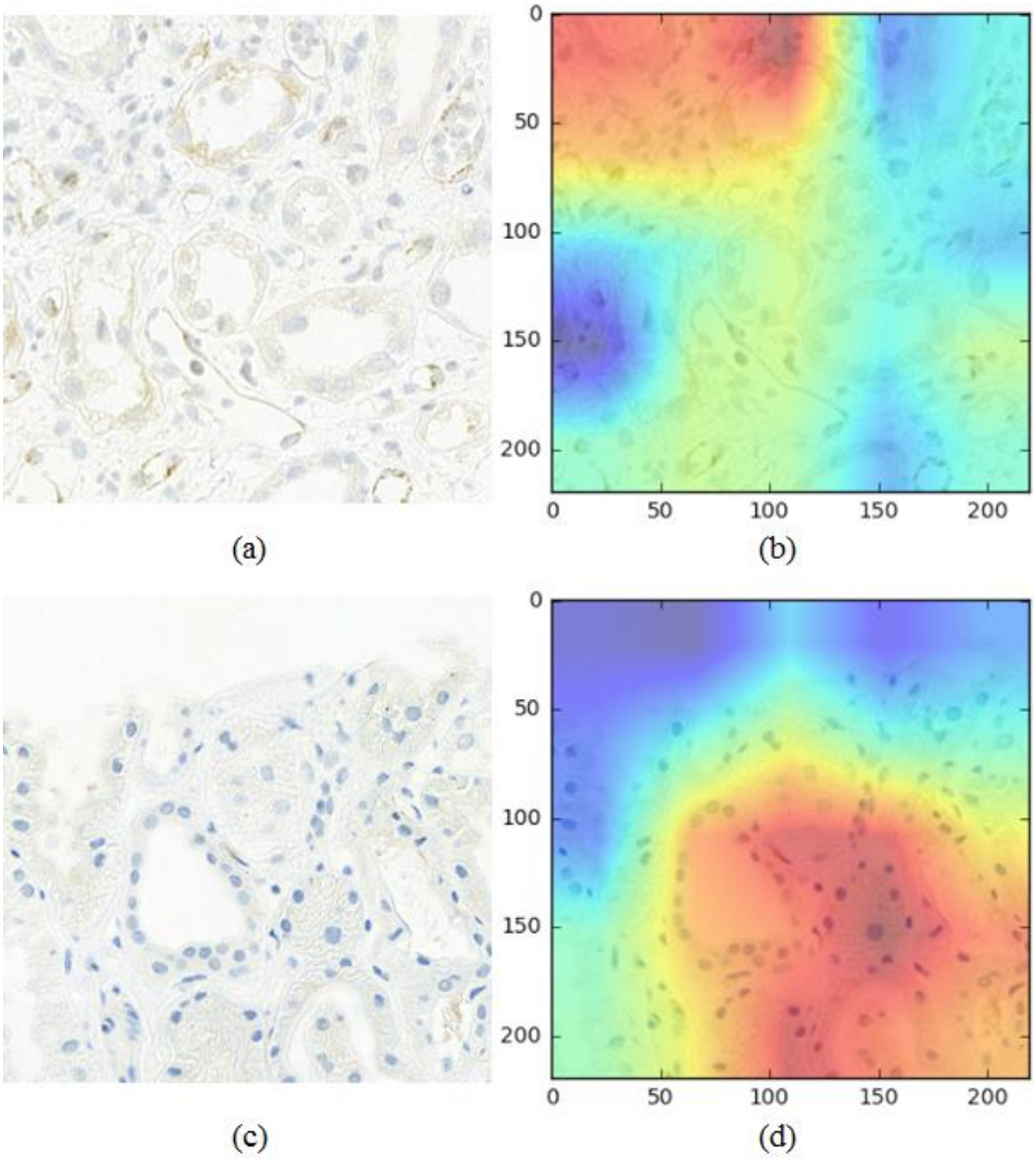



*Figure 3-6. Examples of CAM result for true positive cases in ROI classification. In CAM result, red color shows high confidence of exist of the ROI. (a) An input patch where many of PTC exist and (b) the corresponding CAM result showing high confidence for the left upper region. (c) An input patch where many of PTC exist and (d) the corresponding CAM result showing high confidence for overall region.*

3.2.2. PTC detection

The performances of validations using margin sizes of 0–70 pixels to detect C4d positive

and negative PTCs on manual labeled data were compared in Figure 3-7 and Table 3-2 with Faster R-CNN detection algorithm. FROC scores and overall sensitivities for the detection of C4d positive and negative PTCs increased as margin sizes increased. However, overall sensitivities and FROC scores in detecting C4d positive PTCs were optimal at margin sizes of 50 pixels, decreasing at 60 pixels (Figure 3-7(a)). Similarly, overall sensitivities and FROC scores in detecting C4d negative PTCs scores were optimal at 40 pixels (Figure 3-7(b)). FROC scores were highest for models trained with margin sizes of 50 and 40 pixels for the detection of C4d positive and negative PTCs, respectively.

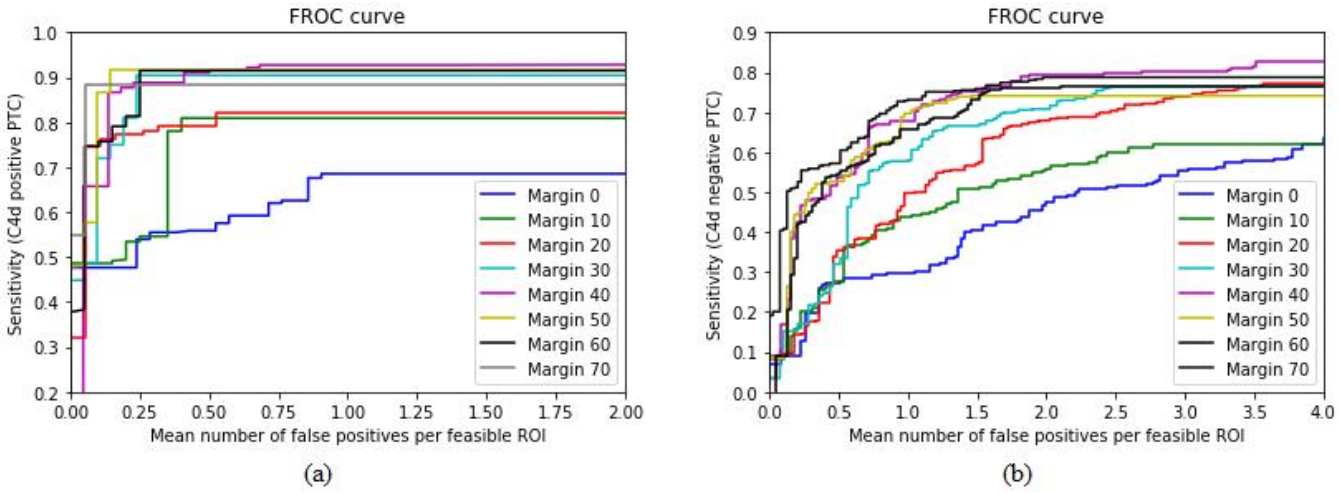



*Figure 3-7. FROC comparisons at different size of margin on manual labeled data. Results for detection of (a) C4d positive and (b) negative PTC.*

*Table 3-2. The sensitivities and FROC scores for Faster R-CNN detection of C4d positive and negative PTC with various margin sizes (0 to 70) at different mean number of false positives per feasible ROI.*

| **Mean** | **Margin size for detection of C4d positive PTC** |
|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **of FPs** | **0** | **10** | **20** | **30** | **40** | **50** | **60** | **70** |
| 0.125 | 0.4767 | 0.4862 | 0.7627 | 0.7200 | 0.6575 | 0.8667 | 0.7571 | 0.8851 |
| 0.250 | 0.5397 | 0.5345 | 0.7740 | 0.9045 | 0.8883 | 0.9167 | 0.8136 | 0.8851 |
| 0.500 | 0.5587 | 0.8092 | 0.7910 | 0.9045 | 0.9106 | 0.9167 | 0.9148 | 0.8851 |
| 1.000 | 0.6854 | 0.8092 | 0.8192 | 0.9045 | 0.9274 | 0.9167 | 0.9148 | 0.8851 |
| 2.000 | 0.6854 | 0.8092 | 0.8192 | 0.9045 | 0.9385 | 0.9167 | 0.9148 | 0.8851 |
| 4.000 | 0.6854 | 0.8092 | 0.8192 | 0.9045 | 0.9385 | 0.9167 | 0.9148 | 0.8851 |
| 8.000 | 0.6854 | 0.8092 | 0.8192 | 0.9045 | 0.9385 | 0.9167 | 0.9148 | 0.8851 |
| **Score** | 0.6166 | 0.7238 | 0.8006 | 0.8781 | 0.8856 | **0.9095** | 0.8778 | 0.8851 |

| **Mean** | **Margin size for detection of C4d positive PTC** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **of FPs** | **0** | **10** | **20** | **30** | **40** | **50** | **60** | **70** |
| 0.125 | 0.0919 | 0.0935 | 0.0976 | 0.1519 | 0.2306 | 0.0944 | 0.0918 | 0.4087 |
| 0.250 | 0.1258 | 0.2019 | 0.1453 | 0.1706 | 0.4663 | 0.4485 | 0.4248 | 0.5549 |
| 0.500 | 0.2733 | 0.2783 | 0.3541 | 0.3200 | 0.5388 | 0.5272 | 0.5465 | 0.5720 |
| 1.000 | 0.2952 | 0.4367 | 0.4978 | 0.5774 | 0.6789 | 0.6969 | 0.6571 | 0.7293 |
| 2.000 | 0.4655 | 0.5553 | 0.6792 | 0.7080 | 0.7920 | 0.7412 | 0.7611 | 0.7876 |
| 4.000 | 0.6217 | 0.6195 | 0.7743 | 0.7633 | 0.8274 | 0.7412 | 0.7633 | 0.7876 |
| 8.000 | 0.7257 | 0.6195 | 0.7743 | 0.7655 | 0.9004 | 0.7412 | 0.7633 | 0.7876 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Score** | 0.3713 | 0.4006 | 0.4746 | 0.4938 | **0.6334** | 0.5700 | 0.5725 | 0.6611 |

The CNN detection models trained with margins of 40 pixels for the detection of C4d positive and negative PTCs were tuned to maximum sensitivity to generate as much data by deep-learning-assisted labeling as possible. Deep-learning-assisted candidate labeled data were generated by the CNN detection models, which have a recall and precision of 0.8821 and 0.9384, respectively, for the detection of C4d positive PTCs, and of 0.8094 and 0.7108, respectively, for the detection of C4d negative PTCs. The characteristics of manual and deep-learning-assisted labeling differed slightly, in that manual labeled data only included masks fitted to both classes, whereas deep-learning-assisted labeling also included masks that were slightly misplaced locally. Figure 3-8 (a) and (b) shows inter- and intra-observer variations, respectively, between subset 1 (manual labeling) and subset 2 (deep-learning-assisted labeling). To validate the feasibility of using deep-learning-assisted labeling, FROC scores and sensitivities were compared in models trained with data by manual labeled, data by deep-learning-assisted labeling, and both together at different mean numbers of false positive PTCs per feasible ROI with two different type of detection algorithm. In detecting C4d positive and negative PTCs, the Faster R-CNN model showed better accuracies than those of YOLO v2 model. In addition, both models trained by subset 2 and fusion dataset including subset 1 and subset 2 showed better accuracies (Figure 3-8 (c) and (d) and Table 3-3).

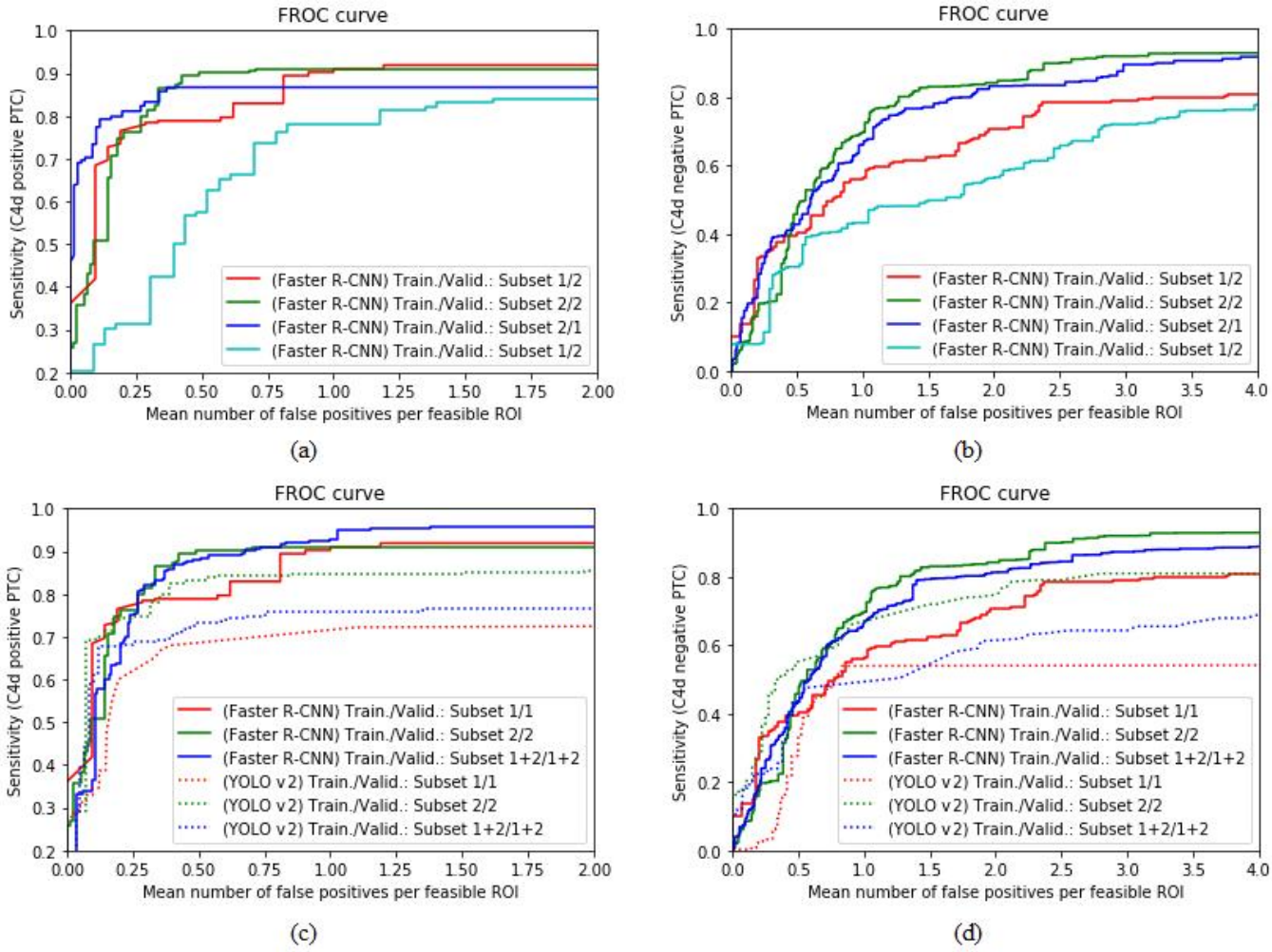



*Figure 3-8. FROC comparisons for validation of feasibility of using deep-learning-assisted labeling. FROC comparisons to show inter- and intra-observer variation between different validation set for detection of (a) C4d positive and (b) negative PTC with Faster R-CNN algorithm. FROC comparisons to validate effectiveness of deep-learning-assisted labeling for detection of (c) C4d positive and (d) negative PTC with Faster R-CNN and YOLO v2 algorithms.*

*Table 3-3. The sensitivities and FROC scores for Faster R-CNN and YOLO v2 detections of C4d positive and negative PTC with different detection models trained by different dataset at*

*different mean number of false positives per feasible ROIs (0 to 2 and 0 to 8 for detection of positive and negative PTC, respectively). Model 1: trained by subset 1, Model 2: trained by subset 2, Model 3: trained by fusion of subset 1 and 2.*

| Mean of FPs | Faster R-CNN | | | | | |
|---|---|---|---|---|---|---|
| | Detection model for C4d positive PTC | | | Detection model for C4d negative PTC | | |
| | **Model 1** | **Model 2** | **Model 3** | **Model 1** | **Model 2** | **Model 3** |
| 0.125 | 0.6970 | 0.5495 | 0.5768 | 0.1387 | 0.0863 | 0.1148 |
| 0.250 | 0.7803 | 0.6923 | 0.7510 | 0.3333 | 0.1969 | 0.2405 |
| 0.500 | 0.7886 | 0.8791 | 0.8817 | 0.3966 | 0.4579 | 0.4343 |
| 1.000 | 0.9024 | 0.9451 | 0.9253 | 0.5615 | 0.6969 | 0.6644 |
| 2.000 | 0.9187 | 0.9478 | 0.9585 | 0.7082 | 0.8424 | 0.8131 |
| 4.000 | 0.9187 | 0.9478 | 0.9647 | 0.8075 | 0.9294 | 0.8887 |
| 8.000 | 0.9187 | 0.9478 | 0.9647 | 0.8563 | 0.9294 | 0.8910 |
| **Score** | 0.8463 | 0.8442 | **0.8603** | 0.5431 | **0.5913** | 0.5781 |

| Mean of FPs | YOLO v2 | | | | | |
|---|---|---|---|---|---|---|
| | Detection model for C4d positive PTC | | | Detection model for C4d negative PTC | | |
| | **Model 1** | **Model 2** | **Model 3** | **Model 1** | **Model 2** | **Model 3** |
| 0.125 | 0.3864 | 0.7009 | 0.6736 | 0.0058 | 0.2034 | 0.1928 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.250 | 0.6284 | 0.7479 | 0.6795 | 0.0032 | 0.3644 | 0.2240 |
| 0.500 | 0.6817 | 0.8333 | 0.7329 | 0.2945 | 0.5512 | 0.4494 |
| 1.000 | 0.7124 | 0.8462 | 0.7567 | 0.5394 | 0.6617 | 0.4761 |
| 2.000 | 0.7221 | 0.8547 | 0.7565 | 0.5423 | 0.7480 | 0.6121 |
| 4.000 | 0.7444 | 0. 8761 | 0.7864 | 0.5423 | 0.8100 | 0.7106 |
| 8.000 | 0.7444 | 0.8846 | 0.7864 | 0.5423 | 0.8100 | 0.7345 |
| **Score** | 0.6599 | **0.8112** | 0.7388 | 0.3528 | **0.5926** | 0.4856 |

The CNN detection models for detecting C4d positive and negative PTCs trained with different amount of training dataset were compared as shown in Figure 3-9. The performances in detection of C4d positive and negative PTC were shown to be saturated at around 300 slides.

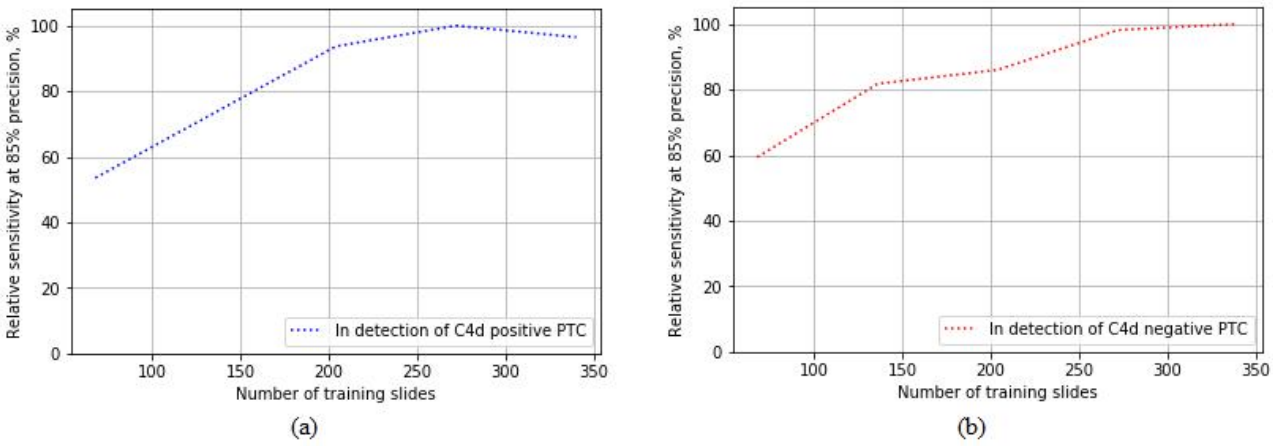



*Figure 3-9. Relative sensitivities comparisons of detection models trained with different amount of training data for detecting (a) C4d positive and (b) negative PTC with Faster R-CNN detection algorithm.*

### 3.3. Discussions

To develop clinically applicable system, deep-learning-assisted labeling with more efficiency, enhancing the detection model with pathologists' insights, combining CNN based classification and detection for a fully automated system were developed in this paper. Training the CNN detection models with enlarged masks surrounding PTC region is inspired by the actual pathologists' insights, which enhanced the detection performance highly (Figure 3-7 and Table 3-2). Deep learning models generally need massive data for training. To overcome the problem of small dataset, we tried to determine a feasibility of deep-learning-assisted labeling that is made from independent dataset with low-labor, which could alleviate massive manual labor. In our experiments, using deep-learning-assisted labeling for training, the performance of the detection model was enhanced (Figure 3-8 (c) and (d) and Table 3-3) compared with only dataset by using manual labeling only, because deep-learning-assisted labeling help draw PTC masks more robustly with less variations such as inter- and / or intra-observer, illumination, and degree of staining. In addition, massive dataset would improve the performance of the CNN detection models. In the stress test, we showed that 380 slides were sufficient to train CNN detection model for finding positive and negative PTC.

Deep learning on pathologic images could depend on various scanning conditions such as not only illumination, but also the degree of staining, different equipment, and so on. To overcome these problems, pre-trained network which has been already trained with bunch of tremendous number of images having a variety of complex variation, histogram normalization that is one of stain normalization method, and doubled dataset by deep-learning-assisted labeling were used to train our deep learning model to have robustness of them.

The most important aspects of application of this system to pathology are full automation for objective diagnosis and alleviation of manual labor. This study proposed a fully automated two-step CNN system for the diagnosis of allograft rejection. The first step consists of the use of a CNN classification model to identify feasible ROIs in all tissue regions and the second step consists of the use of a CNN detection model to identify and count C4d positive and negative PTCs, a marker of allograft rejection in kidney transplant recipients. These findings

suggest that this system may be applicable to most tasks in digital pathology.

Classification of all tissue regions as feasible or non-feasible ROIs using the CNN classification model is practical, as pathologists cannot determine all feasible ROIs in a tissue sample and have difficulty identifying negative PTCs. By contrast, the CNN classification model can precisely evaluate the entire specimen, and the CNN detection model can accurately count the numbers of C4d positive and negative PTCs in all feasible ROIs. Determining both C4d positive and C4d negative PTCs may alter clinical diagnoses.

In addition, two kinds of performance comparisons were conducted. Firstly, the performances of models trained with different size of margin including PTC region were compared. Enlarged mask with a certain size improved detection CNN model, which method was mimicked by a real clinical experience. This novelty including surrounding regions could be used widely for similar tasks. Secondly, the performances of models trained with data by manual labeling and data by deep-learning-assisted labeling were compared. Additionally, false positive cases detected by the detection model were used to fine-tuning the model as the second label data, which might enhance any of other CNN based methods such as classification, segmentation, and detection methods. Nevertheless, many false positives in Glomerulus occurs in C4d positive and negative PTC. We could use false positive reduction by training the CNN model for detection of the Glomerulus so that the predicted box could be excepted if those are detected in the Glomerulus regions. The generation of data by deep-learning-assisted labeling and confirmation by expert pathologists may help improve the performance of these models. Pathologic labeling is very difficult, even for expert pathologists, whereas the deep-learning-assisted method generated relatively robust labels.

Several obstacles should be overcome before clinical application. The sample size (380 slides) is about 1.2 times the average renal allograft biopsy per year in this center, which is one of the largest medical center in South Korea. Though it is also relatively larger than other studies related to pathologic assessment using convolutional neural network, we will try to evaluate the performance of this more with wild dataset from larger data. Also, all cases were recruited from a single center using only one slide scanner, which could lead to less variations

such as background illumination or degree of staining. To evaluate the robustness of this method, further studies with multi-center could be needed. In addition, comparisons of the performance and outcomes of this method with those of pathologists are needed to determine the clinical effectiveness of this system.

## 4. Metastases classification in sentinel lymph nodes on frozen tissue section

### 4.1. Materials and methods

#### 4.1.1. Subjects

During routine surgical procedure for breast cancer in our institution, the excised sentinel lymph nodes were immediately submitted for frozen section. All of the sentinel lymph nodes were cut into 2-mm slices, entirely embedded in OCT (Optimum Cutting Temperature) compound, and frozen in -20 to -30℃. For each lymph node, 5μm-thick frozen sections were cut and one or two sections were picked up on glass slides and stained with hematoxylin and eosin (H&E). In this study, a total of 297 digital slides of sentinel lymph nodes from 132 patients were retrospectively collected. Among those, 144 slides were made from sentinel lymph nodes of patients who had received neoadjuvant therapy (48.5%). The slides were divided into a training set, a development set and a validation set (157, 40, and 100 digital slides, respectively). Patient demographics are summarized in Table 4-1. The slides were scanned using a digital microscopy scanner (Pannoramic 250 FLASH; 3DHISTECH Ltd., Budapest, Hungary) in MIRAX format (.mrxs) and with a resolution of 0.221μm per pixel. The institutional review board for human investigations at Asan Medical Center (AMC) approved the study protocol with removal of all patient identifiers from the images and they waived the requirement for informed consent, in accordance with the retrospective design of this study.

*Table 4-1. Clinicopathologic characteristics of the patients.*

| | | **Training set (n = 157)** | **Development set (n = 40)** | **Validation set (n = 100)** | **P-value*** |
|---|---|---|---|---|---|
| Age (median and range) | | 50 (28 – 80) | 49 (30 – 68) | 47 (34 – 75) | |
| Sex | Female | 157 (100%) | 40 (100%) | 100 (100%) | 1 |
| Metastatic carcinoma | Present, size > 2mm | 68 (43.3%) | 14 (35%) | 40 (40%) | 0.158 |
| | Present, size ≤ 2mm | 35 (22.3%) | 5 (12.5%) | 15 (15%) | |
| | Absent | 54 (34.4%) | 21 (52.5%) | 45 (45%) | |
| Neoadjuvant systemic therapy | Not received | 80 (51.0%) | 28 (70%) | 45 (45%) | 0.027 |
| | Received | 77 (49.0%) | 12 (30%) | 55 (55%) | |
| Histologic type | IDC | 149 (94.9%) | 32 (80%) | 86 (86%) | 0.005** |
| | ILC | 8 (5.1%) | 5 (12.5%) | 11 (11%) | |
| | MC | 0 (0%) | 0 (0%) | 3 (3%) | |
| | metaplastic carcinoma | 0 (0%) | 3 (7.5%) | 0 (0%) | |
| Histologic grade | 1 or 2 | 118 (75.2%) | 34 (85%) | 86 (86%) | 0.074 |

| | | | | |
|---|---|---|---|---|
| | 3 | 39 (24.8%) | 6 (15%) | 14 (14%) |

* P-values, calculated using the $\chi 2$ test.

** For the histologic type, a $\chi 2$ test was conducted between IDC and non-IDC.

4.1.2. Reference standard

All the imaging datasets were segmented manually by one rater, and their annotations were approved by two clinically expert pathologists with six and 20 years' of experience in breast pathology. Regions of metastatic carcinoma larger than 200μm in the greatest dimension were annotated as Cancer with the in-house labeling tool, as shown in Figure 4-1.

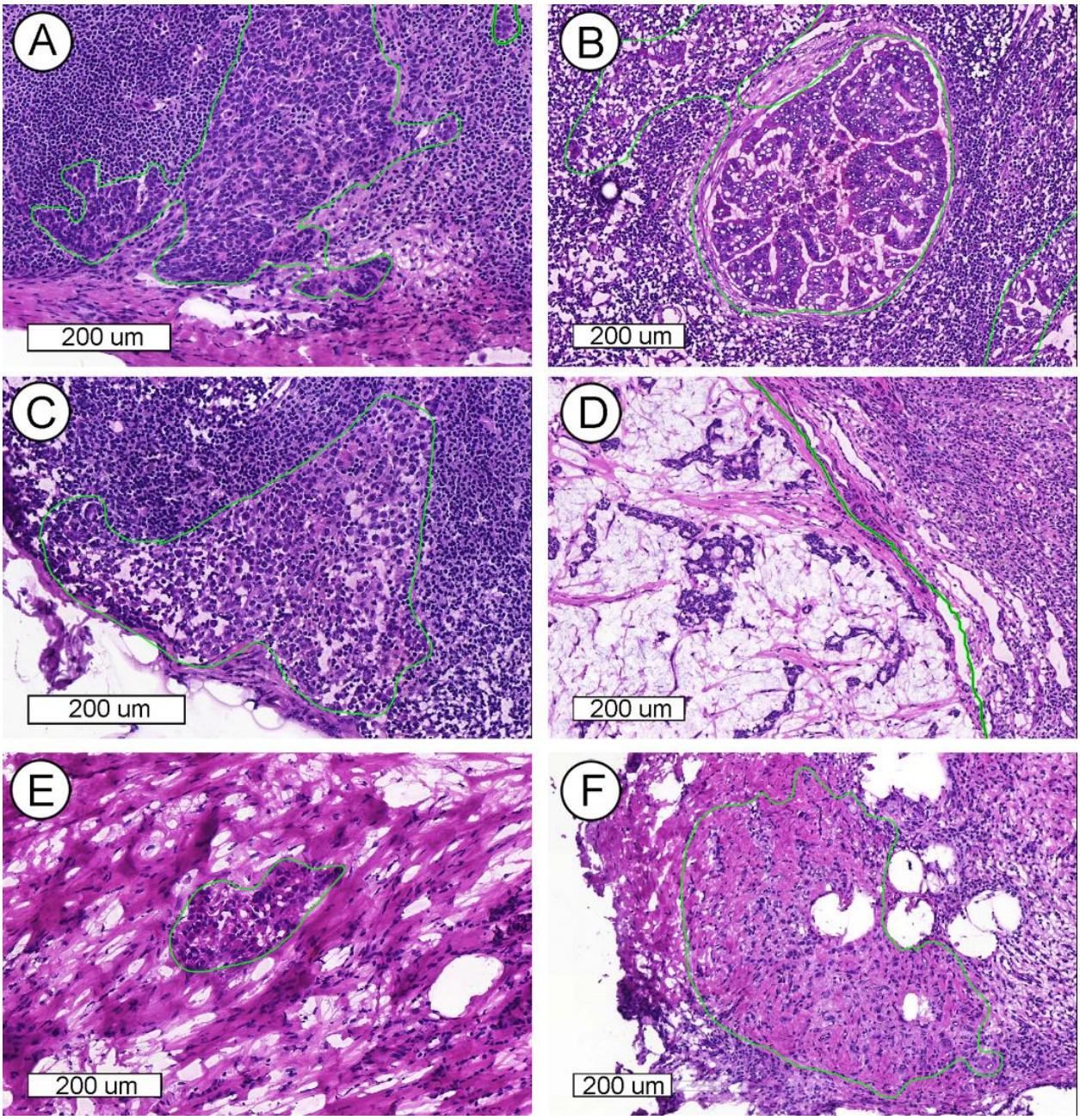



*Figure 4-1. Representative microscopic images of various metastatic carcinomas with annotation. (A) Invasive ductal carcinoma, histologic grade 2, consists of medium-sized tumor cells with moderate glandular formation. (B) Invasive ductal carcinoma, histologic grade 3, shows large-sized tumor cells with poor glandular formation. (C) Tumor cells are small- to medium-sized and poorly cohesive in invasive lobular carcinoma. (D) Mucinous carcinoma*

*contains abundant extracellular mucin. (E) & (F) Invasive ductal carcinoma after neoadjuvant systemic therapy shows fragmented clusters of tumor cells (E) or singly scattered, atypical tumor cells (F) in the fibrotic background (Hematoxylin and eosin).*

4.1.3. Methods

To train the CNN based classification and segmentation model, state of the art architectures, Inception v3[128] (Ours 1) and EfficientNet [131]backbone based feature pyramid network[132] (Ours 2), were selected to compare performance. Patch extraction in training and validation were conducted at the slide level 4. For inferencing the heat maps, confidence vales predicted before the argmax operation were selected for both classification and segmentation model. Gaussian blurring[133] was conducted as a post processing. The probability to determine whether tumor regions are included in the digital slide was selected by the maximum confidence of the heat map.

The algorithms were assessed for classifying between "metastasis slide" or "normal slide". Receiver operating characteristic analysis was performed to measure the AUC (area under the curve).

4.1.4. Challenge environment

The challenge platform developed by Kakao was used to allocate two GPUs to each team. All of the participants were allowed to access only paths of digital slides and corresponding mask images with Kakao platform. Docker image files that enables any of deep learning platform to run were used to train models and inference development and validation sets. Each team was given two GPUs (P40) resources for training models. Kakao platform used CUDA 9,0 and cuDNN 7.

During the first stage for four weeks, participants were given 197 digital slides as the training and development set for six weeks. The training set (157 digital slides) with annotated masks was given for training the model, while the development set (40 digital slides) without

masks was given for tuning the model. Model performance calculated by the evaluation matrix was listed on the leader board after inferencing the development set which was used for tuning the model. During the second stage for two weeks, the participants were given 100 additional digital slides for final evaluation of their models with the optimal model derived from the development set.

#### 4.1.5. Challenge participants

Forty-five participants who were interested in digital pathology or machine learning registered for this challenge within four weeks from the beginning of November 2018. Ten participants were selected according to their inner commitments in accordance with the limited platform environment. Ten participants were composed of students, researchers, and doctors experienced in medical image analysis using machine learning or deep learning. Only five participants submitted their results on the leaderboard. The methodologic description is summarized in Table 4-2. All of the participants selected only deep learning as the main architecture such as Inception v3 and Inception-ResNet [134] for classification of the tumor patch or U-net[135] for segmentation of the tumor region. In one team which ranked high, random forest regression[136] was used to inference confidence by extracting high level features including the number of tumor regions, percentage of the tumor region over the entire tissue region, the area of the largest tumor regions, etc., from the heat map generated using the deep learning method. Detailed descriptions of each algorithm are listed in Table 4-2.

*Table 4-2. Algorithm descriptions and hyper parameters.*

| Team | Architecture | Input patch size (Slide layer level) | Optimization (Learning rate) | Augmentation | Pre-processing | Post-processing; Inference for confidence |
|---|---|---|---|---|---|---|
| Fiffeb | Inception v3, RFC | 256×256×3(6) | SGD (0.9) | Color noise, horizontal flip, random rotation | Otsu thresholding, tumor patch selection (tumor: >90%, non-tumor <20%) | None; RFC output |
| DoAI | U-Net | 512×512×3(0) | SGD (1e-1, decay 0.1 each 2 epochs) | Rotation, horizontal and vertical flip, | None | De-noising for false positive reduction; CNN output |
| Golden Pass | U-Net, Inception v3 | 256×256×3(4) | Adam (1e-3, 5e-4) | Rotation, horizontal and vertical flip, brightness (0.5~1) | Otsu thresholding, tumor patch selection (tumor=100%) | None; Max value for heat-map |
| SOG | Simple CNN | 300×300×3(4) | Adadelta (1e-3) | None | None | None; CNN output |
| Aron Research | Inception-ResNet | 299×299×3(8) | Adam (1e-3, decay 0.1 each 10 epochs) | Rotation, adding noise for saturation, hue, and contrast | While pixel thresholding | Gaussian smoothing; Mean value for heat-map |
| Ours 1 | Inception v3 | 448×448×3(0) | SGD (0.9) | Horizontal and vertical flip, random rotation, random noise, smoothing | Tumor patch selection (tumor: >90%, non-tumor <10%), stain normalization | Gaussian smoothing; Max value for heat-map |
| Ours 2 | Feature pyramid network | 448×448×3(0) | SGD (0.9) | Horizontal and vertical flip, random rotation, noise, rotation, smoothing | Patch selection (40% < tumor < 60%), stain normalization | Gaussian smoothing; Max value for heat-map |

4.1.6. Fine-tuning

To validate the feasibility of using pre-trained model, three types of initial weights such as random initial, the ImageNet, and the CAMELYON were used to train or fine-tune the CNN based classification models with different ratio of dataset (20%, 40%, 60%, 80%, and 100%). We trained CNN based classification models and inferenced heat maps for the same manners same as Our 1 except for the type of initial weights and different number of dataset. Random initial and the ImageNet pre-trained model were offered by Keras, and the CAMELYON pre-trained model was obtained by training a classification based CNN model with 400K patches. A total of 197 WSIs was used to train, validation, and test. All WSIs was split into 6:2:2 per patient.

4.2. Results

Model performances were listed in Figure 4-2 and Table 4-3. Five teams submitted their results on the leader board in phase 1 and phase 2. For the development set which consisted of 40 digital slides, the top three algorithms (Fiffeb, DoAI, and GoldenPass) and ours (Ours 1 and Ours 2) showed 0.945, 0.986, 0.985, 0.986, and 0.985 AUC, while the others showed approximately 0.55 AUCs. For the validation set which consisted of 100 digital slides, the Ours 1 showed the highest AUC 0.917 in the validation set compared with other teams such as the Fiffeb, DoAI, GoldenPass teams, and Ours 2 at AUC 0.805, 0.776, 0.760, and 0.861, respectively. Average times of the first three teams and ours in validation set were 10.8, 0.6, 3.9, 15.7, and 2.7 minutes, respectively.

*Table 4-3. Performance and average time (minute) comparison for classification of tumor slide.*

| Team | AUC | | Time (min.) |
|---|---|---|---|
| | **Development set** | **Validation Set** | |
| Fiffeb | **0.986** | **0.805** | 10.8 |
| DoAI | 0.985 | 0.776 | 0.6 |
| GoldenPass | 0.945 | 0.760 | 3.9 |
| SOG | 0.595 | 0.540 | - |
| Aron Research | 0.525 | 0.470 | - |
| Ours 1 | 0.986 | 0.917 | 15.7 |
| Ours 2 | 0.985 | 0.861 | 2.1 |

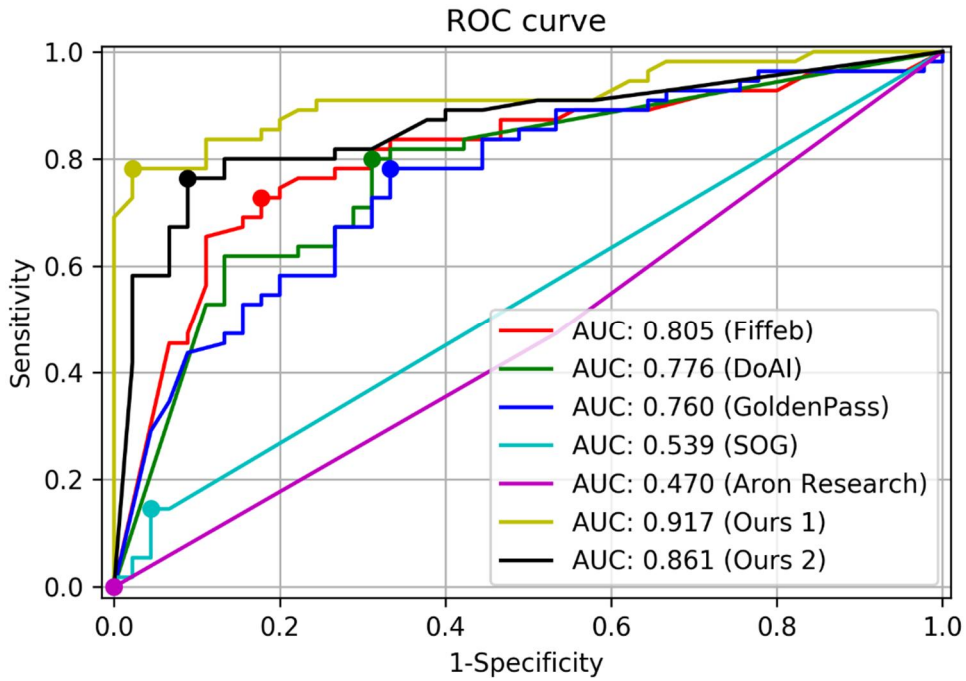



*Figure 4-2. ROC comparisons of models trained by five algorithms for the validation set and cutoff threshold value of each algorithm. The cutoff threshold value is dotted on each ROC curve.*

For more detailed analysis, each algorithm was evaluated with the cutoff threshold determined by the Youden index [137] from the ROC curve in the validation set in terms of the accuracy (ACC).

In addition, model performance comparisons with clinical information for more detail, such as the metastatic tumor size (smaller or larger than 2mm in the greatest dimension), whether patients had received neoadjuvant systemic therapy, histologic type of tumor, and the histologic grade of the tumor was measured, as shown in Table 4-4. Ours 1 and Ours 2 showed the highest 0.900 TPR and the lowest FNR in lymph nodes with larger metastatic tumors. In lymph nodes obtained from patients who had received neoadjuvant systemic therapy, Ours 1 showed the highest TPR (0.862) and TNR (0.962). In terms of the histologic type, all algorithms listed in Table 4-4 showed higher TPR and TNR in the invasive lobular carcinoma

group than in the invasive ductal carcinoma group. When comparing performance between the histologic grades, Fibffeb and DoAI showed higher TPR, but Ours 1 and Ours 2 showed higher TNR in grade 1 or 2 while Ours 1 and Ours2 showed higher TPR and TNR in grade 3.

*Table 4-4. Performance comparison of the first three teams and ours for determining the clinicopathologic characteristics of tumors.*

| Team | Metastatic tumor size | | | | Neo-adjuvant therapy | | | |
|---|---|---|---|---|---|---|---|---|
| | ≤2mm (n=33) | | >2mm (n=22) | | Not received (n=45) | | Received (n=55) | |
| | TPR | FNR | TPR | FNR | TPR | TNR | TPR | TNR |
| Fiffeb | 0.600 | 0.400 | 0.775 | 0.225 | 0.731 | 0.842 | 0.724 | 0.808 |
| DoAI | **0.667** | **0.333** | 0.850 | 0.150 | **0.808** | 0.737 | 0.793 | 0.654 |
| GoldenPass | **0.667** | **0.333** | 0.825 | 0.175 | **0.808** | 0.632 | 0.759 | 0.692 |
| Ours 1 | 0.467 | 0.533 | **0.900** | **0.100** | 0.692 | **1.000** | **0.862** | **0.962** |
| Ours 2 | 0.400 | 0.600 | **0.900** | **0.100** | 0.692 | 0.947 | 0.828 | 0.885 |

| Team | Histologic type | | | Histologic grade | |
|---|---|---|---|---|---|
| | IDC | ILC | MC | 1 or 2 | 3 |

| | (n=86) | | (n=11) | | (n=3) | | (n=86) | | (n=14) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TPR | TNR | TPR | TNR | TPR | TNR | TPR | TNR | TPR | TNR |
| Fiffeb | 0.723 | 0.795 | 0.833 | **1.000** | 0.500 | **1.000** | 0.735 | 0.838 | 0.667 | 0.750 |
| DoAI | **0.766** | 0.667 | **1.000** | 0.800 | **1.000** | **1.000** | **0.816** | 0.676 | 0.667 | 0.750 |
| GoldenPass | **0.766** | 0.641 | **1.000** | 0.800 | 0.500 | **1.000** | 0.796 | 0.649 | 0.667 | 0.750 |
| Ours 1 | **0.766** | **0.974** | **1.000** | **1.000** | 0.500 | **1.000** | 0.776 | **0.973** | **0.833** | **1.000** |
| Ours 2 | 0.745 | 0.897 | 0.833 | **1.000** | **1.000** | **1.000** | 0.755 | 0.919 | 0.833 | 0.875 |

Among the 100 slides in the validation set, 57 slides were correctly categorized by five algorithms including the top three teams, Ours 1, and Ours 2 (35 slides, true positive; 22 slides, true negative), four slides were incorrectly categorized as positive (false positive) by the five algorithms, and six slides were incorrectly categorized as negative (false negative) by the five algorithms, as shown in Figure 4-3. All of the four false positive slides were obtained from patients with invasive ductal carcinoma, histologic grade 2, and two slides were from neoadjuvant systemic therapy patients. Similarly, all of the six false negative slides were obtained from patients with invasive ductal carcinoma, i.e. five from histologic grade 2 patients and one from a histologic grade 3 patient, and three were from neoadjuvant systemic therapy patients. Four of the six false negative slides had micrometastases. The size range of metastatic carcinoma in the false negative slides was 0.13 mm to 4.45 mm.

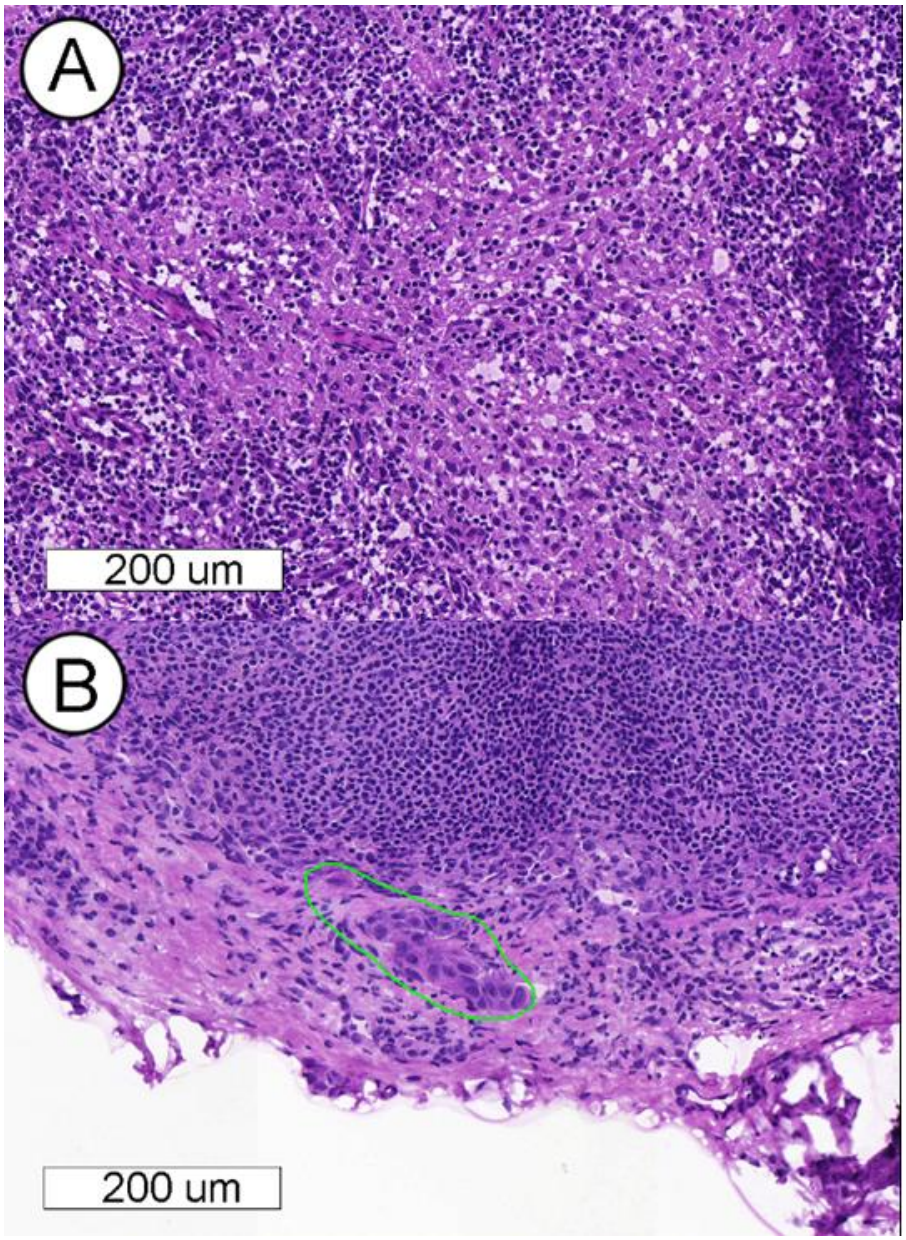



*Figure 4-3. Representative microscopic images of false-positive (a) and false-negative (b) cases. (a) Reactive histiocytes show abundant, eosinophilic cytoplasm and can be misinterpreted as metastatic carcinoma. (b) A very small focus of metastatic carcinoma*

*(approximately 200 μm in the greatest dimension) is seen and which was missed by all five of the teams.*

To validate the performance difference on different level selection for identification of metastasis, additional experiment was conducted by training each model using different level of patches such as level 2, 4, 5, 6, and 7. Models trained with high levels, i.e. 2 or 4, showed higher performance than that of models trained with low levels, i.e. 5, 6, and, 7 as shown in Figure 4-4.

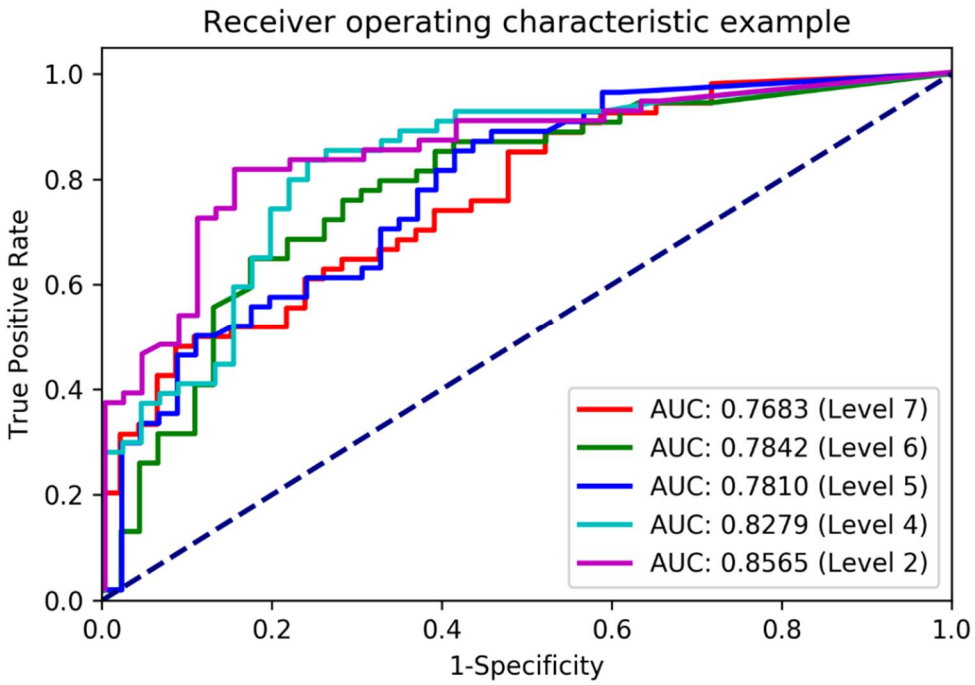



*Figure 4-4. ROC comparisons of models trained with different level of patch.*

To validate the feasibility of using pre-trained model, we trained CNN based classification models in different ways. Random initial weights for full-training and the ImageNet- and the CAMLEYON-based pre-trained weights for fine-tuning were used to train models with

different ratio of dataset (20%, 40%, 60%, 80%, and 100%). Figure 4-5 and 4-6 shows loss comparisons in test dataset of different type of models with different ratio of dataset. Losses of a model based on the initial weight with the 40% of dataset was reduced faster than that of a model with 20% of the dataset as shown in Figure 4-5 (a). As the number of dataset was increased, the minimum losses of models were lower and saturated. It tendency was observed in case of the ImageNet pre-trained model as shown in Figure 4-5 (b). Losses of the models based the initial weight and the ImageNet pre-trained model began at the loss 0.8 while losses of the models based on the CAMELYON pre-trained model began at the loss 0.5 or lower as shown in Figure 4-5 (c). All losses except for that with 20% of the dataset were converged at the same losses for all models as shown in Figure 4-6 (c). AUCs were increased as more dataset are used for models based on all type of initial weights as shown in Table 4-5. In case of using only 20% of the dataset, a model based on the CAMELYON pre-trained model showed higher AUC than that of others while AUCs of others with more than 20% of the dataset were comparable.

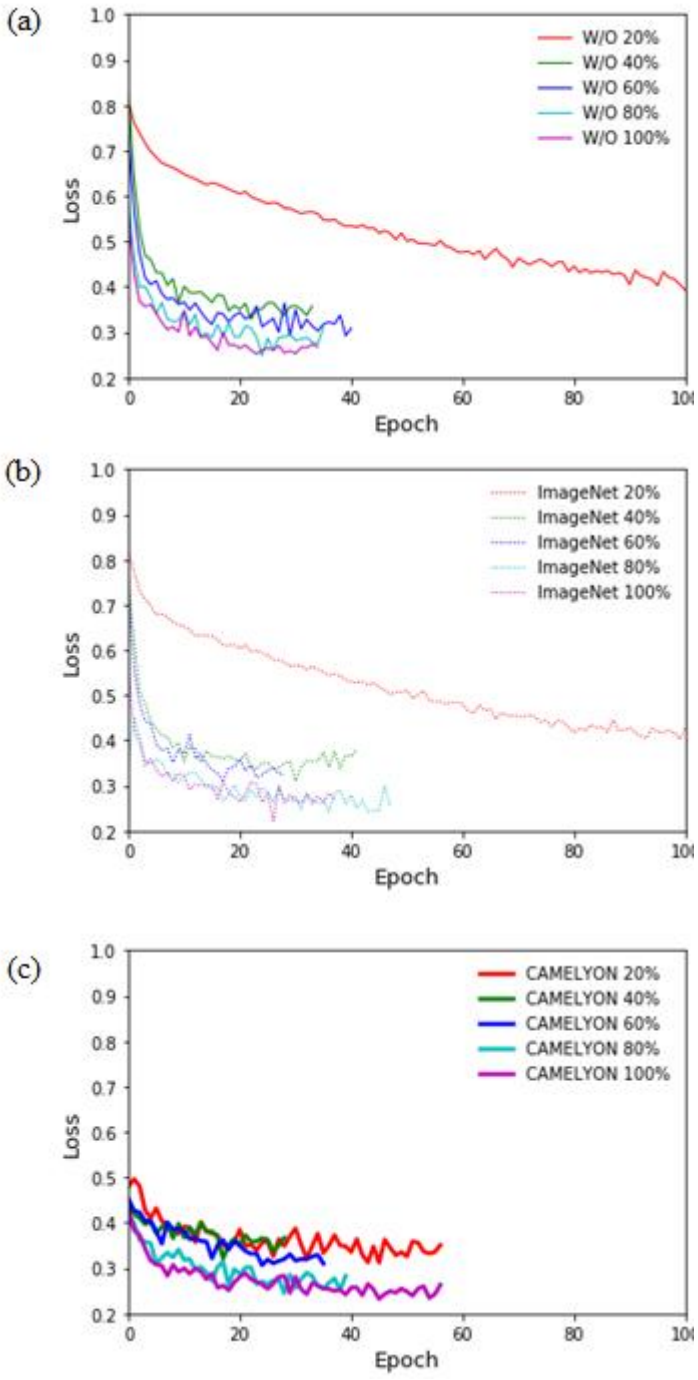



*Figure 4-5. Loss comparisons with different ratio of dataset. CNN based classification models were trained based on (a) initial weight, (b) the ImageNet pre-trained model, (c) the CAMELYON pre-trained model.*

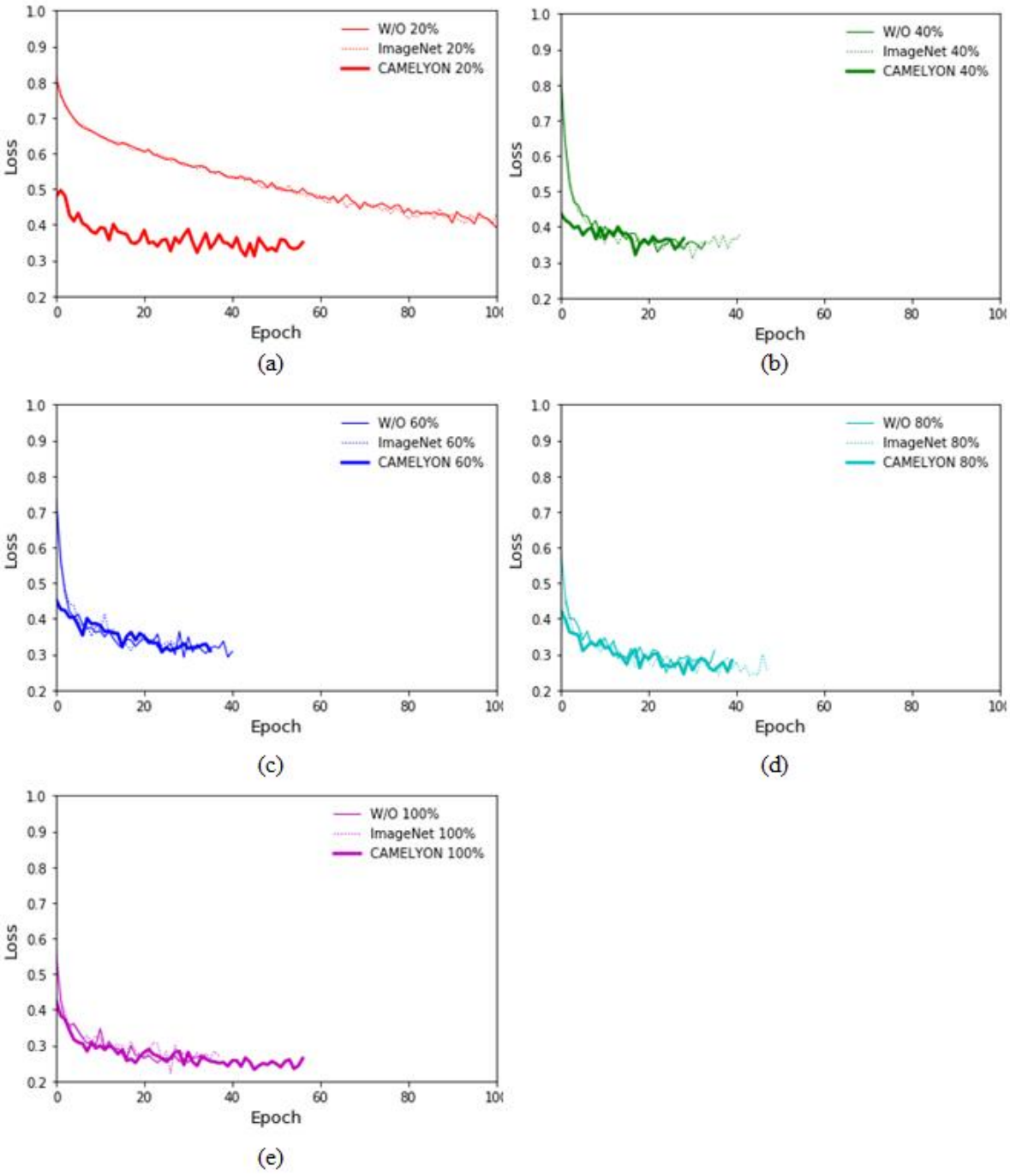



*Figure 4-6. Loss comparisons with different type of models. CNN based classification models were trained with different ratio of dataset such as (a) 20%, (b) 40%, (c) 60%, (d) 80%, (e) 100%.*

*Table 4-5. AUC comparison of models with different initial weights and ratio of dataset.*

| Ratio of dataset (%) | Initial weights | | |
|---|---|---|---|
| | Random | ImageNet pre-trained model | CAMLEYON pre-trained model |
| 20 | 0.8229 | 0.8223 | 0.8422 |
| 40 | 0.8628 | 0.8518 | 0.8639 |
| 60 | 0.8814 | 0.8751 | 0.8771 |
| 80 | 0.8892 | 0.8951 | 0.8867 |
| 100 | 0.9001 | 0.8993 | 0.9036 |

4.3. Discussions

In this current study, all of the participants adopted CNN-based methods as the main idea such as the classification or segmentation network, and which showed high performance at 0.917, 0.861, 0.805, 0.776, and 0.760 in terms of AUC for the Ours 1, Ours 2, and the top three algorithms.

Interestingly, deep learning models trained with high magnification level based patch that was extracted at the level 0 showed higher performance than that of models trained without the patches at level 0 and stain normalization. Moreover, in all algorithms, AUC were lower in the validation set compared to that in the development set. This might be due to the difference in patient demographics, particularly with regard to neoadjuvant systemic therapy. In the validation set, the number of slides obtained from patients after neoadjuvant systemic therapy was significantly higher than that in the development set, as shown in Table 4-1. Neoadjuvant systemic therapy often causes fibrosis and macrophage infiltration in the tumor area and fragmentation and/or scattering of tumor clusters[69], and which can lead to difficulty in histologic examination. It might be suggested that this neoadjuvant systemic therapeutic effect caused a decrease of AUC in the validation set.

Inference time is also key point with this challenge so that methods can be adopted in

routine clinical practice. Two different types of patch-based CNN methods, classification and segmentation network, have shown pros and cons. The number of outputs of the classification network in this challenge is same with the number of classes that the model classifies input patch into (i.e. 1 or 2) by encoding all input dimensions to compressed features for a precise decision. In case of segmentation network, the number of outputs is same with the number of input dimensions (i.e. 448×448=200,704), which is approximately 100K or 100K times more than that of classification network. It is a factor reducing computational time. In our results, the first placed team using only classification network showed 0.3 higher AUC than that of the second placed team using only segmentation network, but too slow to deploy this into the real clinical routine while the computational time of the second placed team took 18.8 times faster than that of the first placed team. Ensemble of those different types of CNN networks should be considered to enhance model performance in routine clinical practice.

Next, we compared model performances according to the clinicopathologic factors of the patients. It is generally known that in manual examination of intraoperative sentinel lymph node biopsy, false negative results are more likely in micrometastases and favorable and/or lobular histology[138]. In the validation set, the top three teams showed better performances in lymph nodes with macrometastatic tumor, and which is consistent with manual examination and the CAMELYON16 study. Lymph nodes which were obtained from non-neoadjuvant systemic therapy patients also revealed better performances, as discussed above. Lymph nodes from invasive ductal carcinoma patients revealed better TPR in all top three teams and better TNR in the top one team than those from invasive lobular carcinoma patients, although the number of slides from invasive lobular carcinoma patients is limited. This is in accordance with the general results in manual examination and the CAMELYON16 study. In the CAMELYON16 study, 29 among 32 teams showed higher AUC in the invasive ductal carcinoma set than in the non-invasive ductal carcinoma set. In addition, tumors of histologic grade 1 or 2 showed higher TPR in the top three teams, but lower TNR in two of the three teams than tumors of histologic grade 3, and which requires further studies.

We found that some cases were wrongly categorized by the first three teams. All of six false

negative cases showed small-sized metastatic carcinoma, and which could result in false negativity. In contrast, four false positive cases did not reveal any common clinicopathologic feature. However, we assume that reactive histiocytic infiltration or prominent germinal centers in lymph nodes might cause false positivity. Manual confirmation is probably necessary, and so a screening tool that would expedite this process might have broad appeal.

In experiments on feasibility of pre-trained model for identification of metastasis, we found that the CAMLEYON pre-trained model for fine-tuning could make the model training faster than that without the CAMELYON pre-trained model. In case of using small dataset (40 WSIs), the losses of the model based on the CAMLEYON pre-trained model were lower than that without the CAMLEYON pre-trained model and it showed higher AUC. In case of using more than 40 WSIs, model performances were comparable to each other, which means that any of pre-trained model was not useful to enhance performance in terms of AUC although the model training is done fast.

Our study has some strong significance compared to previously reported studies about possible usefulness of deep learning algorithm in diagnosis of sentinel lymph node metastasis [44,68]. First, we used digital slides from frozen sections which were made intraoperatively, while previous studies used FFPE sections. Since frozen sections have lower quality due to tissue artifact compared with FFPE sections, it is more difficult to examine frozen sections than FFPE sections. However, what is used to determine the surgical extent intraoperatively in the real world is frozen sections, not FFPE sections. Therefore, we suggest that studies of the deep learning algorithm with sentinel lymph nodes would be more practical if frozen sections are used. Second, our dataset includes a high proportion (48.5%) of post–neoadjuvant patients. The role of neoadjuvant therapy in breast cancer treatment has been increasing these days, but it is much more difficult to histologically diagnose sentinel lymph node metastasis after neoadjuvant therapy[69]. During case selection, we included more post–neoadjuvant cases than clinical setting with an intention of making our dataset unique and more useful. To reduce false positive or false negative issues technically, the deep learning models should be re-trained with those regions and different hyper-parameters such as class weights or loss weights. Those

regions with different hyper-parameters have deep learning models intensively trained as strong positive regions with this strategy. Applications using these methods can be adopted in routine clinical practice by showing attention map with augmented reality and training itself robustly with false positive cases selected by pathologists with on-line learning.

Our contest has several limitations. First, only paths to access the training, development, and validation sets were given to participants, which means that they had no way to check the heat map generated by their models as all dataset contests provided were not available in public. Participants were not allowed to check processing in the middle of training for the same reason. Only less than 1MB log data could be saved and given to participants for the purpose of debugging after training processing to check if and how the training is going well. It was also not available how much time was spent for training and analyses. Second, only 2 GPUs were given to each participant, and it could be limited resource, although this constraint makes participants fair. Third, we did not perform immunohistochemistry to confirm metastatic carcinoma on frozen section slides. On the contrary to FFPE sections, multiple frozen sections which were made from the same tissue fragment showed quite different shapes due to the tissue artifact. Therefore, immunohistochemistry is not as helpful in frozen sections as in FFPE sections to annotate tumor cells. In addition, it is impossible to retrospectively perform immunohistochemistry on frozen sections. Instead, when we annotate tumor cells in frozen sections, we review matched FFPE sections with cytokeratin immunohistochemistry in order to minimize annotation error. Finally, the high proportion of post–neoadjuvant cases or cases with micrometastases could have negatively affected the diagnostic accuracy of algorithms in this study. It would have been nicer if we could divide the dataset into multiple groups and develop different algorithms based on patients' information, such as neoadjuvant status, histologic type, or histologic grade of tumor. However, it was impossible due to the limited number of digital slides. We hope to expand our dataset and include such analysis in our further study.

Possibly because of the characteristics of our dataset and the above limitations, even the top three algorithms in this study showed lower performance than the other first prized in

CAMELYON16, and lower diagnostic accuracy than average of pathologists [139]. However, we believe that it is worth holding a digital pathology challenge using frozen tissue sections. For adjusting algorithms into routine clinical practice, HeLP is preparing another challenge competition to handle other problems such as localization of micro-metastasis and processing time.

Although deep learning performance goes beyond expert physicians, it is not possible to diagnose itself due to its characteristic, i.e. black box. Therefore, so far, it can be used as computer aided system to help doctors diagnose. For example, virtual reality technology can help making quack and accurate decision, or alert a doctor who misses critical parts.

## 5. Discussions

Pathologic diagnosis mainly depends on visual scoring by pathologists, which is susceptible to inter- and/or intra observer variations, i.e. diagnostic discrepancies. Moreover, samples of the biopsy tissues should be closely examined through the microscope. Even experienced pathologists need to zoom-in and out for all over tissue regions while moving, which is time-consuming, as well. To overcome those issues, we tried to adjust the deep learning technique with image processing-based pre-/ and post-processing. Without patch-based approach, each WSI is so huge that the input image cannot be used as input of any of CNN models. To access the all regions of each WSI, patch-based approach was used by splitting all regions into tiled patches at level 0. A bunch of tiled patches from each WSI were used to train CNN based classification, segmentation, and detection models, and WSIs in test set are also tiled into many of patches in order to inference heat map. Although each WSI can be used by resizing them into relatively small size of image, it is hard for CNN models to extract low- and high-level of features due to quite low resolution of input image. All patches with label data to define if it contains tumor or non-tumor per patch level or pixel level were used to train the CNN classification or segmentation models, which is supervised learning approach. Labeling data by pathologists is important to train the model as supervised learning, but time-consuming to get a bunch of labeling data. To overcome this issue, deep leaning-based method can be used to get labeling data efficiently by training the CNN models.

To improve the CNN models, false positives classified, segmented, or detected by the model can be re-used by training the CNN model with the false positive data as the second label. Additionally, pre-trained model trained with other dataset which domain is not quietly different can be used to enhance the model performance when a total of dataset is less, which could not only reduce the training time, but also enhanced performance.

We adjusted our methods into two different tasks; 1) *A fully automated system for prediction of renal allograft rejection*, 2) *Metastases classification in sentinel lymph nodes on frozen tissue section.* In the first task, we proposed a fully automated system to predict renal allograft rejection and two novel methods to enhance deep learning-based detection model. Drawbacks

such as inter- and/or intra-observer variations may be overcome by an automatic method of PTC scoring using two types of trained models (window classification and PTC detection). Classification, detection, and scoring comparisons showed that this method yielded reasonable results when evaluating stained giga-pixel digital slides. Use of this system may be feasible diagnostically in detecting other diseases and conditions. In the second task, we investigated different types of deep learning methods to validate the feasibility of using deep learning methods. Deep learning based classification or segmentation methods might be helpful in the frozen diagnosis of intraoperative, sentinel lymph node biopsy. We held a contest during six weeks to resolve the problem for classification of digital pathology slides with metastases in hematoxylin and eosin–stained frozen tissue sections of SLNs in breast cancer patients. The top three participant teams achieved very high AUCs in the development set while they performed slightly lower AUC in the validation set. Further studies are required in order to increase the accuracy and decrease the time consuming required to apply the deep learning algorithm in the clinical setting.

## 6. Conclusions

The fully automated systems using deep learning on two tasks including prediction of renal allograft rejection and metastases classification in sentinel lymph nodes on frozen tissue section were developed and evaluated. The system of the first task is highly reliable, efficient, and effective, making it applicable to real clinical workflow. The investigation of the system of the second task might be helpful for efficient training, and fast and accurate diagnosis in the frozen diagnosis in intraoperative biopsy.

## 7. Bibliography


1. Bharati MH, Liu JJ, MacGregor JF. Image texture analysis: methods and comparisons. *Chemometrics and intelligent laboratory systems.* 2004;72(1):57-71.
2. Dundar MM, Fung G, Krishnapuram B, Rao RB. Multiple-instance learning algorithms for computer-aided detection. *IEEE Transactions on Biomedical Engineering.* 2008;55(3):1015-1021.
3. Schoepf UJ, Schneider AC, Das M, Wood SA, Cheema JI, Costello P. Pulmonary embolism: computer-aided detection at multidetector row spiral computed tomography. *Journal of thoracic imaging.* 2007;22(4):319-323.
4. Bauer S, Wiest R, Nolte L-P, Reyes M. A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine & Biology.* 2013;58(13):R97.
5. Yoshida H, Näppi J. CAD in CT colonography without and with oral contrast agents: progress and challenges. *Computerized Medical Imaging and Graphics.* 2007;31(4-5):267-284.
6. Summers RM. Improving the accuracy of CTC interpretation: computer-aided detection. *Gastrointestinal Endoscopy Clinics.* 2010;20(2):245-257.
7. Lodwick GS. Computer-aided diagnosis in radiology: A research plan. *Investigative Radiology.* 1966;1(1):72-80.
8. Giger ML. Computerized analysis of images in the detection and diagnosis of breast cancer. Paper presented at: Seminars in Ultrasound, CT and MRI2004.
9. Giger ML, Karssemeijer N, Schnabel JA. Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer. *Annual review of biomedical engineering.* 2013;15:327-357.

10. Rao VM, Levin DC, Parker L, Cavanaugh B, Frangos AJ, Sunshine JH. How widely is computer-aided detection used in screening and diagnostic mammography? *Journal of the American College of Radiology.* 2010;7(10):802-805.
11. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology.* 2001;220(3):781-786.
12. Davatzikos C, Fan Y, Wu X, Shen D, Resnick SM. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiology of aging.* 2008;29(4):514-523.
13. Kim D, Burge J, Lane T, Pearlson GD, Kiehl KA, Calhoun VD. Hybrid ICA–Bayesian network approach reveals distinct effective connectivity differences in schizophrenia. *Neuroimage.* 2008;42(4):1560-1568.
14. Mitchell TM, Shinkareva SV, Carlson A, et al. Predicting human brain activity associated with the meanings of nouns. *science.* 2008;320(5880):1191-1195.
15. Bartels P, Thompson D, Bibbo M, Weber J. Bayesian belief networks in quantitative histopathology. *Analytical and quantitative cytology and histology.* 1992;14(6):459-473.
16. Basavanhally AN, Ganesan S, Agner S, et al. Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. *IEEE Transactions on biomedical engineering.* 2009;57(3):642-653.
17. Petushi S, Garcia FU, Haber MM, Katsinis C, Tozeren A. Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer. *BMC medical imaging.* 2006;6(1):14.
18. Al-Kadi OS. Texture measures combination for improved meningioma classification of histopathological images. *Pattern recognition.* 2010;43(6):2043-2053.

19. Kong J, Cooper L, Kurc T, Brat D, Saltz J. Towards building computerized image analysis framework for nucleus discrimination in microscopy images of diffuse glioma. Paper presented at: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society2011.
20. Thiran J-P, Macq B. Morphological feature extraction for the classification of digital images of cancerous tissues. *IEEE Transactions on biomedical engineering.* 1996;43(10):1011-1020.
21. Gil J, Wu H, Wang BY. Image analysis and morphometry in the diagnosis of breast cancer. *Microscopy research and technique.* 2002;59(2):109-118.
22. Mouroutis T, Roberts SJ, Bharath AA. Robust cell nuclei segmentation using statistical modelling. *Bioimaging.* 1998;6(2):79-91.
23. Gurcan MN, Pan T, Shimada H, Saltz J. Image analysis for neuroblastoma classification: segmentation of cell nuclei. Paper presented at: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society2006.
24. Tutac AE, Racoceanu D, Putti T, Xiong W, Leow W-K, Cretu V. Knowledge-guided semantic indexing of breast cancer histopathology images. Paper presented at: 2008 International Conference on BioMedical Engineering and Informatics2008.
25. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine.* 2015;175(11):1828-1837.
26. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine.* 2007;356(14):1399-1409.

27. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Paper presented at: Advances in neural information processing systems2012.
28. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. Paper presented at: 2009 IEEE conference on computer vision and pattern recognition2009.
29. Farabet C, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence.* 2012;35(8):1915-1929.
30. Tompson JJ, Jain A, LeCun Y, Bregler C. Joint training of a convolutional network and a graphical model for human pose estimation. Paper presented at: Advances in neural information processing systems2014.
31. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2015.
32. Mikolov T, Deoras A, Povey D, Burget L, Černocký J. Strategies for training large scale neural network language models. Paper presented at: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding2011.
33. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine.* 2012;29.
34. Sainath TN, Mohamed A-r, Kingsbury B, Ramabhadran B. Deep convolutional neural networks for LVCSR. Paper presented at: 2013 IEEE international conference on acoustics, speech and signal processing2013.
35. Leung MK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. *Bioinformatics.* 2014;30(12):i121-i129.
36. Xiong HY, Alipanahi B, Lee LJ, et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science.* 2015;347(6218):1254806.

37. Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings. *arXiv preprint arXiv:1406.3676.* 2014.
38. Jean S, Cho K, Memisevic R, Bengio Y. On using very large target vocabulary for neural machine translation. *arXiv preprint arXiv:1412.2007.* 2014.
39. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Paper presented at: Advances in neural information processing systems2014.
40. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama.* 2016;316(22):2402-2410.
41. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115.
42. Williams BJ, Bottoms D, Treanor D. Future-proofing pathology: the case for clinical adoption of digital pathology. *Journal of clinical pathology.* 2017;70(12):1010-1018.
43. Araújo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using convolutional neural networks. *PloS one.* 2017;12(6):e0177544.
44. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama.* 2017;318(22):2199-2210.
45. Golden JA. Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen. *Jama.* 2017;318(22):2184-2186.
46. Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports.* 2018;8.

47. Behrmann J, Etmann C, Boskamp T, Casadonte R, Kriegsmann J, Maaβ P. Deep learning for tumor classification in imaging mass spectrometry. *Bioinformatics.* 2017;34(7):1215-1223.

48. Xu Y, Jia Z, Wang L-B, et al. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC bioinformatics.* 2017;18(1):281.

49. Wang S, Chen A, Yang L, et al. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Scientific reports.* 2018;8(1):10393.

50. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports.* 2016;6:26286.

51. Regele H, Böhmig GA, Habicht A, et al. Capillary deposition of complement split product C4d in renal allografts is associated with basement membrane injury in peritubular and glomerular capillaries: a contribution of humoral immunity to chronic allograft rejection. *Journal of the American Society of Nephrology.* 2002;13(9):2371-2380.

52. Haas M, Loupy A, Lefaucheur C, et al. The Banff 2017 Kidney Meeting Report: Revised diagnostic criteria for chronic active T cell–mediated rejection, antibody-mediated rejection, and prospects for integrative endpoints for next-generation clinical trials. *American Journal of Transplantation.* 2018;18(2):293-307.

53. Brazdziute E, Laurinavicius A. Digital pathology evaluation of complement C4d component deposition in the kidney allograft biopsies is a useful tool to improve reproducibility of the scoring. Paper presented at: Diagnostic pathology2011.

54. Gibson I, Gwinner W, Bröcker V, et al. Peritubular capillaritis in renal allografts: prevalence, scoring system, reproducibility and

clinicopathological correlates. *American Journal of Transplantation.* 2008;8(4):819-825.

55. Mengel M, Chan S, Climenhaga J, et al. Banff initiative for quality assurance in transplantation (BIFQUIT): reproducibility of C4d immunohistochemistry in kidney allografts. *American Journal of Transplantation.* 2013;13(5):1235-1245.

56. Otsu N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics.* 1979;9(1):62-66.

57. Harrison TR, Kasper DL, Fauci AS. *Harrison's Principles of Internal Medicine 19th Ed.* McGraw-Hill AccessMedicine; 2015.

58. Hayes SC, Janda M, Cornish B, Battistutta D, Newman B. Lymphedema after breast cancer: incidence, risk factors, and effect on upper body function. *Journal of clinical oncology.* 2008;26(21):3536-3542.

59. Fleissig A, Fallowfield LJ, Langridge CI, et al. Post-operative arm morbidity and quality of life. Results of the ALMANAC randomised trial comparing sentinel node biopsy with standard axillary treatment in the management of patients with early breast cancer. *Breast cancer research and treatment.* 2006;95(3):279-293.

60. Lyman GH, Temin S, Edge SB, et al. Sentinel lymph node biopsy for patients with early-stage breast cancer: American Society of Clinical Oncology clinical practice guideline update. *J Clin Oncol.* 2014;32(13):1365-1383.

61. Manca G, Rubello D, Tardelli E, et al. Sentinel lymph node biopsy in breast cancer: indications, contraindications, and controversies. *Clinical nuclear medicine.* 2016;41(2):126-133.

62. Galimberti V, Cole BF, Zurrida S, et al. Axillary dissection versus no axillary dissection in patients with sentinel-node micrometastases (IBCSG 23–01): a phase 3 randomised controlled trial. *The lancet oncology.* 2013;14(4):297-305.

63. Giuliano AE, Ballman KV, McCall L, et al. Effect of axillary dissection vs no axillary dissection on 10-year overall survival among women with invasive breast cancer and sentinel node metastasis: the ACOSOG Z0011 (Alliance) randomized clinical trial. *Jama.* 2017;318(10):918-926.

64. Wang J, Tang H, Li X, et al. Is surgical axillary staging necessary in women with T1 breast cancer who are treated with breast-conserving therapy? *Cancer Communications.* 2019;39(1):25.

65. Donker M, van Tienhoven G, Straver ME, et al. Radiotherapy or surgery of the axilla after a positive sentinel node in breast cancer (EORTC 10981-22023 AMAROS): a randomised, multicentre, open-label, phase 3 non-inferiority trial. *The lancet oncology.* 2014;15(12):1303-1310.

66. Celebioglu F, Sylvan M, Perbeck L, Bergkvist L, Frisell J. Intraoperative sentinel lymph node examination by frozen section, immunohistochemistry and imprint cytology during breast surgery–a prospective study. *European journal of cancer.* 2006;42(5):617-620.

67. Chen Y, Anderson KR, Xu J, Goldsmith JD, Heher YK. Frozen-Section Checklist Implementation Improves Quality and Patient Safety. *American journal of clinical pathology.* 2019;151(6):607-612.

68. Bandi P, Geessink O, Manson Q, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE transactions on medical imaging.* 2018;38(2):550-560.

69. Honkoop AH, Pinedo HM, De Jong JS, et al. Effects of chemotherapy on pathologic and biologic characteristics of locally advanced breast cancer. *American journal of clinical pathology.* 1997;107(2):211-218.

70. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification.

Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition2016.

71. Babaie M, Kalra S, Sriram A, et al. Classification and retrieval of digital pathology scans: A new dataset. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops2017.
72. Bizzego A, Bussola N, Chierici M, et al. Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. *PLoS computational biology.* 2019;15(3):e1006269.
73. Vu QD, Graham S, Kurc T, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology.* 2019;7.
74. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics.* 2016;7.
75. Bishop CM. *Pattern recognition and machine learning.* springer; 2006.
76. Michie D, Spiegelhalter DJ, Taylor C. Machine learning. *Neural and Statistical Classification.* 1994;13.
77. Magerman DM. Statistical decision-tree models for parsing. Paper presented at: Proceedings of the 33rd annual meeting on Association for Computational Linguistics1995.
78. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics.* 1991;21(3):660-674.
79. Cortes C, Vapnik V. Support-vector networks. *Machine learning.* 1995;20(3):273-297.
80. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. *Neural processing letters.* 1999;9(3):293-300.

81. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. *Machine learning.* 1997;29(2-3):131-163.
82. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics.* 2001;17(8):754-755.
83. Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: A tutorial. *Computer.* 1996;29(3):31-44.
84. Yao X. Evolving artificial neural networks. *Proceedings of the IEEE.* 1999;87(9):1423-1447.
85. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2016.
86. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology.* 2018;290(1):218-228.
87. Dunnmon JA, Yi D, Langlotz CP, Ré C, Rubin DL, Lungren MP. Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology.* 2018;290(2):537-544.
88. Cicero M, Bilbily A, Colak E, et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Investigative radiology.* 2017;52(5):281-287.
89. Islam MT, Aowal MA, Minhaz AT, Ashraf K. Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850.* 2017.
90. Shiraishi J, Li Q, Suzuki K, Engelmann R, Doi K. Computer-aided diagnostic scheme for the detection of lung nodules on chest radiographs: Localized search method based on anatomical classification. *Medical Physics.* 2006;33(7Part1):2642-2653.

91. Chaya Devi S, Satya Savithri T. On Segmentation of Nodules from Posterior and Anterior Chest Radiographs. *International journal of biomedical imaging.* 2018;2018.
92. Pesce E, Withey SJ, Ypsilantis P-P, Bakewell R, Goh V, Montana G. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. *Medical image analysis.* 2019;53:26-38.
93. de Hoop B, De Boo DW, Gietema HA, et al. Computer-aided detection of lung cancer on chest radiographs: effect on observer performance. *Radiology.* 2010;257(2):532-540.
94. Kumar D, Wong A, Clausi DA. Lung nodule classification using deep features in CT images. Paper presented at: 2015 12th Conference on Computer and Robot Vision2015.
95. Hua K-L, Hsu C-H, Hidayati SC, Cheng W-H, Chen Y-J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy.* 2015;8.
96. Setio AAA, Ciompi F, Litjens G, et al. Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. *IEEE transactions on medical imaging.* 2016;35(5):1160-1169.
97. Jaeger S, Karargyris A, Candemir S, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging.* 2013;33(2):233-245.
98. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology.* 2017;284(2):574-582.
99. Kalinovsky A, Kovalev V. Lung image Ssgmentation using deep learning methods and convolutional neural networks. 2016.
100. Ngo TA, Carneiro G. Lung segmentation in chest radiographs using distance regularized level set and deep-structured learning and inference. Paper

presented at: 2015 IEEE International Conference on Image Processing (ICIP)2015.

101. Cernazanu-Glavan C, Holban S. Segmentation of bone structure in X-ray images using convolutional neural network. *Adv. Electr. Comput. Eng.* 2013;13(1):87-94.

102. Middleton I, Damper RI. Segmentation of magnetic resonance images using a combination of neural networks and active contour models. *Medical engineering & physics.* 2004;26(1):71-86.

103. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE transactions on medical imaging.* 2016;35(5):1240-1251.

104. Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJ, Išgum I. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE transactions on medical imaging.* 2016;35(5):1252-1261.

105. Prasoon A, Petersen K, Igel C, Lauze F, Dam E, Nielsen M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. Paper presented at: International conference on medical image computing and computer-assisted intervention2013.

106. Geras KJ, Wolfson S, Shen Y, et al. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047.* 2017.

107. Dhungel N, Carneiro G, Bradley AP. Automated mass detection in mammograms using cascaded deep learning and random forests. Paper presented at: 2015 international conference on digital image computing: techniques and applications (DICTA)2015.

108. Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Scientific reports.* 2016;6:27327.

109. Kooi T, Litjens G, Van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis.* 2017;35:303-312.
110. Schaumberg AJ, Rubin MA, Fuchs TJ. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. *BioRxiv.* 2018:064279.
111. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718.* 2016.
112. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing.* 2016;191:214-223.
113. McNutt M. Journals unite for reproducibility. American Association for the Advancement of Science; 2014.
114. Litjens G, Bandi P, Ehteshami Bejnordi B, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience.* 2018;7(6):giy065.
115. Gupta S, Mazumdar SG. Sobel edge detection algorithm. *International journal of computer science and management Research.* 2013;2(2):1578-1583.
116. Perona P, Malik J. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on pattern analysis and machine intelligence.* 1990;12(7):629-639.
117. Van Dokkum PG. Cosmic-ray rejection by Laplacian edge detection. *Publications of the Astronomical Society of the Pacific.* 2001;113(789):1420.
118. Gustafsson H, Claesson I, Nordholm S. Signal noise reduction by spectral subtraction using linear convolution and casual filtering. Google Patents; 2001.

119. Tai S-C, Yang S-M. A fast method for image noise estimation using laplacian operator and adaptive edge detection. Paper presented at: 2008 3rd International Symposium on Communications, Control and Signal Processing2008.
120. Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2016.
121. Kim J, Kwon Lee J, Mu Lee K. Deeply-recursive convolutional network for image super-resolution. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2016.
122. Graham B. Fractional max-pooling. *arXiv preprint arXiv:1412.6071.* 2014.
123. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer.* 2005;27(2):83-85.
124. Skodras A, Christopoulos C, Ebrahimi T. The jpeg 2000 still image compression standard. *IEEE Signal processing magazine.* 2001;18(5):36-58.
125. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics.* 2013;4.
126. Ciompi F, Geessink O, Bejnordi BE, et al. The importance of stain normalization in colorectal tissue classification with convolutional networks. Paper presented at: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)2017.
127. Racusen LC, Halloran PF, Solez K. Banff 2003 meeting report: new diagnostic insights and standards. *American journal of transplantation.* 2004;4(10):1562-1566.

128. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2016.
129. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2017.
130. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2016.
131. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv preprint arXiv:1905.11946.* 2019.
132. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2017.
133. Wink AM, Roerdink JB. Denoising functional MR images: a comparison of wavelet denoising and Gaussian smoothing. *IEEE transactions on medical imaging.* 2004;23(3):374-387.
134. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. Paper presented at: Thirty-First AAAI Conference on Artificial Intelligence2017.
135. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Paper presented at: International Conference on Medical image computing and computer-assisted intervention2015.
136. Liaw A, Wiener M. Classification and regression by randomForest. *R news.* 2002;2(3):18-22.
137. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3(1):32-35.

138. Akay CL, Albarracin C, Torstenson T, et al. Factors impacting the accuracy of intra-operative evaluation of sentinel lymph nodes in breast cancer. *The breast journal.* 2018;24(1):28-34.

139. Houpu Y, Fei X, Yang Y, et al. Use of Memorial Sloan Kettering Cancer Center nomogram to guide intraoperative sentinel lymph node frozen sections in patients with early breast cancer. *Journal of surgical oncology.* 2019;120(4):587-592.

Abstract (In Korean)

병리학은 질병을 최종진단하는 학문이며, 이를 위해 병리사들이 생검 조직을 전자 현미경을 사용하여 면밀히 봐야 한다. 이러한 진단을 위해 숙련된 병리사들도 현미경에 보이는 영상에 대해 확대, 축소를 반복적으로 하면서 모든 영역을 살펴봐야만 한다. 게다가 병리적 진단에 있어서 많은 시간을 필요로 하는 노동과 관찰자 내, 외 오차로 인해 불확실 할 수 있으며 병리사들의 시각적인 평가에 의존적일 수 밖에 없다. 수술 중 얻어지는 생검 조직인 경우 정확하면서도 빠른 결정이 필요하기도 하다. 이러한 문제들을 극복하기 위해, 두가지 주제에 대해 딥러닝을 활용한 빠르고 정확한 병리진단을 하는 모델에 관한 연구 방법론을 제공하고자 한다. 1) 스캔된 동결 절편 조직 슬라이드에서의 림프 노드에서의 암 전이 여부를 분류한다. 2) 스캔된 포르말린으로 고정된 조직 슬라이드에서의 신이식 거부반을 예측한다.

첫번째로, 두 종류의 합성곱 신경망 기반 알고리즘을 활용한 전자동 시스템을 제안한다. 제안하는 전자동 시스템은 두 부분으로 나뉜다. 동결 절편 조직 슬라이드에서 1) 관심영역을 분류하는 부분과 2) C4d 염색된 고리주변의 모세혈관 (PTC)과 C4d 염색되지 않은 PTC를 검출하는 부분이다. 표지 영역 크기를 늘리는 방법에 대한 최적의 변수를 구하기 위해 염색된 PTC와 염색되지 않은 PTC의 표지된 영역의 크기를 다양하게 실험하여 검출 모델의 성능을 평가하였다. 표지 영역의 크기는 염색된 PTC와 염색되지 않은 PTC에 대해 각각 50, 40픽셀을 늘림으로써 최상의 검출 성능을 보였다. 또한, 탐지 모델의 성능을 높히기 위한 효과적인 데이터 수집을 위해 딥러닝 기반 표지 툴을 활용한 표지 데이터 사용의 적합성 여부를 검증하였다. 이 전자동 시스템은  신이식 거부반응

예측 시스템에 있어서 신뢰성이 높고 효율적이어서 실제 임상 작업흐름에 적용할 수 있었다.

두번째 주제를 위해, 암 전이유무를 자동 분류하는 동일한 목적을 위한 서로 다른 CNN 방법론 (영상 분류, 물체 분할)에 대한 실험을 했으며 2018 HeLP challenge에 참여한 팀 결과들과의 비교를 통해 방법론들간의 특성을 분석하였다. CNN 기반 분류 방식들이 분할 방식들보다 높은 AUC를 보인 반면 시간 측면에서 분류 방식들은 분할 방식들보다 5배 느리다는 분명한 각각의 장단점을 결과를 통해 나타났다. 또한, 공개 CAMLEYON 공개 데이터셋으로 미리 학습한 선행 모델을 활용하여 모델의 초기값을 주게 되면 80개 이하의 적은 디지털 슬라이드 데이터셋으로 분석을 해야하는 경우, 더 낮은 손실값을 통한 더 정확한 모델로 학습시킬 수 있었다. 하지만 80개 이상의 디지털 슬라이드 데이터셋으로 충분한 데이터가 주어진 경우에는 손실값과 정확도에 있어서 큰 차이를 보이지 않았다.

동결 절편 조직 분석 분석에 있어서 딥러닝을 활용한 두 방법론에 있어서 빠르고 정확한 진단에 도움이 되는 방법들을 분석하였다. 또한, 딥러닝을 활용한 전자동 신이식 거부 반응 예측 세스템은 병리 진단에 있어서 신뢰성이 높고 효율적이며 실제 임상 현장에 적용될 수 있음을 검증하였다.

**중심 단어**: 딥러닝, 디지털 병리, 신장이식 거부반응, 암 전이, 영상 분류, 물체 검출, 물체 분할