UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

REGRESSION AND CLASSIFICATION OF BREAST CANCER DEPICTING ON DIGITAL PATHOLOGY IMAGES USING CONVOLUTION NEURAL NETWORKS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

LINSHENG HE
Norman, Oklahoma
2019

REGRESSION AND CLASSIFICATION OF BREAST CANCER DEPICTING ON DIGITAL PATHOLOGY IMAGES USING CONVOLUTION NEURAL NETWORKS

A THESIS APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY

Dr. Bin Zheng, Chair

Dr. Hong Liu

Dr. Liangzhong Xiang

To my parents and Zhitao Luo,

for your unselfish support, wherever you are.

Acknowledgements

I would like to thank my graduate academic professor, Dr. Zheng. Without him, I cannot finish my master's degree in two years successfully. My thesis would have been unachievable without his support, guidance, and suggestions. Professor Zheng's gentle personality and rigorous academic attitude have benefited me a lot.

I would like to thank my colleagues in our research lab, Yunzhi, Gopi, Nafiseh, Morteza, and Faranak. It's my fortunate to study with them, I have learned a lot from them.

I would like to thank Zhitao Luo, thanks for your encouragement and companionship in the past two years, my life is more wonderful because of you.

I would like to express my deepest gratitude to my parents and family, thanks for your flawless support and love.

Data used in this research was acquired from Sunnybrook Health Sciences Centre with funding from the Canadian Cancer Society and was made available for the BreastPathQ challenge, sponsored by the SPIE, NCI/NIH, AAPM, Sunnybrook Research Institute.

Table of Contents

List of Tables

List of Figures

Abstract

Computer-Aided Detection and Diagnosis (CAD) of medical images has been developed and tested for the last three decades. It is designed to be an effective time-saving assistant and provide doctors with a wealth of diagnostic information in clinical practice. Traditional CAD techniques utilize the features of manual extraction of images and algorithms that use shallow supervised learning. They have great limitations in medical image classification, effective feature extraction, and segmentation. In order to overcome these limitations, deep learning has emerged as a promising technology in CAD development. In this study, my motivation is to investigate an optimal approach of applying Convolutional Neural Network (CNN), one of the deep learning models, to assist detection of residual breast cancer using histopathology images after Neoadjuvant treatment (NAT). In this process, CNN models were implemented and compared using different optimization objectives and evaluation metrics.

Specifically, the CAD system used two CNN schemes, namely, ResNet and SE-ResNeXt. To detect the residual cancer cells from the pathology images after NAT, a regression CAD system was developed to predict cancer cellularity value followed by applying a concordance evaluation metric to compare and verify the effectiveness of the model. The performance of integrated models for predicting cancer cellularity and providing doctors with second information is almost the same as the performance of the supplementary information given by other doctors.

To distinguish classes of cancer cellularity in the pathology images, a classification CAD system was developed to predict the probability of classes. The classier performance was evaluated using the Receiver Operating Characteristic (ROC)

method. When using a 5-fold cross-validation method to classify the classes of pathology images, the best area under the ROC curve was 0.905 ± 0.075. The results of study indicated that CNN based deep learning is a promising CAD technology, which can significantly improve the diagnostic efficacy of detecting residual breast cancer cells in pathology images after NAT with high performance.

Chapter 1: Introduction

1.1 Breast cancer and screening using radiographic imaging

Breast cancer, which refers to a malignant tumor, is an uncontrolled growth of breast cells. A tumor can be benign (not dangerous to health) or malignant (has the potential to be dangerous). Benign tumors are not considered cancerous. The cells in benign tumors are close to normal in appearance. They grow slowly and do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous because malignant cells can grow rapidly and eventually spread beyond the original tumor to other parts of the body [1]. Thus, it is important to detect and treat malignant tumors in the early stage.

Most breast cancers are a type of carcinoma called adenocarcinoma, which starts in cells that make up glandular tissue. The most common symptom of breast cancer is a new lump or mass. A painless, hard mass that has irregular edges is more likely to be cancer, but breast cancers can be tender, soft, or rounded. They can even be painful [2].

For women in the United States, breast cancer death rates are higher than those for any other cancer, besides lung cancer. About 1 in 8 U.S. women (about 12.4%) will develop invasive breast cancer over the course of her lifetime. About 41,760 women in the U.S. are expected to die in 2019 from breast cancer, though death rates have been decreasing since 1989 [3]. These decreases are thought to be the result of treatment advances, earlier detection through screening, and increased awareness.

Imaging tests using the modalities, such as Mammography, Magnetic Resonance Imaging (MRI) and Ultrasound, are clinically accepted in aiming to detect and analyze the tumor which is still not visible from the epidermis. However, imaging tests have limits;

no imaging test can show a single cancer cell or even a few. Other than this, it is a quite difficult task for the certified radiologists in breast imaging to interpret the mammogram with high detection sensitivity and specificity. In fact, especially in the breast area, the dense fibro-glandular tissues overlapping as well as the large heterogeneity of breast lesions will produce different diagnostic results from different radiologists due to the high inter-reader variability [4]. Thus, in the breast cancer screening environment, high recall rates of radiologists will result in a large number of biopsies to determine whether the tumor is malignant or benign based on histopathology test and analysis, which is the best and gold-standard method in the clinical practice.

1.2 Breast cancer diagnosis in pathology

Pathology has always been regarded as a "bridge discipline" between basic medicine and clinical medicine, which fully demonstrates its irreplaceably important role in medicine, which is determined by the nature and tasks of pathology. Specifically, pathology lays a scientific theoretical foundation for mastering the nature, diagnosis, treatment, and prevention of diseases and directly participates in clinical diagnosis and treatment. As a field of general inquiry and research, the four major tasks of pathology include cause (etiology), mechanisms of development (pathogenesis), structural alterations of cells (morphologic changes), and the consequences of changes (clinical manifestations) [5].

The methods of pathological test are diverse, such as biopsy, blood analysis, dissection, and other applications of medical microscopy. When conducting pathological diagnosis of breast cancer, the pathologists collect the tissue samples extracted by biopsy,

which include the use of different methods of local excision, clamping, needle aspiration, and scraping, removal, etc. to acquire samples of the patient's diseased tissue.

With the development of natural science, medical science has gradually formed many sub-disciplines, and their common purpose and tasks are through studying the life activities of normal and diseased organisms from different fields and in different ways, to prevent diseases and to protect human health. Therefore, pathology is closely related to anatomy, histology, embryology, physiology, biochemistry, parasitology, microbiology, etc. in basic medicine.

1.2.1 Histopathology

Histological observation is one of the most commonly used means of observing and studying diseases. By making the diseased tissue into slices several micrometers thick, staining them with different materials, and then observing the microscopic lesions with a microscope, the resolution of the naked eye observation is improved hundreds of times, and the understanding of diseases and lesions is deepened. The pathologists usually utilize histopathology which is combining histological observation and pathology to diagnose diseased tissue after biopsy in breast cancer.

Histopathology contrasts with cytopathological methods which use free cells or tissue fragments and aid in the diagnosis of certain infectious diseases. Histopathological examination of the tissue begins with surgery, biopsy or autopsy [6]. Once the targeted breast tissue region is biopsied, the samples taken are studied manually under the microscope by a pathologist. And typically, the tissue is removed from the organism and placed in a fixative to stabilize the tissue to prevent decay.

In the entire field of diagnostic histopathology, the role of pathologists in tumor diagnosis is unparalleled. Unfortunately, patients or non-professionals know nothing about this and often take it for granted that their diagnostician is a surgeon, clinician or oncologist. In fact, for any patient with swelling or mass, a histopathological report is the primary factor in diagnosing, predicting clinical progression, and determining treatment. However, many patients are reluctant to carry out the histopathology test because to take biopsy is an invasive examination. Therefore, histopathology is often performed after the image test and surgery.

Breast histopathology can be used not only to detect the benign and malignant tumors after surgery, but can also be used to diagnosis the stage of breast cancer, follow up chemotherapy or radiation therapy to prevent cell proliferation, and identify whether cancer cells have been completely eliminated.

1.2.2 Neoadjuvant treatment and cancer cellularity

Neoadjuvant treatment (NAT) of breast cancer is an option for patients with locally advanced disease. In addition to the treatment's effect on tumor size, NAT may alter the cancer cellularity. Tumor response to the therapy provides useful information for patient management and can guide decisions about subsequent therapy. Tumor size may not decrease, but the overall cellularity may be markedly reduced, making residual cancer cellularity an important factor in assessing response.

Cellularity within the tumor bed is defined as the percentage area of the overall tumor bed that is comprised of tumor cells (invasive or in situ). The histopathological examination of the tissue sections after surgery to estimate the residual tumor and the

assessment of cellularity is an important component of tumor burden assessment. It's a very helpful feedback for NAT of breast cancer.

In the current clinical practice, cancer cellularity is manually estimated by pathologists on haematoxylin and eosin (H&E) stained slides, the quality and reliability of which might be impaired by inter-observer variability which potentially affects prognostic power assessment in NAT trials [7, 8]. Although this procedure is also qualitative and time-consuming in the current practice, cellularity fraction of cancer is a better prognostic indicator to illustrate the therapy results without segmenting the bound of the cancer cell or indexing the cell's characteristic individually [9].

1.3 Digital pathology

Whenever using optical microscopes or electron microscopes, pathologists who diagnose pathology images need to operate microscopes manually. The data size of pathology images is huge when compared to radiology images, in order to save time costs, pathologists often use their experience to determine suspected areas for diagnosis and observation, rather than for the whole image. Therefore, reading such a big size of images by pathologists is a tedious and high error rate job. However, with the rapid development of the information age in these decades, the emergence of digital pathology can effectively revolute this situation. Digital pathology images are obtained through medical instruments, such as Aperio Digital Pathology Slide Scanner, a commercialized available whole slide scanner to produce digital images, which can display whole digital slides in computer monitors in less than 60 seconds. It also can change the magnification manually in computers which is the same as observing the glass slide under a conventional microscope.

Meanwhile, relatively objective diagnostic reports often need to be diagnosed by more than one pathologist. Because different pathologists often have different diagnosis results, we call it inter-reader variability. If two pathologists are not in the same laboratory, it is difficult for them to give a diagnosis of the pathological image under a conventional microscope at the same time. Along with the development of digital pathology, images can be transmitted in the network without being constrained by space. Also, digital pathology can assist telemedicine (or tele-pathology) to diagnosis jointly and even to better serve remote areas without or with few highly experienced pathologists.

Digital pathology is a disruptive technology, as technology becomes more cost-effective, digital pathology is becoming more common. Digital pathology will undoubtedly allow pathologists to make more accurate and consistent diagnoses in the near future.

1.4. Computer-aided detection and diagnosis

Once a digital image has been acquired, Computer-aided Detection and Diagnosis (CAD) system can be leveraged to analyze the information they hold. Over the past decade, dramatic increases in computational power and improvement in image analysis algorithms have allowed the development of powerful CAD approaches to biomedical image data. Just as with digital radiology over two decades ago, digitized histopathology has now become amenable to the application of computerized image analysis and machine learning techniques for an accurate diagnosis.

For instance, a Whole Slide Images (WSI) printed at 600 dpi could fill 70 dull 8.5" × 11" pages, to predict the cellularity of every 512 × 512 patches at a rate of 5s/patch. Using the CAD method, global scanning of the WSI and giving the diagnosis for each

patch only takes less than five minutes, but when labeling by pathologists, they may need 12 hours or more. A WSI figure is shown in Figure 1.

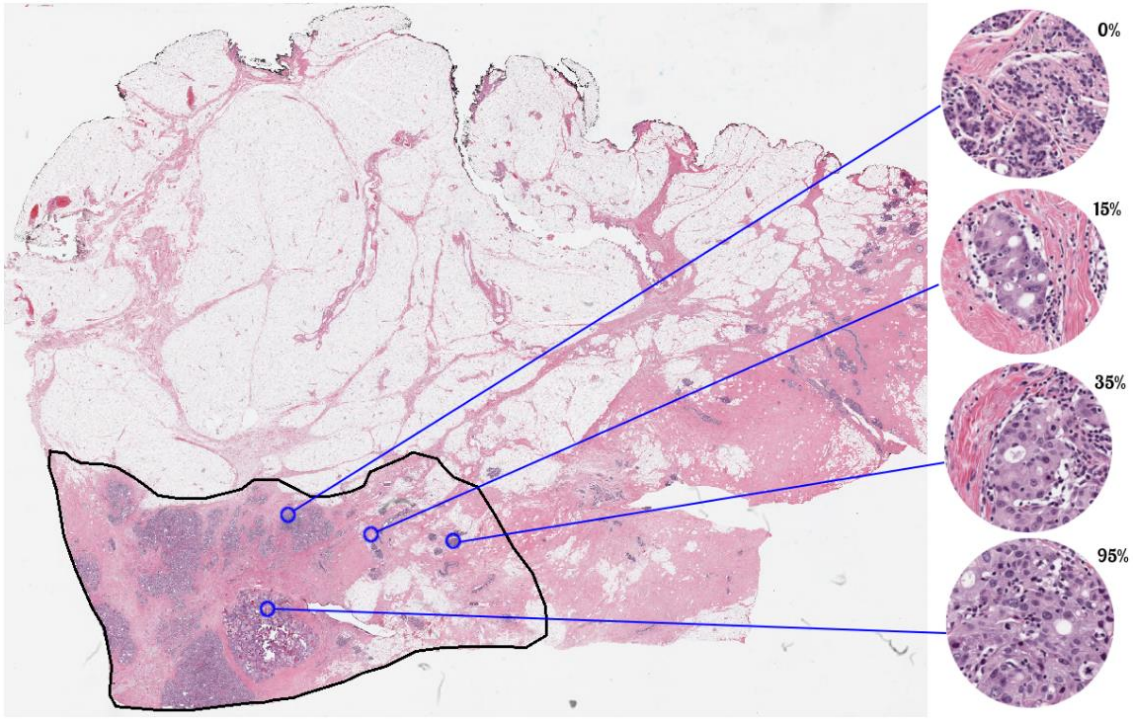


Figure 1. The WSI and four examples in different cancer cellularity.

CAD as a second opinion system includes prognosis diagnosis, evaluation, and assessment. Different pathologists have different results in each slice. Only with improved consistency, will the prognostic value and outcomes be increased. So, we can build a CAD system to minimize this heavy work. The CAD system can pay attention to the slices on which pathologists have big differences in opinion from the result of CAD system and also give a second opinion to pathologists who have different diagnosis results. CAD systems reduce the workload of specialists and cost, contributing to diagnosis efficiency and effectiveness.

1.4.1 Research status

There has already been researching about an analysis of cell and microscopy images in 1965, a very long time ago. Mendelsohn et al. initially demonstrated the morphological analysis of cells and chromosomes [10]. Over the last decade, clinical diagnosis and research were interested in digital pathology, a progressive technology with the rapid development of the Internet. Besides digital pathology, a large digital repository of tissue slides for medical students and pathologists is a huge educational resource. Using CAD tools to address specifically targeted biomarkers or segment and classify the cells in pathology were advanced in recent years as well. In this study, I am going to pay more attention to CAD field in the Artificial Neural Networks (ANN).

Compared to traditional methods, ANN methods do not require hand-crafted functional design and feature extraction, however, they also can scale well to large data sets, and can be easily applied to other applications. ANN-based methods, especially those based on Convolutional Neural Networks (CNN), have received much attention in the field of histopathology images analysis because they have better performance in some applications than traditional methods.

Wang D. et al. created a 27-layer deep network correlated with pathologists to identify metastatic breast cancer [11]. Han Z. et al. achieved a BiCNN model which combined some traditional feature descriptors, such as PFTAS and GLCM, to automated classify the images' property [12]. Bayramoglu N. et al. developed a system which uses conditional generative adversarial networks to virtually stain the unstained histopathology [13]. Sari CT. et al. produced a model to extract features in histopathological classification field [14]. The research [13, 14] above created

unsupervised ANNs to analyze features automatically. Motlagh NH. et al. used a simple ResNet CNN model to classify binary benign-malignant assessment in breast cancer imaging [15].

There are too many examples to mention. But, through the research and development of CNN models by many scientists in the past decades, I will research the breast cancer histopathology image after NAT treatment by training and generated on CNN model.

1.5 Objective of this study

Although CAD of digital pathology images has been attracting great research and development interest in the last decades, it still faces many technical challenges in automatically identifying the diseased regions, conducting cell segmentation, and selecting the optimal hand-craft image features to develop machine learning classifiers for disease diagnosis. In order to better address these challenges, researchers have investigated the feasibility of using deep learning approaches to develop new CAD schemes of digital pathology images.

In this study, two quantitative CAD systems were utilized to automatically predict the cancer cellularity and classes within tumor bed in WSI breast cancer histopathology images of surgical or biopsy specimens acquired from the breast patients after NAT (i.e., neoadjuvant chemotherapies). Accurately detecting residual cancer cells and distinguishing (or diagnosing) cancer categories based on cancer cellularity levels is important for clinicians (surgeons and oncologists) to select or determine the optimal treatment strategy for the individual patients after NAT. The CAD systems also aim to minimize pathologist overloading work and give a second opinion for diagnosis. For this

purpose, we took the following steps in this study. First, a series of preprocessing methods were developed to automatically augment the size of datasets and decrease overfitting, which will influence the accuracy of predictions. Second, different schemes were compared by computing performance, and two CAD systems were used by cross-validation to identify the optimal CNN models in solving the different problems. Third, different outcome layers were analyzed to fit different evaluation metrics and circumstances. Fourth, the CNN models were ensembled from trained processes, and calculated and discussed the prediction effectiveness, advantages, and disadvantages from the results. Finally, the potentials for further improvement were discussed to amend the prediction performance.

The details of the datasets, the procedures of CAD systems and the study results are presented in the following sections.

Chapter 2: Neural networks

2.1 ANN

Just like 90 billion neurons form a complex neural network system in the human brain, the concept of neural network construction is inspired by the operation of biological (human or other animals) neural network [16]. Each connection, such as a synapse in a biological brain, can pass signals from one artificial neuron to another. Artificial neurons that receive the signals can process it and then signal other artificial neurons to connect it.

The neural network itself is not an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs [17]. Typically, artificial neurons are aggregated into layers where different layers can perform different types of conversions on their inputs. The output of each artificial neuron is calculated by activation function of the sum of its inputs. Rectified linear unit (ReLU) is one of the most common activation functions; it can be described as outputting the maximum number between 0 and x:

$$f(x) = max(0, x)$$

The connection between artificial neurons is called the edge. Artificial neurons and edges often have weights that adjust as learning progresses. The weight increases or decreases the strength of the signal at the junction. The signal may propagate from the first layer (input layer) to the last layer (output layer) after traversing the layers multiple times. The commonly used ANNs in the traditional machine learning field, which used a limited number of subjectively defined or hand-craft image features, is shown in Figure 2. The ANN has a 3-layer structure including an input layer with neurons represented by

selected image features, a hidden layer to adjust or increase discriminatory power, and an output layer linked with neurons represented different classes.

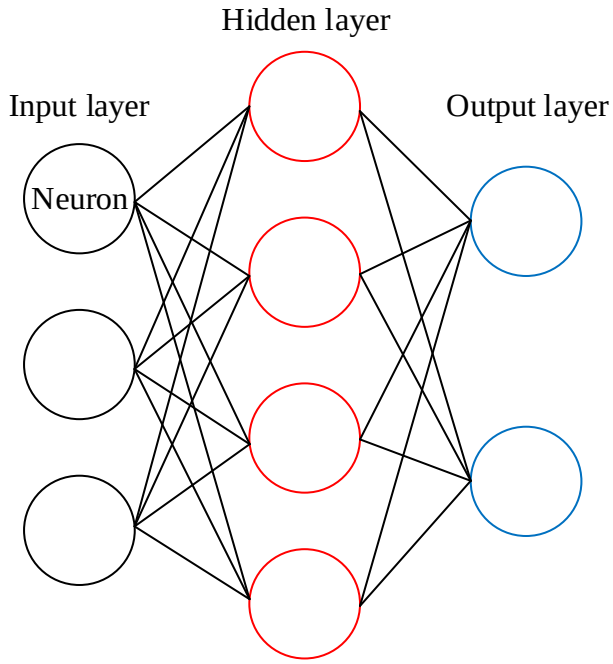


Figure 2. Simplest ANN architecture.

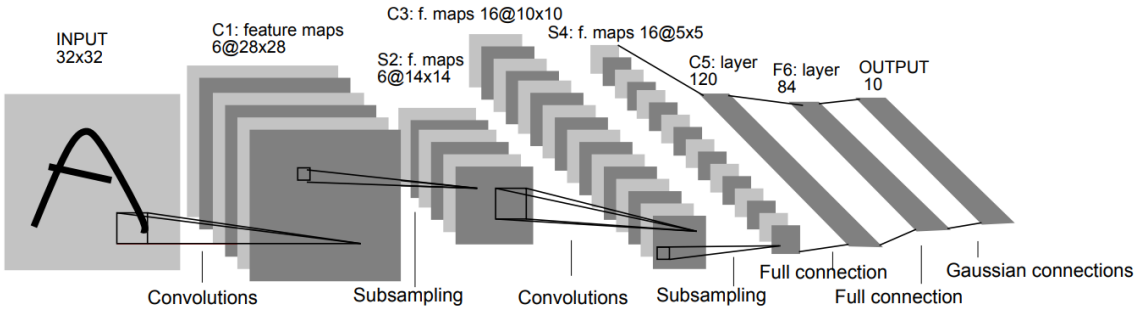
ANNs have been used on a variety of fields, including computer vision, natural language processing, image segmentation and classification, and medical diagnosis. Unquestionably, ANN is an ideal machine learning tool widely optimized and used in variety of CAD schemes of medical images including digital pathology images [18,19].

2.2 CNN

The challenge of developing an optimal conventional ANN is the difficulty in identifying a small set of effective and not redundant image features. In order to address and overcome this challenge, developing feature-less CNN has been attracting great research interest recently. A CNN is quite similar to an ANN: they are all composed of neurons, which have the weight and bias of learning ability. However, a CNN includes many input neurons and many hidden neuron layers. Each neuron gets input data and

performs a convolutional operation before performing an activation function. The network still has a loss function and a fully connection (FC) layer often performed in the last layers.

The biggest difference between the CNN and the ANN is the receptive field. Each convolutional neuron was proposed to connect to a small sub-region of the previous layer. The size of the connection is called the neuron's receptive field, and its shape is square, and size is a hyperparameter (actually the spatial size of the filter). This architecture ensures the strongest response to a spatially local input pattern. The earliest CNN architecture is LeNet in the last decade of the last century, shown in Figure 3.



(A six-layer structure: C1 + S2 + C3 + S4 + C5 + F6)

Figure 3. Architecture of LeNet, the first successful CNN application [20].

The CNN has a batch filter that continuously scrolls in the picture. It only collects a small pixel area at each time, after collecting all the information, the output value can be understood as a volume, has a higher height with the size of batch filter. It will be inadvertently losing some information. When the volume is set, retain more information and the compression work is given to the max pooling layer. Such additional work can effectively improve accuracy.

Pooling layer also used to reduce the spatial size and the number of parameters, to overcome overfitting. In addition to max pooling, average pooling was often used

historically but has recently fallen out of favor compared to max pooling, which performs better in practice [21].

The following are some of the more well-known structures in the field of CNNs, and are sorted by accuracy from low to high which is as same as the order of years of birth [22].

LeNet [20], the first successful CNN application was implemented by Yann LeCun in the 1990s. LeNet laid out three core ideas of the CNN: local receptive field, weight sharing, and downsampling.

AlexNet [23] has been successfully applied to a larger variety of computer vision tasks. It was developed by Alex Krizhevsky et al. and won the ImageNet ILSVRC challenge in 2012. The network structure is very similar to LeNet, but deeper and larger. In terms of connection design, the number of convolutional layers is increased to 5 layers, the convolutional layer scales are 11x11, 5x5, 3x3, and the FC layer is increased to 3 layers. In terms of operators, ReLU was used instead of a sigmoid as the activation function. Besides, Dropout and LRN were implemented to reduce overfitting and to normalize the data.

ZF Net [24], the network invented by Matthew Zeiler and Rob Fergus, won the ILSVRC 2013 competition. It implements the improvement of AlexNet by modifying the hyperparameters in the structure, specifically increasing the size of the intermediate convolutional layer, making the step size and filter size of the first layer smaller.

GoogLeNet [25] (Inception Net) was the winner of ILSVRC 2014. It mainly makes a big improvement in the convolutional layer and significantly reduces the number of parameters in the network (60M in AlexNet, only 4M in the GoogLeNet). While the

network deepens, the convolutional layer network is broadened. The influence of increasing the number of filters is as same as the influence of increasing the number of channels, and the widening network can be simply considered to be a wider structure.

VGGNet [26] got the runner-up placement in ILSVRC 2014. Its demonstrated that the depth of the network is a key part of the excellent performance of the algorithm. Because a 5 × 5 convolution layer can be equivalent to two layers of 3 × 3 convolutions so that the size of convolution in VGGNet is all 3 × 3, that makes the network unit modularization. However, VGGNet consumes more computing resources and uses more parameters, resulting in more memory usage. Most of the parameters are from the first FC layer. It was later discovered that even if these FC layers were removed, which will significantly reduce the number of parameters, there will be no effect on performance.

It is not hard to find that the performance of a CNN is improved by the deeper depth and wider width of networks [27]. The deeper depth means more layers in the CNN architecture, and the wider width means more channels in a single layer. However, deep networks often have gradient explosions that require very good initialization of parameters; deep networks also can overfit and increase the test loss. The emergence of Batch Normalization [28] can normalize the output of each layer so that the gradient can remain stable after being transmitted in the reverse layer without being too small or too large.

Moreover, as the depth of the network increases, the accuracy becomes saturated and then decreases rapidly. This degradation is not caused by overfitting nor by gradient explosions, but because the network is so complicated that it is difficult to achieve the ideal error rate by unconstrained stocking training, which means adding more layers to

the appropriate depth model leads to higher training errors [27,29]. The currently widely used optimization methods, whether SGD [30, 31], Adam[32], or RMSProp [33], are unable to achieve theoretically optimal convergence results as the network depth increases.

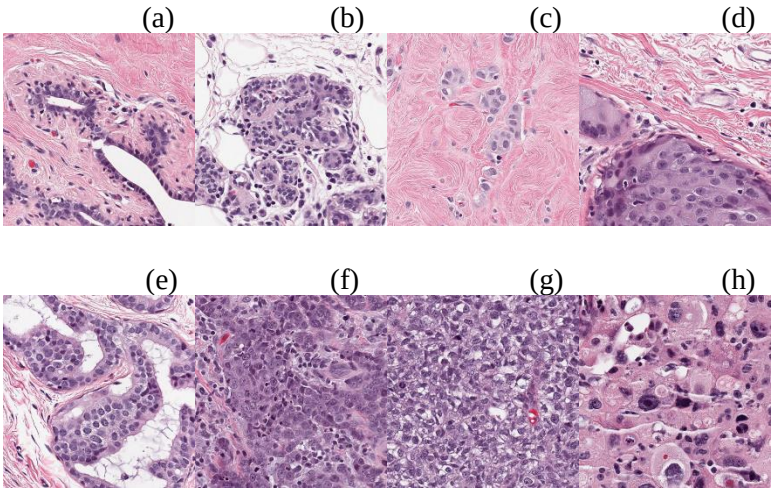
Chapter 3: Image datasets

This study used a publicly available image dataset, which was collected and ensembled from the Sunnybrook Health Sciences Centre, Toronto, Canada. It contains 3698 patch images with a size of 512 × 512 pixels. These patch images were acquired from 96 WSIs, which have been stained with H&E, a commonly used stain in pathology for prominent cells and connective tissue. After NAT therapy, 96 WSIs were extracted from excised specimens of 64 patients with residual invasive breast cancer. WSIs were prepared and scanned using a whole slice digitizer of pathology images at 20× magnification (0.5μm/pixel). The two pathologists have more than 10 years of practical experience in reading and diagnosis of histopathology images of breast lesion specimens at the University of Toronto. They were involved in interpreting these images to build "ground-truth" in detecting residual cancer cells and scoring image slices based on their subjectively graded cellularity values [34]. The Canadian Cancer Society provided research funding to support assembling this image dataset. This dataset was made publicly available for research purposes and was used as a common dataset in a competition to develop CAD systems of histopathology images of breast lesion specimens namely, the SPIE-AAPM-NCI BreastPathQ: Cancer cellularity challenge in SPIE 2019 [35]. Figure 4 shows a number of randomly selected samples of image slices from the dataset.

3.1 The training set and validation set

The whole dataset was divided into three independent subsets. The training set has 2579 patches in 63 WSIs. The validation set has 185 patches in 6 WSIs. In the both sets, the truth ground was labeled by a single pathologist. The label is called cancer

cellularity, it is defined in Chapter 1 as the percentage area of the overall tumor bed that is comprised of tumor cells (invasive or in situ). Plus, the number is a percentile decimal number between 0.00 and 1.00, where 0 corresponds to 0% cellularity, and 1 corresponds to 100% cellularity.



(a-b) Cellularity Scores are 0; (c) Cellularity Score is 0.1; (d) Cellularity Score is 0.35; (e) Cellularity Score is 0.65; (f) Cellularity Score is 0.9; (g-h) Cellularity Scores are 1.0.

Figure 4. Samples of cancer cellularity values.

Table 1 and Figure 5 illustrate the statistics of truth ground in the training set and validation set. As we can see in Table 1, the distribution of cellularity value is relatively uniform. In Figure 3, it is clear that the labels from 0 to 0.1 and 0.9 to 1.0 are stepped by 0.01 approximately; the labels between 0.1 and 0.9 are stepped by 0.05.

Table 1. Cancer cellularity value and distribution in classes.

	Cancer cellularity value and distribution in classes			
	0%	1 – 30%	31 – 70%	>70%
Training (Pathologist A)	670 (28.0%)	775 (32.4%)	597 (24.9%)	352 (14.7%)
Validation (Pathologist A)	31 (16.8%)	65 (35.1%)	68 (36.8%)	21 (11.4%)

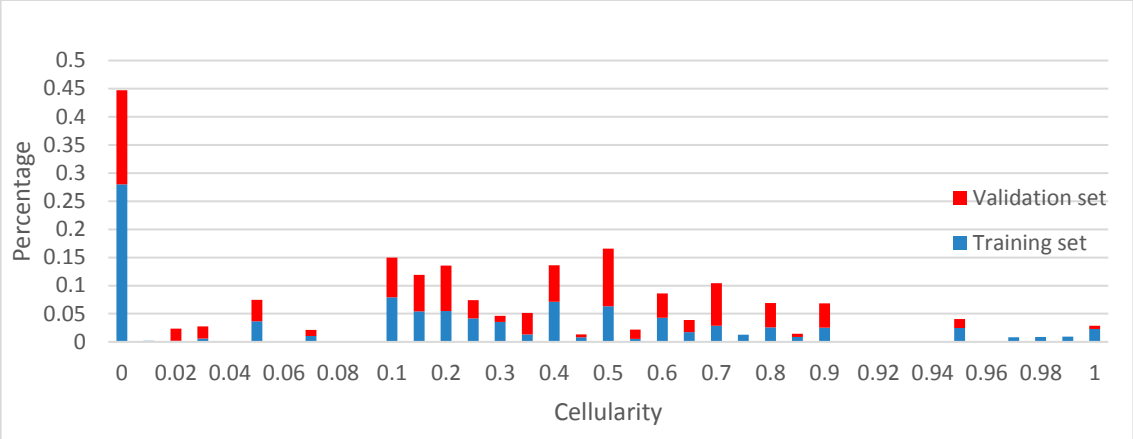


Figure 5. Percentage of cellularity value in the training set and validation set.

3.2 The testing set

The testing set has been prepared and scanned in an identical manner as the training set, which includes 1119 patches in 27 WSIs acquired from 18 patients. In the testing set, the truth ground was labeled by two pathologists and one of the pathologists is the same person as the pathologist who marked the training set. This testing dataset is used to objectively evaluate performance of CAD schemes developed by different research groups in the SPIE competition. Thus, the truth ground is invisible or not released to the researchers. The competition organizer will use this testing dataset to test the CAD schemes submitted to the competition and report the performance results to the researchers only. Table 2 demonstrates the statistics of the testing set labeled by two pathologists.

3.3 ImageNet

Since a deep CNN typically involves a huge number of weights or link coefficients, it needs to be trained using a very big dataset to yield or maintain high robustness (avoiding overfitting). The number of images in this histopathology dataset is too small to fully train a complete CNN. Thus, a transfer learning method was applied

and used in this study and the initial epoch of the CNN was pretrained by a published large dataset named as an ImageNet [36]. ImageNet is an image dataset organized according to the WordNet hierarchy [37] and consists 1000 classes. Images of each concept are quality-controlled and human-annotated. The models were trained on 1.28 million training images, evaluated on 50k validation images, and tested on 100k testing images.

Table 2. Cancer cellularity value and distribution in classes.

	0%	1 – 30%	31 – 70%	>70%
Test (Pathologist A)	242 (21.6%)	225 (20.1%)	301 (26.9%)	353 (31.4%)
Test (Pathologist B)	237 (21.1%)	312 (27.8%)	375 (33.5%)	197 (17.6%)

Chapter 4: ResNet and SE-ResNeXt

In order to apply proper CNN architectures or models, based on the literature review above and knowledge learned, ResNet and SE-ResNeXt were selected as the CNN structures applying to develop the new CAD schemes of digital histopathology images. Those two structures have performed very well in the field of image processing for the past five years. To detect residual breast cancer cells and classify cancer severity based on the cancer cellularity values in this study, the selected structures have relatively different complexity, which can be compared to the performance when the database is not big enough.

4.1 Training a ResNet based CNN model

ResNet (Residual Neural Network) [38] was proposed by Kaiming He et al. from the Microsoft Research Institute. ResNet was successfully trained and won the championship in all five main tracks of ILSVRC & COCO 2015 competitions (ImageNet: classification, detection, localization, and COCO: detection and segmentation). At the same time, the parameter quantity of ResNet is lower than VGGNet, the runner-up CNN scheme. The structure of ResNet can accelerate the training of neural networks very quickly while the accuracy of the model is also greatly improved. Also, ResNet consists of some similar blocks of the convolutional layer, which has strong expandability. It can be applied not only to previous networks but also to the subsequent networks. A simple block of ResNet-50 is shown in Figure 6.

ResNet mainly solves the degradation in a deeper network. The main idea of ResNet is to add a direct connection channel (shortcut) to the network. The previous network structure was a non-linear transformation of the performance inputs, while the

Residual network allowed a certain percentage of the output of the previous network layer to be preserved and passed directly to the later layers. ResNet solved this problem to a certain extent. By directly transferring the input information to the output and protecting the integrity of the information, the entire network only needs to learn the part of the input and output differences, simplifying the learning objectives and difficulty.

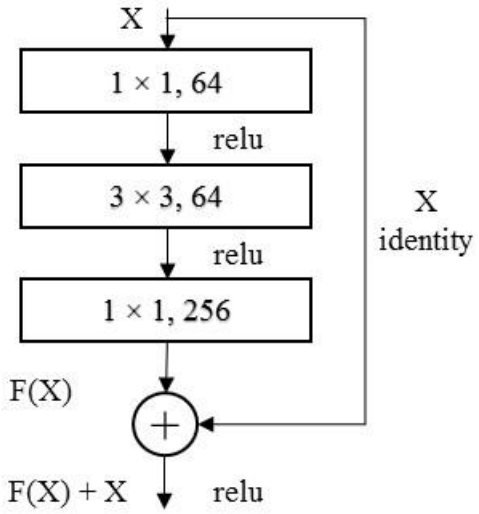


Figure 6. The block of ResNet-50.

With the most basic ResNet block, and through stacking, different layers of the ResNet scheme can be generated. The ResNet-50 simple blocks contain three convolutional layers, as shown in Figure 6 and Table 3. The two shortcuts between the three blocks in conv2 stage are directly added to the output of the convolutional network. Meanwhile, between two different stages, ResNet connected the convolutional network and shortcut by using a 1 × 1 convolution kernel. The whole architecture of ResNet-50 including the number of parameters and Floating-point Operations Per Second (FLOPs) is shown in Table 3.

The last layer of ResNet is composed of the global average pool [39], FC layer and softmax. The global average pool is mainly used to solve the problem of full

connection. The main purpose is to make the feature map of the last layer into a mean value pool of the whole graph to form a feature point and form these feature points into the final feature vector. The global average uses a simple average to establish the connection between the feature map and the category which is a very effective way to reduce overfitting. For instance, in ResNet-50, the output is operated from 7 × 7 × 2048 to 1 × 2048 by the global average pool.

When training on the classification of an ImageNet dataset that has 1000 classes, the output after the global average pool is a 2048-dimension vector. FC layer acts as a "classifier" throughout the ResNet, mapping the learned "distributed feature representation" to the role of the sample marker space which is 1000 classes in ImageNet. In practical applications, the FC layer can be implemented by convolution operations.

Softmax is used in the multi-classification process to map the output of multiple neurons to the (0,1) interval, the output with the highest probability (the value corresponds to the largest) or the probabilities of classes can be the final output result. The equation of Softmax is shown below.

$$S_i = \frac{e^i}{\sum_j e^j}$$

In ResNet-50, the error rate of the largest softmax, called top-1 error, is 20.74% and the error rate of the largest five softmax, called top-5 error, is 5.25%. ResNet is the state-of-art CNN scheme in 2016.

Table 3. The architectures of ResNet-50 and ResNeXt-50 when training on ImageNet dataset.

stage	output	ResNet-50	ResNeXt-50 (32×4d)
conv1	112×112	7 × 7, 64, stride 2	
conv2	56×56	3 × 3 max pool, stride 2	
		$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128, C=32 \\ 1\times1, 256 \end{bmatrix} \times 3$
conv3	28×28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256, C=32 \\ 1\times1, 512 \end{bmatrix} \times 4$
conv4	14×14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512, C=32 \\ 1\times1, 1024 \end{bmatrix} \times 6$
conv5	7×7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 1024 \\ 3\times3, 1024, C=32 \\ 1\times1, 2048 \end{bmatrix} \times 3$
	1×1	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax
# params.		25.5×10^6	25.0×10^6
FLOPs		4.1×10^9	4.2×10^9

4.2 ResNeXt

ResNeXt [40] proposes a strategy between common convolution and depth separable convolution: packet convolution, which balances the two strategies by controlling the number of packets (cardinality). The idea of packet convolution comes from Inception network [41]. Unlike Inception, which needs to manually design each

branch, the topologies of branches of ResNeXt are same. Finally, after cardinality architecture, ResNeXt also combined the shortcut connection and residual network.

The simple block in ResNeXt-50 (30×4d) has 32 cardinalities, the size of the set of transformations. The numbers in each rectangle of Figure 7 represent numbers of channels, filter size, and numbers of out channels, it roughly has the same complexity of block in Figure 6.

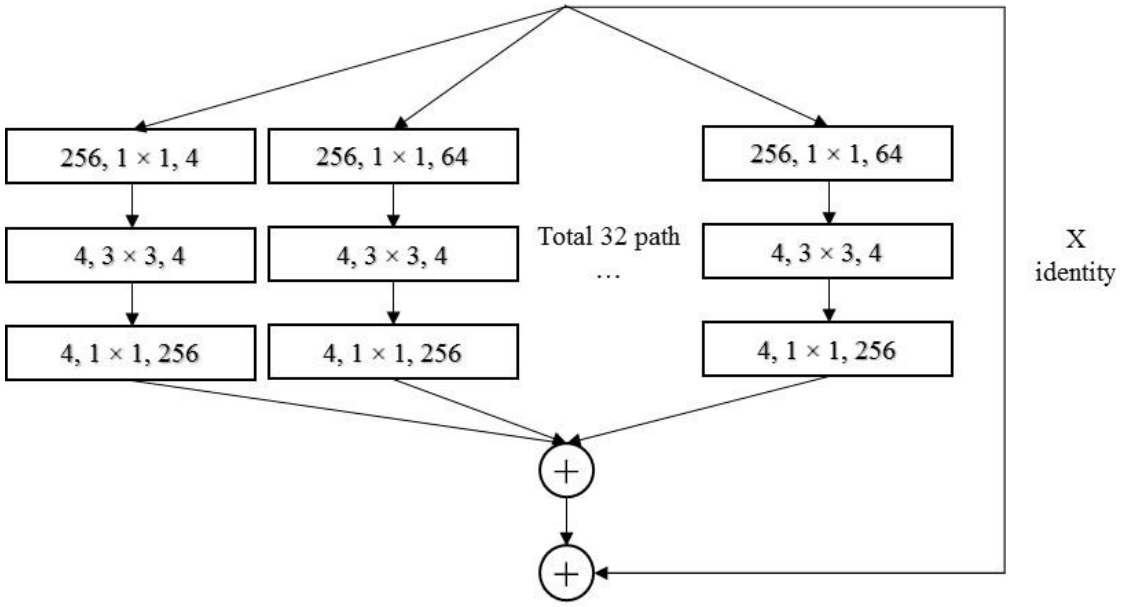


Figure 7. A simple aggregated residual transformation block.

The architectures of two blocks illustrated in Figure 7 and Figure 8 are similar. In Figure 7, thirty-two paths of convolutional blocks were concatenated initially, and then the shortcut of input was concatenated as well, the paths here are all the same topology. On the contrary, Figure 8 is combining the idea of group convolution proposed in AlexNet [23]; the width is grouped and convoluted in the same layer, the input and output channels of 32 groups are all 4. This algorithm can reduce the amount of calculations by combining only after convolution in one convolution layer. In the ResNeXt paper [40], through experiments, the author proves that the two architectures are completely equivalent, and

the results of the two structures are exactly the same. However, using the Figure 8 model can speed up the training and be more concise, so ResNeXt uses the architecture of Figure 8 as the block.

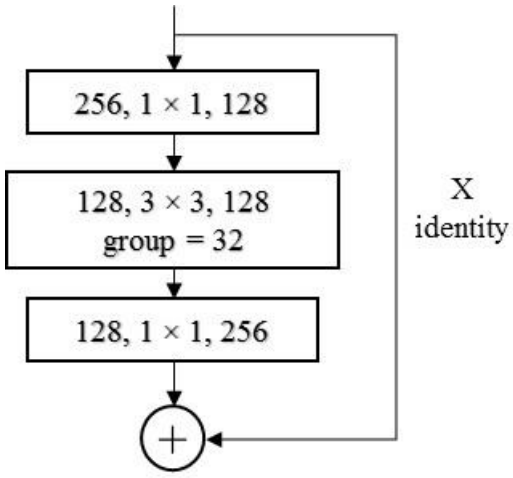


Figure 8. The block of ResNeXt-50. Implemented as grouped convolutions [23].

Experiments demonstrate that increasing cardinality is a more effective way of gaining accuracy than going deeper or wider, especially when depth and width start to give diminishing returns for existing models. In particular, a 101-layer ResNeXt is able to achieve better accuracy than ResNet-200 [42] but has only 50% complexity. The detailed comparisons between ResNet-50 and ResNeXt-50 are shown in Table 3.

4.3 Training a SE-ResNeXt based CNN model

SENet [43], the winner of the last ImageNet competition Image Classification mission in 2017 reduced top-5 error to 2.251% on the ImageNet dataset, with the original best score being 2.991%.

It is obvious that there are already a lot of attempts to improve the performance of the network in the spatial dimension, for instance, ResNet and ResNeXt. Besides, it is possible that the network can be considered to improve performance from other parts,

such as considering the relationship between feature channels. Specifically, based on learning loss, enhance the weight of useful features according to the degree of importance, and suppress the features that are not useful for the current task. The Squeeze-and-Excitation (SE) is a method to enhance and suppress the features. SENet is not a complete network structure, but a substructure that can be embedded in other classification or detection models to recalibrate the original features.

Figure 9 shows the basic idea of SE. The most left cube has three parameters, H, W, and C, which have no relationship with the size of output and numbers of channels. C represents the number of features, H and W are the sizes of features. In the Squeeze stage, through a global pooling, feature compression is performed along the spatial dimension, and each two-dimensional feature is turned into a real number that has a global receptive field to some extent, this process does not change the feature numbers C. In the Excitation stage, as demonstrated in Figure 10, the r is a scaling parameter, which is 16 in the SENet paper [43]. The purpose is to reduce the number of channels and reduce the amount of calculation. After two FC layers and two activation functions, the SE architecture achieves the purpose of learning feature weights.

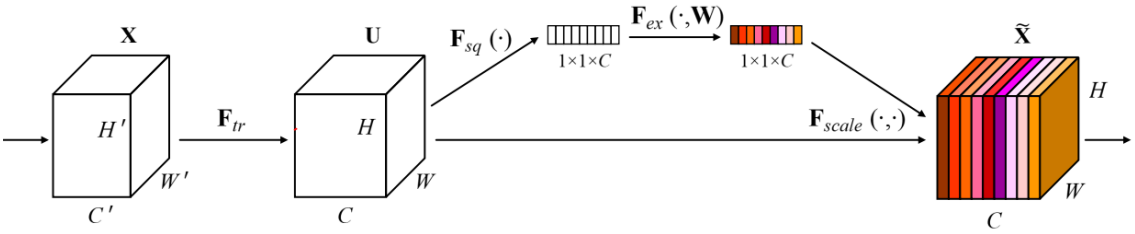


Figure 9. A Squeeze-and-Excitation block.

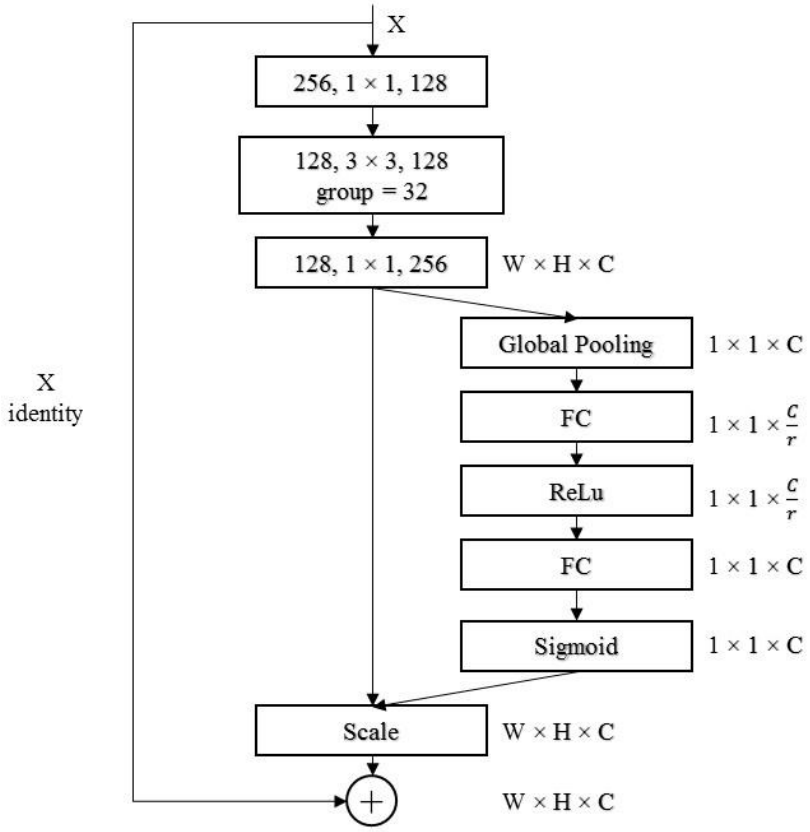


Figure 10. The block of SE-ResNeXt-50.

Finally, in a reweight operation, the weight of the output of the Excitation can be regarded as the importance of each feature channel after the feature selection, and then weight-by-channel weighted to the previous feature to complete the pair in the channel dimension. The architecture of SE-ResNeXt-50 is illustrated in Table 4.

Table 4 shows that the SENet construct is very simple and easy to deploy without introducing new functions or layers. In addition, it has good characteristics in terms of model and computational complexity, additional model parameters only exist in two new FCs. So, the FLOPs which was re-implemented by SENet paper [43] between ResNet-50 and SE-ResNeXt-50 is not differing greatly.

Table 4. The architectures of ResNet-50 and SE-ResNeXt-50 when training on ImageNet dataset.

Stage	output	ResNet-50	SE-ResNeXt-50 (30×4d)
conv1	112×112	7 × 7, 64, stride 2	
conv2	56×56	3 × 3 max pool, stride 2	
		$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1\times1, 128 \\ \text{conv}, 3\times3, 128, C=32 \\ \text{conv}, 1\times1, 256 \\ \text{fc}, [16,256] \end{bmatrix} \times 3$
conv3	28×28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} \text{conv}, 1\times1, 256 \\ \text{conv}, 3\times3, 256, C=32 \\ \text{conv}, 1\times1, 512 \\ \text{fc}, [32,512] \end{bmatrix} \times 4$
conv4	14×14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} \text{conv}, 1\times1, 512 \\ \text{conv}, 3\times3, 512, C=32 \\ \text{conv}, 1\times1, 1024 \\ \text{fc}, [64,1024] \end{bmatrix} \times 6$
conv5	7×7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1\times1, 1024 \\ \text{conv}, 3\times3, 1024, C=32 \\ \text{conv}, 1\times1, 2048 \\ \text{fc}, [128,2048] \end{bmatrix} \times 3$
	1×1	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax
FLOPs		3.86×10^9	3.87×10^9

Chapter 5: Experimental methods

In this study, the experiments for SPIE competition and follow-up are discussed. First, the experiment for SPIE competition, which is predicting the cancer cellularity value. Second, the follow-up experiment is predicting the classes of cancer cellularity. After selected two CNN structures, ResNet and SE-ResNeXt, the following experiments were conducted to build the final integrated models and evaluate its performance.

5.1 Pre-processing

5.1.1 Transfer learning

CNN needs to train on the dataset and get useful information from the data, which in turn translates information into corresponding weights and bias in network. The weights and bias are parameters, not fixed values, that need to be iteratively optimized. Although using a large amount of time and hardware resources, the optimal value of these parameters can be found, a more effective way is to extract these weights and translate to our target neural networks. This process is called transfer learning [44].

Because of the not big enough dataset, transfer the parameters from the same model which was trained on other similarity big datasets to this CNN model before training is an effective way. And then fine-tuning the models by training and iterating on these parameters. In this study, pre-trained models which contain the parameters of all layers were used from ImageNet [37], a set of containing 1000 classes annotated daily images. And pathology images were used as the input of first layer of CNN models.

5.1.2 Data augmentation

When training a machine learning model, the goal is to get an objective function that contains the optimal network parameters. In CNN, if the parameter can predict the

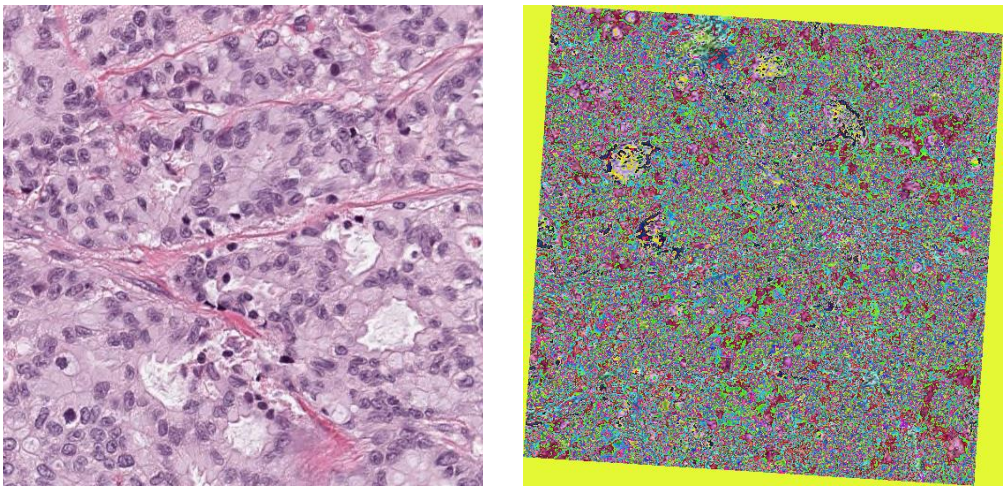
result even if it performs a variety of transformations on the dataset, it will be a robust and optimized network.

In addition, data augmentation can not only perform transformations on the image but also expand the size of the dataset to minimize the influence of overfitting. When training on the training set, the CNN predict and fit the results precisely and flawlessly, but when testing on the verification set and the testing set, it cannot get the same good result as training. In this circumstance, CNN was considered to have been overfitting. The main reason for this phenomenon is the presence of noise in the training data or the small scale of the training dataset.

In specifically, the dataset and truth grounds in this study are 512 × 512 square images and cancer cellularity, a real number between 0 and 1. When augmenting the dataset, the relationship between the image and the truth ground cannot be changed. If random cutting and random rotation without expanding are performed, the cancer cellularity value in the image will be transformed without the truth ground, so that it will be an inconsequent augmentation.

Implementation of data augmentation includes: random flips (horizontal, vertical and both), random rotation (-5 to 5 degrees, expand space area), random grayscale (10% probability), color jitter (saturation = 0.2, hue = 0.25), and color normalize (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).

Finally, after data augmentation transforms, convert the original image size 512 × 512 to the input size 224 × 224 that matches the model. A comparison of example from the training set and its data augmentation output are shown in Figure 11.



(left: Original; right: After augmentation)

Figure 11. The original data and augmentation outputs.

5.2 Model training and validation

In addition to the model structures, details of the convolutional layer, and model complexity introduced in CNN section, the specific parameters in networks, subtle changes of the models, and methods of ensemble models will be explained in this section.

5.2.1 Cross-validation

Cross-validation is a statistical analysis method used to verify the performance of a model. The basic idea is splitting a part of the training set as a new training set, and another part as a validation set. Use the new training set to train the model, and another set for evaluation.

By using cross-validation method, the training set and the testing set can be completely isolated which is a basic rule of machine learning. Only the training set can be used in the training process of the model, and only the testing set can be used to evaluate the merits of the models after training. Moreover, since increased the training

times of data, the generated plurality of models can be integrated together, to eliminate the enormous influence of erroneous data.

Common cross-validation methods are as follows:

Hold-out cross-validation: It's the easiest method to handle, just randomly divide the dataset into two groups. But there is no crossover, just the simplest grouping.

Leave-one-out cross-validation: Almost all samples in each epoch are used to train the model, so the distribution is closest to the original sample. But at the same time, when the numbers of the dataset are large, the calculation cost is very high.

K-fold cross-validation: The original dataset is divided into K groups, and each subset dataset is separately used as a validation set, and the remaining K-1 subset datasets are used as a training set, so that K models obtained, which can avoid overfitting effectively. The final result is persuasive and reasonable.

Finally, K-fold cross-validation method was chosen for both evaluation metrics but applied different K values. For predicting the cancer cellularity and meeting the requirements of the SPIE competition, 7-fold cross-validation and seven individual models were trained on the training set. For predicting the classes of cancer cellularity in the follow-up experiment, 5-fold cross-validation was trained and tested on the training set, the results of 5-fold cross-validation are the final prediction results of the validation set individually.

5.2.2 Models for SPIE competition

In the competition period, the distribution of the database is divided into three phases: 1. Distributing the training set and its labels; 2. Distributing the validation set and its labels; 3. Distributing the testing set without labels. Therefore, it is necessary to use

the cross-validation method to select the optimal models based on the training set in phase 1.

ResNet-50 and SE-ResNeXt-50 were selected from a number of schemes and experiments. For predicting the cancer cellularity, a linear layer was used to transfer the outputs from pre-trained models to a single number.

In phase 1 of SPIE competition, 7-fold cross-validation method was implemented by training on 2394 pathology images. Considering the overfitting and performance of the models, ResNet-50 at epoch 320 and SE-ResNeXt-50 at epoch 196 were chosen; each CNN scheme has 7 sub-models in this phase. In phase 2, competition organizer released a validation set includes 185 images for allowing participants to compare and adjust the model in the middle of the competition, SE-ResNeXt-50 was trained on the training set and tested on the validation set at epoch 384 which is the optimal one in the whole training process. Finally, ensembled 7 sub-models of ResNet-50 at epoch 320 and SE-ResNeXt-50 model at epoch 384 as Model A. Ensembled 7 sub-models of SE-ResNeXt-50 at epoch 196 and SE-ResNeXt-50 model at epoch 384 as Model B. Eight sub-models of each CNN scheme were integrated as the submission models for testing and evaluation of competition.

For both of the network models, through a large number of experiments, Adam [32] was implemented as the optimize function and $\beta 1$ is 0.9 and $\beta 2$ is 0.999 which are the default parameters. The learning rate is adaptable from 10^{-4} to 10^{-6}, the learning rate changes with different epochs, models and different loss functions. And the evaluation metric is concordance metric.

5.2.3 Models for follow-up experiment

Cancer cellularity can be used to illustrate the steps of breast cancer after NAT, the classes of cancer cellularity can be used to perfectly explain the steps that patients belong to. So that in some cases, analyze classes of cancer cellularity is more intuitive and useful than the cancer cellularity values.

To predict the classes, transferred the cancer cellularity to classes is necessary, the labels of images were changed to four discrete names as the classes at first. Class 1 represents no-cancer cells in tumor bed; Class 2 means low-level cancer cellularity in tumor bed; the pathology images in Class 3 and Class 4 are belonging to mid-level and high-level cancer cellularity in tumor bed respectively. Based on the four classes, the last layer of models should have four classes as outputs and implement a softmax layer to calculate the probability of each class which is suitable for loss function.

The training set in phase 1 and the validation set in phase 2 of the SPIE competition were combined to enlarge the dataset of predicting the classes of cancer cellularity. ResNet-50 and SE-ResNeXt-50 were trained by 5-fold cross-validation, and the results of the testing sets of each sub-model from cross-validation were constituted the final prediction results.

Those two models were evaluated by ROC. Trough average the testing results from cross-validation process, ResNet-50 at epoch 160 and SE-ResNeXt-50 at epoch 168 were chosen to be the Model C and Model D which are the optimum models. The optimize function and learning rate is as same as Model A and Model B.

5.3 Evaluation metrics and loss function

Next, for participating in SPIE competition, the Model A and Model B were evaluated using the default evaluation metric defined by the competition organizer and an independent testing dataset. Furthermore, due to the lack of truth ground of independent testing dataset and different objective in the follow-up experiment after SPIE competition, another evaluation metric which is commonly used in medical image-based CAD schemes was implemented.

5.3.1 Concordance metric and Mean Square Error

Under the subjective labeling processing, different pathologists barely have the same diagnosis results. Variability among clinical raters' results makes it difficult to define an unbiased and calibrated reference standard. Concordance metric is a good evaluation metric in this circumstance so that it is the default evaluation metric chosen by the competition organizer.

Specifically, the testing set in this study was labeled by two pathologists. Prediction Probability (P_k) is one of concordance metric. This method ranks two randomly chosen cases in the same order as the reference standard first, and finally averages all of the P_k value using the two sets: prediction set and truth ground. Concordance pair means Case 2 > Case 1 for both sets. Discordance pair means Case 2 > Case 1 for one set and Case 1 > Case 2 for another set. The formula is shown below:

$$P_k = \frac{C + \frac{1}{2}T}{C + D + T}$$

Where C is the number of concordant pairs, D is the number of discordant pairs, and T is numbers of ties in the submitted algorithm results.

Typically, there are two different P_k values between two sets by using a different set as the submitted set. Furthermore, because the P_k metric only focuses on the orders of results in the dataset, the same P_k value cannot be used to prove that the two sets are identical.

Based on this evaluation metric, Mean Square Error (MSE) was used as the loss function to iterate and optimize the models. This statistical parameter is the mean of the sum of the squares of the corresponding point errors between the predicted results and the truth ground. The formula is shown below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

Where n is the size of the dataset, y_i is truth ground and $\hat{y}_i$ is predict result, w_i is the weight of data and all of the weights are 1 in this study.

5.3.2 Receiver Operating Characteristic and cross-entropy

Concordance metric is an evaluation metric that can only predict the order of results in the dataset. To truly understand whether the predicted class of the model is similar to the truth ground's class, the Receiver Operating Characteristic (ROC) curve method was chosen to analyze the validity and rationality of the model.

To apply the ROC curve, the regression scheme in predicting the cancer cellularity values should be transferred to a classification scheme. So that the cancer cellularity value was divided into four classes to facilitate prediction. There are four possible outcomes from a binary classifier. If the outcome from a prediction is positive and as same as the truth ground, then it is called a true positive (TP); however, if the truth ground is negative then it is said to be a false positive (FP). Conversely, a true negative (TN) is when both

the prediction outcome and the truth ground are negative, and false negative (FN) is when the truth ground is positive.

The ROC is a comparison of two operating characteristics (TPR and FPR) as its discrimination threshold is varied. TPR is defined as the ratio of TP and the sum of TP and FN. FPR is defined as 1 – TNR, TNR is the ratio of TN and the sum of TN and FP. We also call TPR as sensitivity and TNR as specificity.

Unlike the regression problem, the truth ground of the classification is not a continuous number, but a discrete name that represents different classes. Therefore, the output of CNN schemes are probabilities belonging to different classes. Through the comparison between multi-classes probabilities and truth ground, the network can iterate out the optimal function. The cross-entropy function describes the distance between two probability distributions so that it can be used for loss functions in multi-classification tasks. The formula is shown below:

$$Loss = -\sum_{i=1}^{n} y log \hat{y}_i$$

Where n is the size of dataset, $\hat{y}_i$ is the probability of prediction and y_i is truth ground which can only be 0 or 1. Unlike the original truth ground, truth ground of classification problem represents whether the corresponding image belongs to a class, where 1 means belonging, 0 means not belonging.

5.4 Implementation

The integrated CNN based CAD systems were developed and tested in this study was implemented with Python based on PyTorch [45]. The experiments were conducted on an Ubuntu 16.04 LTS system with Intel (R) Xeon(R) CPU E5-1620 v4 @ 3.5 GHz

and two GPUs of Nvidia GeForce GTX 1080Ti with 32GB memory. The speed of training six models which are explained above is around one epoch per minute approximately.

Chapter 6: Experimental results

6.1 Results of SPIE competition

When trained ResNet-50 by 7-fold cross-validation method, Figure 12 shows the P_k value changes with the growth of epochs. It is clearly to see that in the whole figure, the P_k value does not have much fluctuations, so that chosen the biggest three P_k value at epoch 320, epoch 340 and epoch 308. Considering the stability of the seven folds in these three epochs and the performance when tested on the validation set in phase 2 of competition, epoch 320 was selected to become the seven sub-models of Model A.

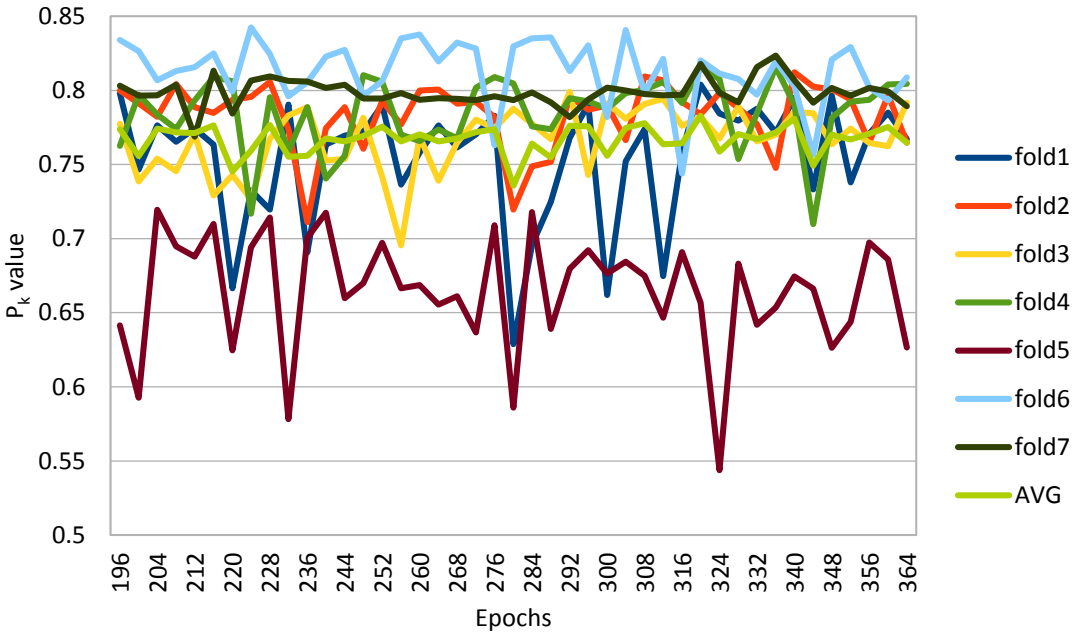


Figure 12. The P_k value of Model A by 7-fold cross-validation.

Different from the chosen in ResNet-50, because of the complexity of SE-ResNeXt-50 is much larger than ResNet-50. When trained SE-ResNeXt-50 by 7-fold cross-validation method, chosen three epochs to compare the performance: the smallest average MSE loss value at epoch 328, shown in Figure 13; the biggest average P_k value

at epoch 312, shown in Figure 14; and the first epoch when started to collect information which is epoch 196.

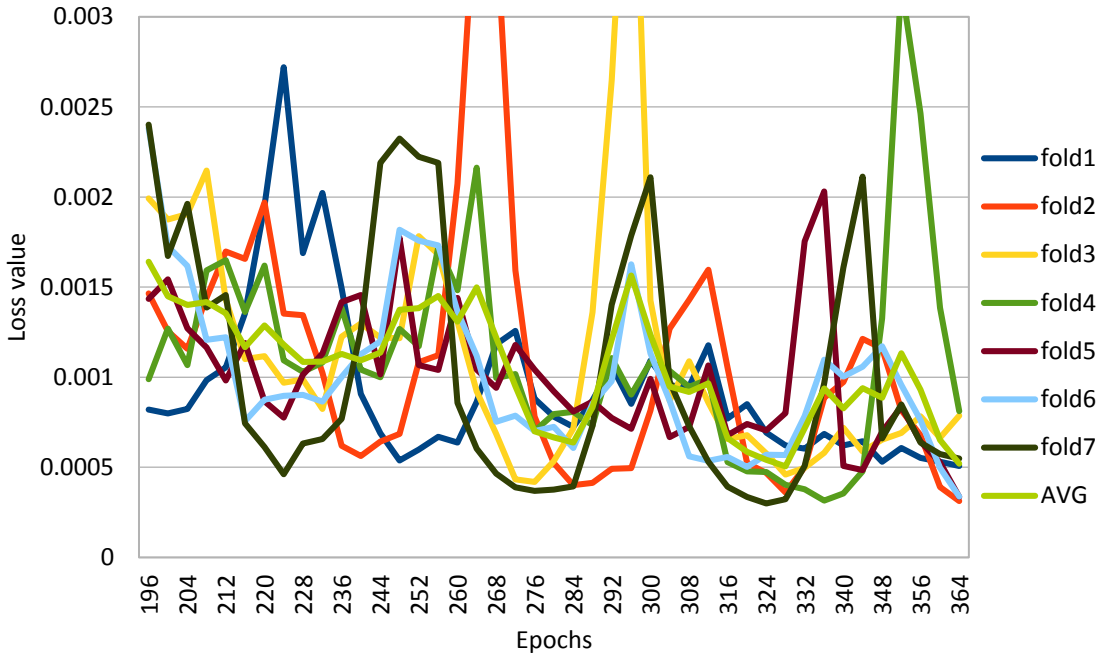


Figure 13. The MSE Loss value of Model B by 7-fold cross-validation.

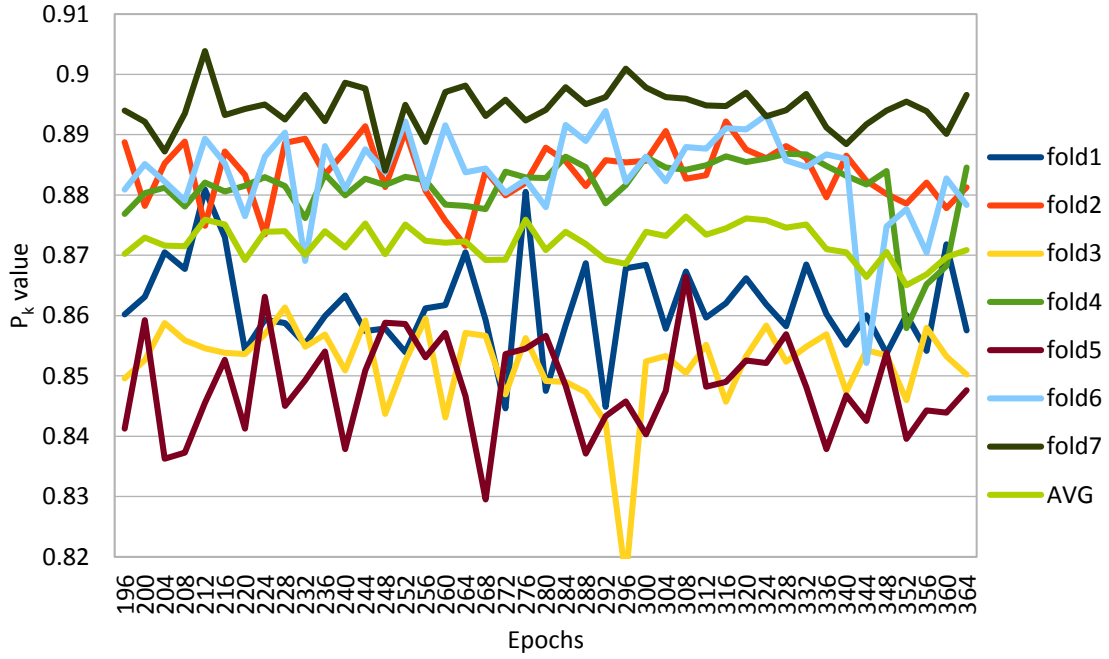


Figure 14. The P_k value of Model B by 7-fold cross-validation.

Finally, when tested on the validation set in phase 2 of competition, epoch 196 got the best prediction and was selected to become the seven sub-models of Model B. Even though it has not been trained to fit the training set very well, the early epoch performed better in the testing set, which shows that the complexity of SE-ResNeXt-50 is not suitable for the small database in this study, and the data deviation will have a great impact. In other words, the epoch that fits better with the training set has produced overfitting.

In the end, a CNN model was trained on the whole training set without cross-validation and tested on the validation set in phase 2. The loss and P_k value of this model is shown in Figure 15 and Figure 16.

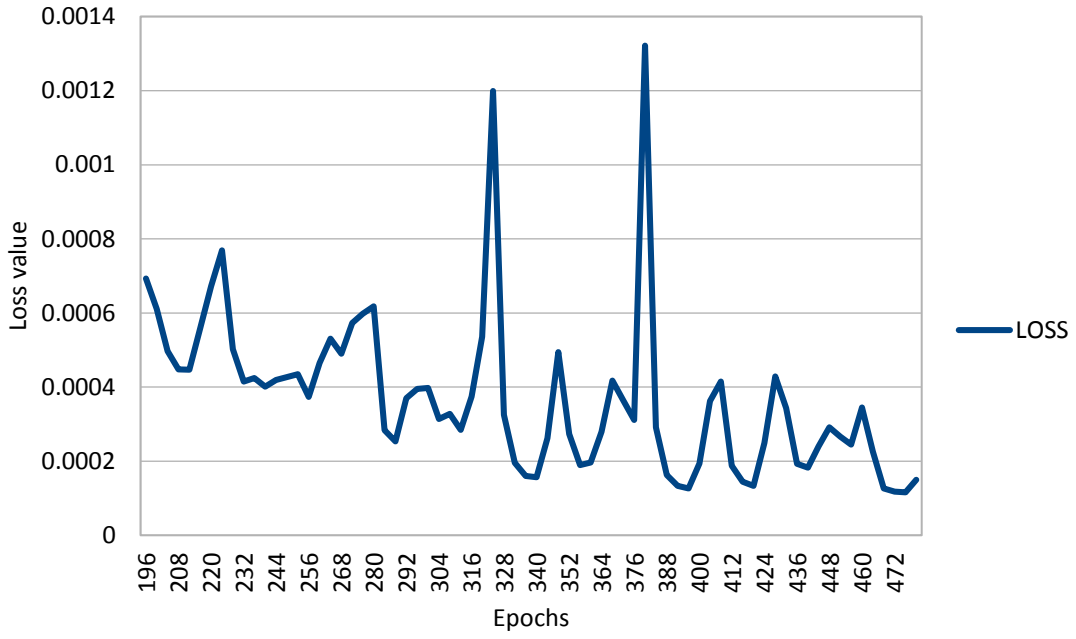


Figure 15. The MSE loss value of SE-ResNet-50 training in the whole training set.

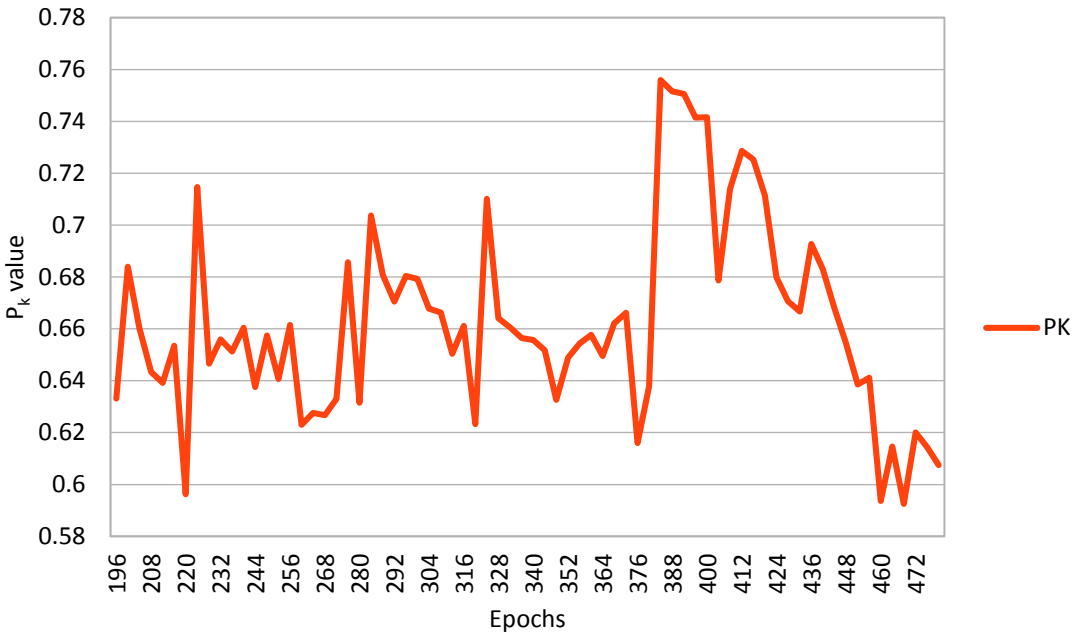


Figure 16. The P_k value of SE-ResNet-50 training in the whole training set.

The model at epoch 384 which have the optimum result P_k value was chosen to make up the last sub-model in Model A and Model B. The final prediction P_k value was calculated by the average of results of eight sub-models in Model A and Model B.

P_k value of Model A is 0.825 when tested on the validation set and 0.92004 when tested on the testing set.

P_k value of Model B got 0.803 when tested on the validation set and 0.92345 when tested on the testing set.

The testing set was labeled by two pathologists, the P_k value is 0.962 when pathologist B as a submitted group and pathologist A as truth ground group, in contrast, when pathologist A as a submitted group and pathologist B as truth ground group, the P_k value is 0.929. The different results are due to the characteristic of the concordance metric.

There are 101 total qualified algorithms from 37 participations submitted to competition organizer. The ranking of participation still does not be released. But Figure 17 shows the distribution information of results in the SPIE challenge competition.

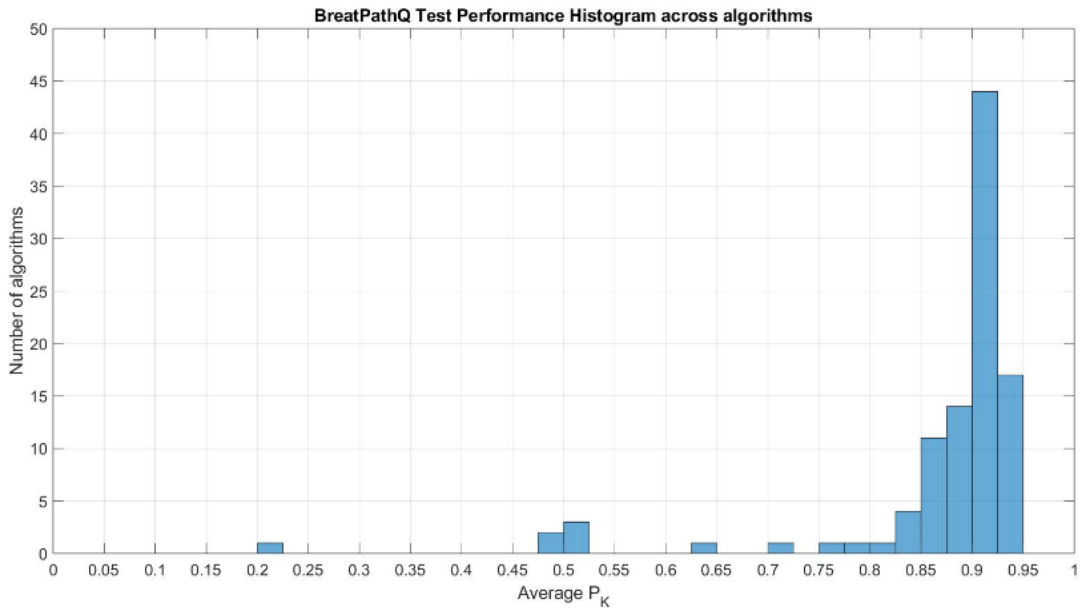


Figure 17. Distribution information of results in the SPIE challenge competition.

6.2 Results of follow-up experiment

Different from the prediction of cancer cellularity value, when set up a CAD system to predict the classes of cancer cellularity, the cancer cellularity should be transferred to four classes first. Because of lacking the labels of the testing set, the models can only get the results of the validation set in the five sub-models through the 5-fold cross-validation method. Finally, the five results were combined to cover the entire training set, in this way, the labels of the training set did not affect the predicted results. By comparing the largest accuracy values, cross-entropy loss values and complexity of models, the epoch 160 of ResNet-50 as Model C and the epoch 168 of SE-ResNeXt-50 as Model D were optimum choices for implemented to the CAD system. The accuracy of Model C is 66.93% and loss value is 2.18×10^{-5} , the accuracy of Model D is 73.05% and

loss value is 9.29×10^{-6}. Related figures are shown in Figure 18 to Figure 21. Figure 22 and Figure 23 shows the ROCs of Model C and Model D.

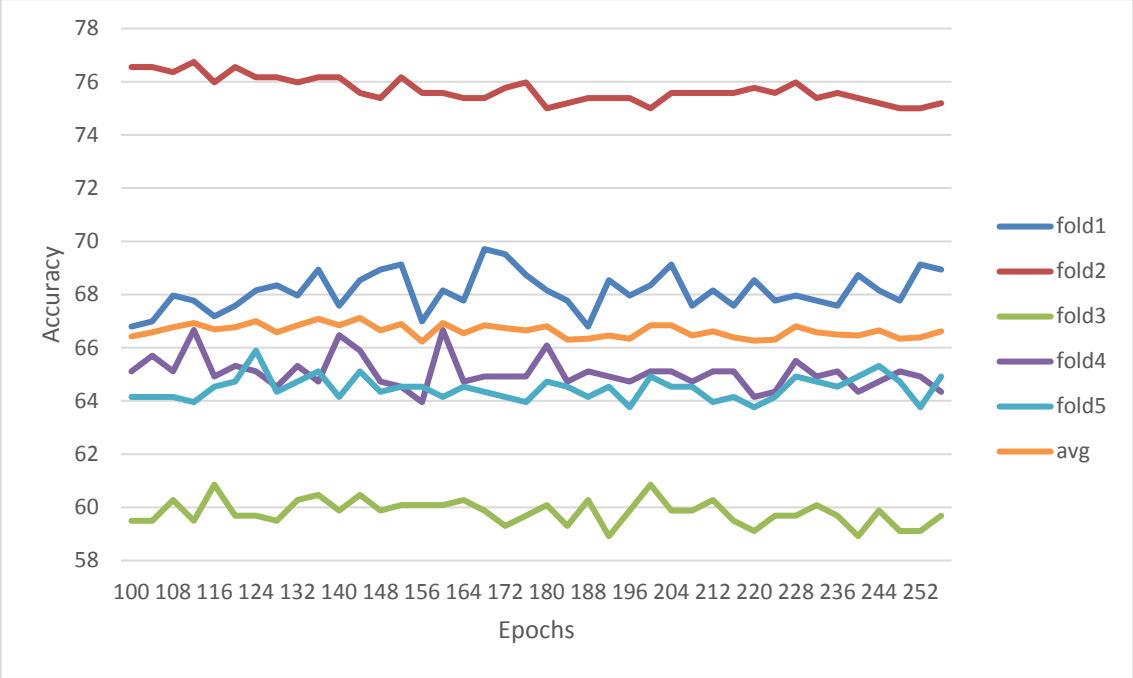


Figure 18. The accuracy of cancer cellularity classification by Model C.

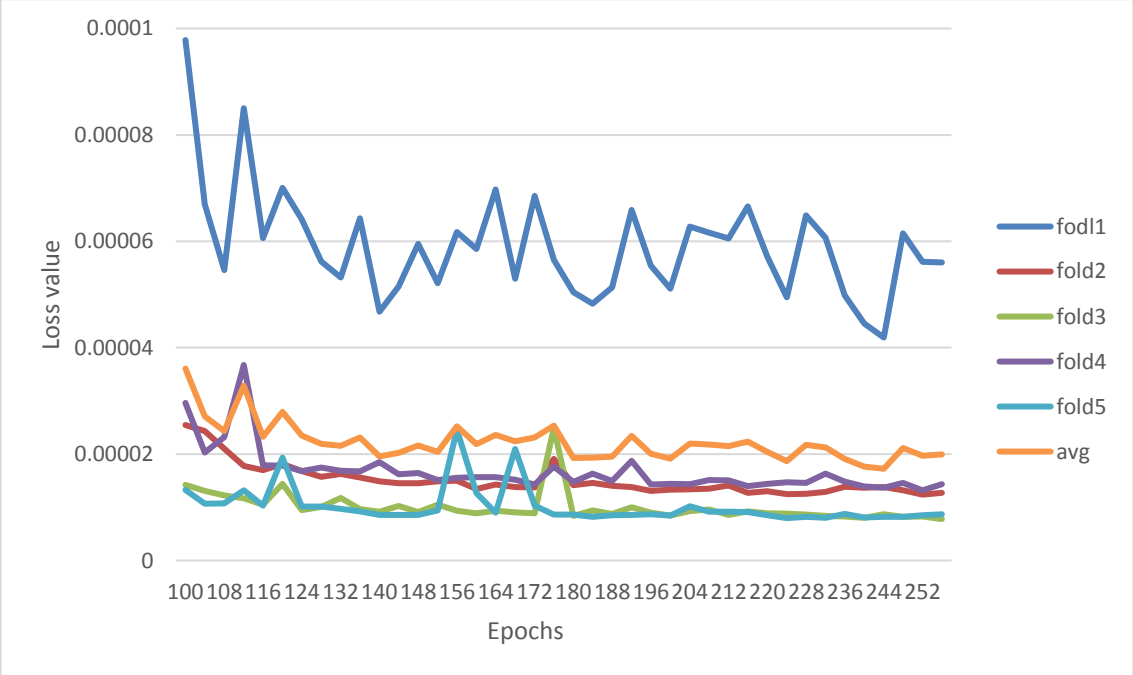


Figure 19. The cross-entropy loss of cancer cellularity classification by Model C.

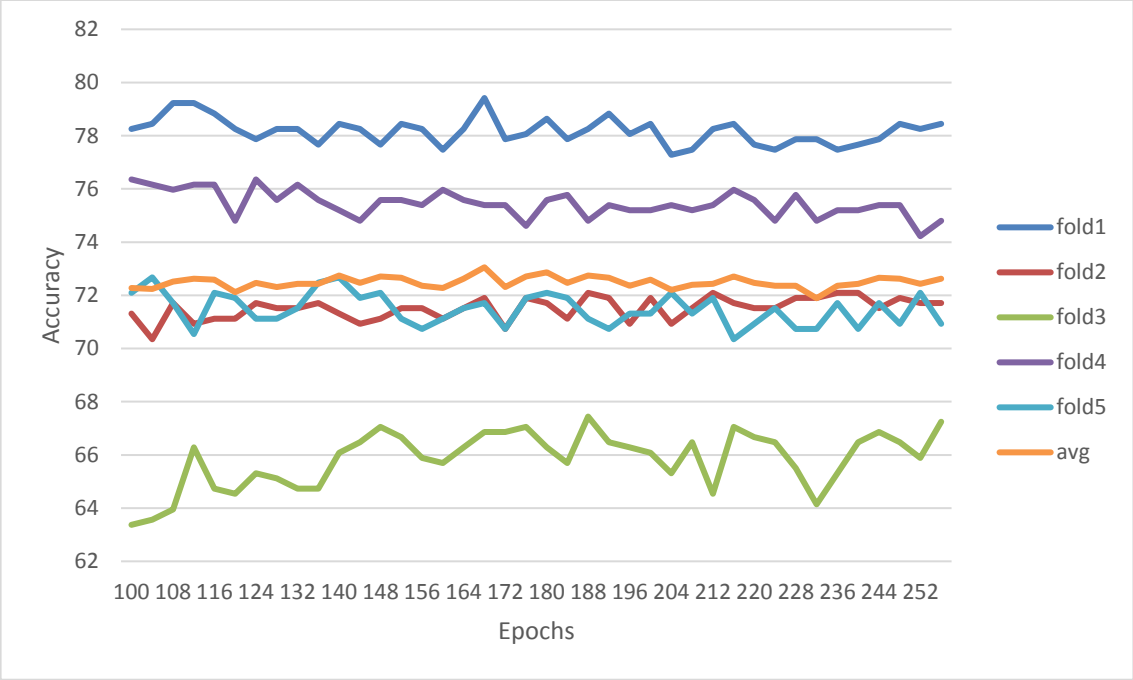


Figure 20. The accuracy of cancer cellularity classification by Model D.

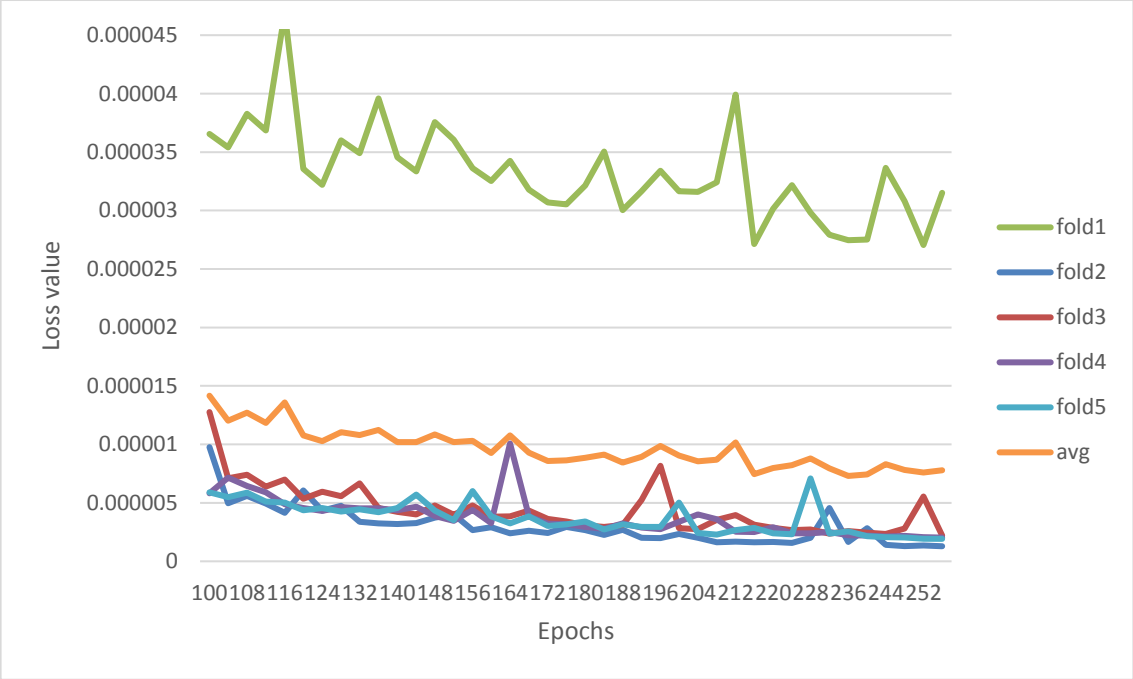


Figure 21. The cross-entropy loss of cancer cellularity classification by Model D.

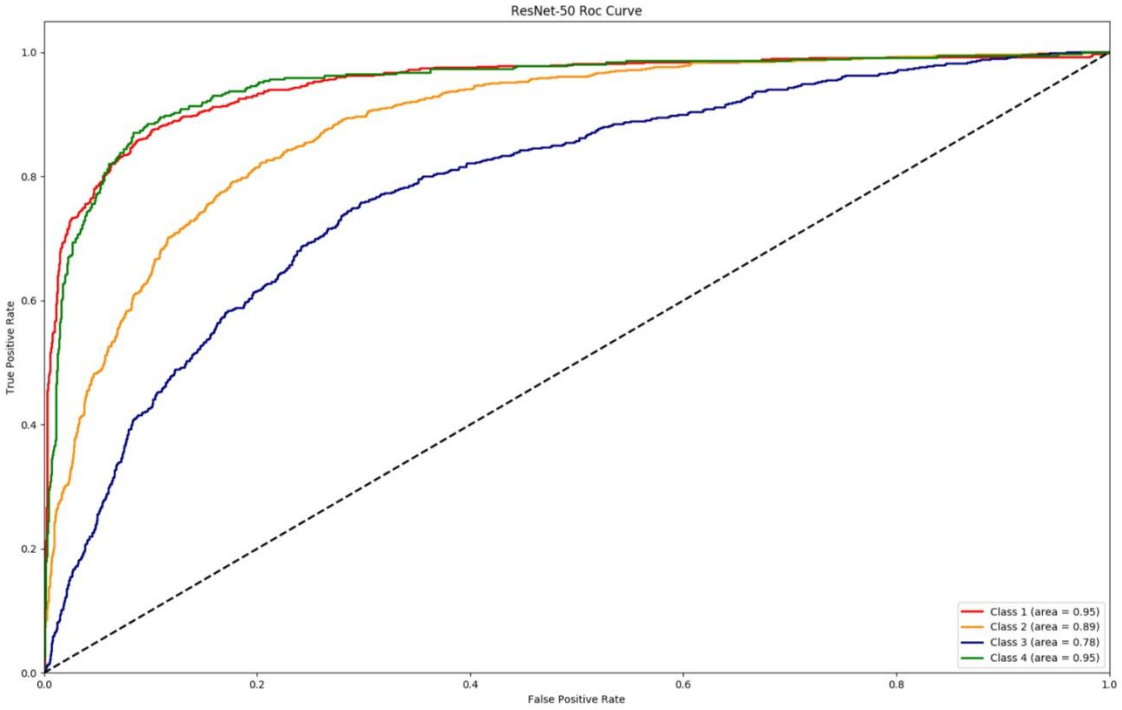


Figure 22. The ROC curve of Model C when predicting the classes.

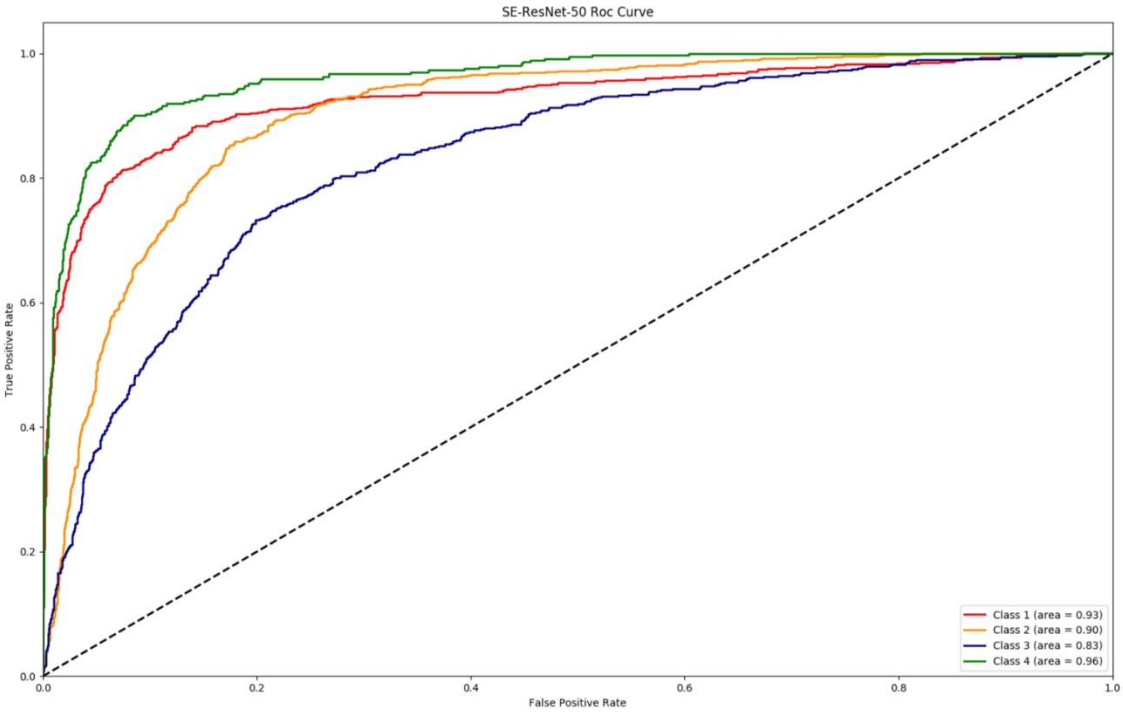


Figure 23. The ROC curve of Model D when predicting the classes.

The evaluation index of the binary classifier, area under the ROC curve (AUC), can be used as an indicator to evaluate models, AUC equals to 1 is the ideal model. For Model C, AUC of four classes are 0.95, 0.89, 0.78 and 0.95, the average AUC value of four classes is 0.893 ± 0.113; For Model D, AUC of four classes are 0.93, 0.90, 0.83 and 0.96, the average AUC value of four classes is 0.905 ± 0.075.

Table 5 and Table 6 summarized the performance of the sensitivity and specificity with different thresholds over the different CNN schemes.

Table 5. Performance of Model C for cancer cellularity classification.

Classes	Threshold = 0.3		Threshold = 0.5		Threshold = 0.7		Threshold = 0.9	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
Class 1	0.595	0.988	0.488	0.995	0.412	0.997	0.245	0.998
Class 2	0.858	0.746	0.814	0.800	0.750	0.848	0.639	0.901
Class 3	0.687	0.757	0.585	0.822	0.481	0.878	0.301	0.938
Class 4	0.820	0.937	0.751	0.956	0.694	0.971	0.539	0.986

Table 6. Performance of Model D for cancer cellularity classification.

Classes	Threshold = 0.3		Threshold = 0.5		Threshold = 0.7		Threshold = 0.9	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
Class 1	0.763	0.947	0.725	0.963	0.665	0.975	0.568	0.987
Class 2	0.904	0.763	0.865	0.802	0.824	0.833	0.755	0.872
Class 3	0.528	0.894	0.444	0.923	0.386	0.945	0.257	0.967
Class 4	0.869	0.933	0.831	0.946	0.815	0.960	0.718	0.976

Chapter 7: Discussion and conclusion

7.1 Discussion and future works

In this study, two CAD systems were implemented for predicting the cancer cellularity value and cancer cellularity classification. For each CAD system, two CNN schemes were compared and combined to some extent to complete the assessment and diagnosis of the problem requirements. In general, during the entire research process of model selection and model evaluation, there are several widely exist problems as follows: overfitting, the complexity of models and the absence of the testing dataset.

First, whatever the system for predicting cancer cellularity value or cancer cellularity classification, the overfitting circumstance always exist. To overcome overfitting, fully use and enlarge the training set, data augmentation and cross-validation were used in the CAD systems. From the results and figures, the model which fitting more and more in the training set, the performance on the testing set is not showing better. Such as in Figure 13 vs. Figure 14, Figure 18 vs. Figure 19, and Figure 20 vs. Figure 21. Although the output values of various loss functions decreased with the increase of the training epochs, the results of evaluation did not increase significantly, instead of were stabilized near a certain value in the form of waves.

So that select the models which are at the early epochs is better for evaluation and assessment. Increase the size of the training set and let the training set covers more and more different cases that may be happened in the assessment processing are other effective ways to solve overfitting. Even if obtaining the pathology images is a time consuming and difficult task, increasing the size of the dataset is an indispensable condition for future research. At the same time, in the future, it is also important to adjust

the structure of the deep learning network models to make it more adaptable to a small number of datasets.

Second, in this study, two CNN schemes were selected to compare the differences. From the results, the performance of SE-ResNeXt-50 is commonly better than the performance of ResNet-50. The same result actually can be seen from the performances of ImageNet. This proves that a better model is effective for improving performance. However, in the field of medical image processing, because the size of the dataset is small, it is difficult to perform deep neural network processing. More complex network structures mean more loss of accuracy in evaluating diverse data. Precisely, because of the diversification and unpredictability of medical images data, there are a lot of requirements for the professional doctors. Therefore, the current CAD system can only exist as a second opinion in the doctor's clinical diagnosis process.

Third, the absence of the testing dataset. Because the testing set in SPIE competition is used to be an objective evaluation dataset, its truth ground is invisible and not released to the researchers, so in the subsequent research, only the cross-validation can be used to aggregate the predicted values of each validation. And finally, use the summarized results of the verification set as an objective output for the entire training set. Although this is an experimental process that completely isolates the training set and the testing set, have not been tested in the additional datasets, making the model cannot fully convincing for evaluating diverse data.

In addition to the above three points, there are still many details that can be improved in future research.

1. Instead of using sub-models at the same epoch, different epochs can be selected for different sub-models, which can make up for the deficiencies between the models and make the model performance more excellent.
2. Applying more data augmentation methods to augment the dataset and eliminate the effects of overfitting is another way.
3. Expand more truth ground information which is outside the medical images. Such as in the pathology image of breast cancer, the age of the patient, the number of days after cancer detected, and the number of days after NAT treatment. Using these truth grounds may have a small impact on the forecast results.
4. Do not use cancer cellularity as the only annotation, positioning the location of cancer cells in pathology images, or segmenting the boundaries of cancer cells, will greatly improve the performance of the model. But at the same time, marking these truth grounds is very time consuming and inefficient. There is no medical value in clinical manifestations.

Of course, researching and developing a better CNN model is a top priority and a key to greatly improving the performance of the CAD system.

7.2 Conclusion

In the past decade, with a significant increase in computing power and digitization in various fields, the CAD system of medical digital images has become possible. Applying CAD system to help doctors make a clinical diagnosis can greatly improve the efficiency and robustness of diagnosis, reduce the time cost, and if applied properly, it is possible for medical teaching, long-distance medical diagnosis, and even gradually independent diagnosis in the future instead of as a second opinion. For this reason, in this

study, four deep learning network models were used to predict and evaluate cancer cellularity information in breast histopathology images after NAT treatment.

As a result of the SPIE competition, the Model B presented in this study is in the top 15% of all valid submissions. The P_k value of Model B is very similar to the P_k value between two pathologists who labeled the testing set individually. From Figure 12 and Figure 14, in the competition phase 1, the P_k value of Model B is greater than the P_k value of Model A as well. It can be said that the Model B developed based on SE-ResNeXt-50 has better performance.

However, in phase 2, since the validation set only contains 185 images, and the distribution of truth ground in the validation set is quite different from the training set, the P_k value of Model B is not better than Model A, because Model A has less overfitting. The differences in the distribution of two datasets also explain why the competition results are much better than the results of the validation.

In the cancer cellularity classification, from the accuracy and cross-entropy loss value shown in Figure 18 to Figure 21, Model D is better than Model C; from the ROC curve, the Model D is larger than Model C in the AUC value of average, Class2, Class3, and Class4, Model C just better in predicting Class 1; from the performance table of sensitivity and specificity, Model D also has a better performance for predicting the classes of cancer cellularity. In summary, Model D has better performance. By developing more models to compare the performance in the future can give more comparisons and references for pathologists' clinical diagnosis.

In conclusion, this study shows that CAD systems developed by CNN schemes such as SE-ResNeXt and ResNet are an effective and efficient method for the diagnostic

of residual breast cancer cells in pathological images after NAT treatment. And let CAD systems to be the second opinion for pathologists is useful and valuable.

References

[1] breastcancer.org. https://www.breastcancer.org/symptoms/understand_bc/what_is_bc

[2] cancer.org. https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/types-of-breast-cancer.html

[3] breastcancer.org. https://www.breastcancer.org/symptoms/understand_bc/statistics

[4] J. J. Fenton, J. Egger, P. A. Carney, G. Cutter, C. D'Orsi, E. A. Sickles, et al., "Reality check: perceived versus actual performance of community mammographers," American Journal of Roentgenology, 2012.

[5] Robbins, Stanley (2010). Robbins and Cotran pathologic basis of disease (8th ed.). Philadelphia: Saunders/Elsevier. ISBN 978-1-4160-3121-5.

[6] Wilson LB (1905). "A method for the rapid preparation of fresh tissues for the microscope". J Am Med Assoc. 45 (23): 1737.

[7] Fraser Symmans W, Peintinge F, Hatzis C, Rajan R, Kuerer H, Valero V, Assad L Poniecka A, Hennessy B, Green M, et al. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. J Clin Oncol 2007;25:4414–4422.

[8] Smits AJJ, Alain Kummer J, de Bruin PC, Bol M, van den Tweel JG, Seldenrijk KA,Willems SM, Johan A. Offerhaus G, de Weger RA, van Diest PJ, et al. The estimationof tumor cell percentage for molecular testing by pathologists is not accurate. ModPathol 2014;27:168–174

[9] Peintinger F, Sinn B, Hatzis C, Albarracin C, Downs-Kelly E, Morkowski J, Gould R,Fraser Symmans W. Reproducibility of residual cancer burden for prognosticassessment of breast cancer after neoadjuvant chemotherapy. Mod Pathol 2015;28:913–920.

[10] Mendelsohn, M.L., Kolman, W.A., Perry, B., Prewitt, J.M., 1965b. Computer analysis of cell images. Postgrad. Med. 38, 567–573.

[11] Wang, Dayong, et al. "Deep learning for identifying metastatic breast cancer." arXiv preprint arXiv:1606.05718 (2016).

[12] Han, Zhongyi, et al. "Breast cancer multi-classification from histopathological images with structured deep learning model." Scientific reports 7.1 (2017): 4172.

[13] Bayramoglu, Neslihan, et al. "Towards virtual h&e staining of hyperspectral lung histology images using conditional generative adversarial networks." International Conference on Computer Vision. 2017.

[14] Sari, Can Taylan, and Cigdem Gunduz-Demir. "Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images." IEEE transactions on medical imaging (2018).

[15] Motlagh, Nima Habibzadeh, et al. "Breast Cancer Histopathological Image Classification: A Deep Learning Approach." bioRxiv (2018): 242818.

[16] van Gerven, Marcel, and Sander Bohte, eds. Artificial neural networks as models of neural information processing. Frontiers Media SA, 2018.

[17] "Build with AI | DeepAI". DeepAI. Retrieved 2018-10-06.

[18] Wu, Yuzheng, et al. "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer." Radiology 187.1 (1993): 81-87.

[19] Reddick, Wilburn E., et al. "Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks." IEEE Transactions on medical imaging 16.6 (1997): 911-918.

[20] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[21] Scherer, Dominik, Andreas Müller, and Sven Behnke. "Evaluation of pooling operations in convolutional architectures for object recognition." International conference on artificial neural networks. Springer, Berlin, Heidelberg, 2010.

[22] Coşkun, Musab, et al. "An overview of popular deep learning methods." Eur J Tech 7.2 (2017): 165-176.

[23] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[24] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.

[25] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[26] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[27] He, Kaiming, and Jian Sun. "Convolutional neural networks at constrained time cost." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[28] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015).

[29] Srivastava, Rupesh Kumar, Klaus Greff, and Jürgen Schmidhuber. "Highway networks." arXiv preprint arXiv:1505.00387 (2015).

[30] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1989.

[31] Bottou, Léon. "Large-scale machine learning with stochastic gradient descent." Proceedings of COMPSTAT'2010. Physica-Verlag HD, 2010. 177-186.

[32] Bottou, Léon. "Large-scale machine learning with stochastic gradient descent." Proceedings of COMPSTAT'2010. Physica-Verlag HD, 2010. 177-186.Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[33] Tieleman, Tijmen, and Geoffrey Hinton. "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude." COURSERA: Neural networks for machine learning 4.2 (2012): 26-31.

[34] Peikari, M., Salama, S., Nofech-Mozes, S. and Martel, A.L., 2017. Automatic cellularity assessment from post-treated breast surgical specimens. Cytometry Part A, 91(11), pp.1078-1087.

[35] SPIE-AAPM-NCI BreastPathQ: Cancer Cellularity Challenge 2019. http://spiechallenges.cloudapp.net/competitions/14

[36] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International journal of computer vision 115.3 (2015): 211-252.

[37] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." (2009): 248-255.

[38] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[39] Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." arXiv preprint arXiv:1312.4400 (2013).

[40] Xie, Saining, et al. "Aggregated residual transformations for deep neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[41] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[42] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In ECCV, 2016.

[43] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[44] Shin, Hoo-Chang, et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning." IEEE transactions on medical imaging 35.5 (2016): 1285-1298.

[45] The tutorial of PyTorch. https://pytorch.org/tutorials/