UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

NOVEL COMPUTER-AIDED DIAGNOSIS SCHEMES FOR RADIOLOGICAL IMAGE ANALYSIS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

XUXIN CHEN

Norman, Oklahoma

2022

NOVEL COMPUTER-AIDED DIAGNOSIS SCHEMES FOR RADIOLOGICAL IMAGE ANALYSIS

A DISSERTATION APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Yuchen Qiu, Chair

Dr. Bin Zheng

Dr. Hong Liu

Dr. Lei Ding

Dr. Si Wu

Acknowledgments

I would like to express my deepest gratitude to my mentor Dr. Yuchen Qiu for his tremendous guidance and support during my PhD study. It was such a privilege that Dr. Qiu offered me to participate in several different projects to explore and gradually find out what my true research interest is. In each project, Dr. Qiu not only provided insightful comments, but also worked with me to solve difficult problems. When no promising results came out of a project, Dr. Qiu never blamed me but instead encouraged me to move on. Swimming in this vast sea of freedom, I was able to safely make mistakes and learn from different mistakes. As a result, my enthusiasm of seeking a successful academic career has been growing stronger and stronger day by day. Moreover, Dr. Qiu has helped me a lot on my way of overcoming severe procrastination that had hindered me from achieving my best since I was in elementary school. Following his advice – start small and start early, now I can well utilize my potential to tackle challenging tasks confidently. In addition, Dr. Qiu's positive traits such as affability, modesty, diligence, and perseverance have influenced me in an imperceptible but constant manner. Therefore, Dr. Qiu is a great mentor to me rather than a supervisor.

Besides my mentor, I cannot thank the committee members enough: Dr. Bin Zheng, Dr. Hong Liu, Dr. Lei Ding, and Dr. Si Wu. I deeply appreciate their time, guidance, support, and encouragement throughout my PhD study. With an impartial and compassionate mind, Dr. Zheng cares and supports every student that needs his help. I am grateful for his discussions and suggestions on the mammogram project as well as paper revising. It was an enjoyable and fruitful experience for me to take Dr. Liu's two classes. The preparation, dedication, and passion demonstrated by Dr. Liu towards teaching and research are inspirational. I am appreciative of Dr.

Wu’s great advice on effectively formulating the innovations/novelties of my research when writing a grant proposal. I treasure Dr. Ding’s insightful, accurate comments, which significantly improved my presentation content and skills.

In the meantime, I want to thank Ximin Wang, my best friend since college. Even though we have been at the on opposite sides of the earth after I arrived at OU, he always helps me in different ways as much as he can. I cannot imagine a wiser, more supportive, more generous, more trustworthy friend than him.

I also want to thank Farid Omoumi for providing me much useful information regarding each of the doctoral program’s procedures. His help made everything much easier! And I want to thank my former and current fellow colleagues and friends: Dr. Yue Du, Dr. Abolfazl Zargari, Dr. Wei Liu, Ke Zhang, Neman Abdoli, Patrik W Gilley. They helped make my life at OU successful and colorful! I cherish those happy moments spent with other friends, Zhihao Zhao, Meijuan Xiao, Lianglv Chen, Siqi Wang, Yue Zhao, Shiyun Tang, Mumu, Shaoyun Yi, Kai Sun, etc.

Last but not least, I am extremely thankful to my mom, dad, and sister for their unconditional love and support!

Table of Contents

Contents

List of Tables

List of Figures

Abstract

The computer-aided diagnosis (CAD) scheme is a powerful tool in assisting clinicians (e.g., radiologists) to interpret medical images more accurately and efficiently. In developing high-performing CAD schemes, classic machine learning (ML) and deep learning (DL) algorithms play an essential role because of their advantages in capturing meaningful patterns that are important for disease (e.g., cancer) diagnosis and prognosis from complex datasets. This dissertation, organized into four studies, investigates the feasibility of developing several novel ML-based and DL-based CAD schemes for different cancer research purposes. The first study aims to develop and test a unique radiomics-based CT image marker that can be used to detect lymph node (LN) metastasis for cervical cancer patients. A total of 1,763 radiomics features were first computed from the segmented primary cervical tumor depicted on one CT image with the maximal tumor region. Next, a principal component analysis algorithm was applied on the initial feature pool to determine an optimal feature cluster. Then, based on this optimal cluster, machine learning models (e.g., support vector machine (SVM)) were trained and optimized to generate an image marker to detect LN metastasis. The SVM based imaging marker achieved an AUC (area under the ROC curve) value of 0.841 ± 0.035. This study initially verifies the feasibility of combining CT images and the radiomics technology to develop a low-cost image marker for LN metastasis detection among cervical cancer patients. In the second study, the purpose is to develop and evaluate a unique global mammographic image feature analysis scheme to identify case malignancy for breast cancer. From the entire breast area depicted on the mammograms, 59 features were initially computed to characterize the breast tissue properties in both the spatial and frequency domain. Given that each case consists of two cranio-caudal and two medio-lateral oblique view images of left and right breasts, two feature pools were built, which contain the computed features from either

two positive images of one breast or all the four images of two breasts. For each feature pool, a particle swarm optimization (PSO) method was applied to determine the optimal feature cluster followed by training an SVM classifier to generate a final score for predicting likelihood of the case being malignant. The classification performances measured by AUC were 0.79±0.07 and 0.75±0.08 when applying the SVM classifiers trained using image features computed from two-view and four-view images, respectively. This study demonstrates the potential of developing a global mammographic image feature analysis-based scheme to predict case malignancy without including an arduous segmentation of breast lesions. In the third study, given that the performance of DL-based models in the medical imaging field is generally bottlenecked by a lack of sufficient labeled images, we specifically investigate the effectiveness of applying the latest transferring generative adversarial networks (GAN) technology to augment limited data for performance boost in the task of breast mass classification. This transferring GAN model was first pre-trained on a dataset of 25,000 mammogram patches (without labels). Then its generator and the discriminator were fine-tuned on a much smaller dataset containing 1024 labeled breast mass images. A supervised loss was integrated with the discriminator, such that it can be used to directly classify the benign/malignant masses. Our proposed approach improved the classification accuracy by 6.002%, when compared with the classifiers trained without traditional data augmentation. This investigation may provide a new perspective for researchers to effectively train the GAN models on a medical imaging task with only limited datasets. Like the third study, our last study also aims to alleviate DL models' reliance on large amounts of annotations but uses a totally different approach. We propose employing a semi-supervised method, i.e., virtual adversarial training (VAT), to learn and leverage useful information underlying in unlabeled data for better classification of breast masses. Accordingly, our VAT-based models have two types of losses,

namely supervised and virtual adversarial losses. The former loss acts as in supervised classification, while the latter loss works towards enhancing the model's robustness against virtual adversarial perturbation, thus improving model generalizability. A large CNN and a small CNN were used in this investigation, and both were trained with and without the adversarial loss. When the labeled ratios were 40% and 80%, VAT-based CNNs delivered the highest classification accuracy of 0.740±0.015 and 0.760±0.015, respectively. The experimental results suggest that the VAT-based CAD scheme can effectively utilize meaningful knowledge from unlabeled data to better classify mammographic breast mass images.

In summary, several innovative approaches have been investigated and evaluated in this dissertation to develop ML-based and DL-based CAD schemes for the diagnosis of cervical cancer and breast cancer. The promising results demonstrate the potential of these CAD schemes in assisting radiologists to achieve a more accurate interpretation of radiological images.

Chapter 1: Introduction

1.1 The Concept of Computer-Aided Diagnosis (CAD) Schemes

Medical image interpretation directly impacts the conclusions of disease diagnosis and treatment in clinical practice. Within the radiological field, radiologists are responsible for accurately interpreting images derived from different imaging modalities, such as plain X-ray, computerized tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), etc. Since medical images usually exhibit high inter-class similarity, it is not uncommon that two different types of lesions (e.g., benign vs. malignant) may appear similar to a radiologist. Meanwhile, the accuracy of detection and diagnosis of cancers and/or many other diseases depends on the expertise of individual radiologist [1]. This can lead to a large inter-reader variability among different radiologists when reading and interpreting medical images. On the other hand, even for the same radiologist, the image interpretation accuracy may experience undesirable fluctuations over time due to factors such as fatigue and analyzing large numbers of images. In addition, the efficiency of image interpretation is another concern in situations where a mass influx of images needs immediate attention, but only limited radiologists are available. In particular, this issue was highlighted at the early stages of COVID-19 outbreak worldwide in 2020. In order to address these clinical challenges, many computer-aided detection and/or diagnosis (CAD) schemes have been developed and tested that aim at providing a “second opinion” to assist clinicians more efficiently read medical images and make diagnostic decisions in a more accurate and objective manner [2]. The scientific rationale of this approach is that using computer-aided quantitative image feature analysis can help overcome many negative factors in the clinical practice, including the wide variations in expertise of the clinicians, potential fatigue of human

experts, and lack of sufficient medical resources. Utilizing CAD prediction as a "second opinion" to facilitate clinicians' final interpretation has been explored in many studies. For example, in paper [3], the authors conducted a clinical evaluation of CAD schemes for determining cancer aggressiveness in prostate MRI, and they found that a combination of CAD prediction and radiologist assessment can achieve higher diagnostic accuracy in differentiating indolent and aggressive cancer than an individual radiologist. In a more recent study, McKinney and colleagues evaluated the complementary role of CAD schemes for predicting breast cancer from mammograms [4]. They found that the proposed scheme could correctly identify many cancer cases missed by radiologists. Furthermore, in the "double-reading process" (standard practice in UK), the model significantly reduced the second reader's workload while maintaining a comparable performance to the consensus opinion.

1.2 Classic Machine Learning and Deep Learning Based CAD schemes

CAD schemes were developed as early as the 1970s [1, 5, 6], and the progress of CAD schemes has accelerated since the middle of 1990s [7] due to the development and integration of more advanced learning algorithms. Looking back at the history of these learning algorithms, we can generally divide them into two stages: the era of classic machine learning and the era of deep learning. Although deep learning related methods have become the mainstream technology in the current CAD field [8, 9], the CAD schemes built with classic machine learning algorithms are still widely used in many medical imaging tasks that have small amounts of images. In the meantime, machine learning based CAD schemes demonstrate relatively better interpretability, since each manually extracted feature carries specific meaning with a mathematical definition. Given the limitations and advantages of classic machine learning and deep learning, my dissertation covers both types of CAD schemes.

1.2.1 ML-Based CAD Schemes

Currently ML-based CAD schemes are still recognized as a useful alternative for several important reasons. First, data forms the basis of both machine learning and deep learning algorithms, but deep learning shows a much stronger reliance on big data. As various types of medical image datasets with increasingly larger size (e.g., at least several hundred images) are being developed to facilitate training larger deep learning models, it is not hard to predict DL-based CAD schemes will be mainstream for a long time in medical image analysis. However, here underlies an important prerequisite that certain amounts of data should be available. What if only limited data is available especially when some rare disease occurs? DL-based approaches cannot achieve satisfactory performance in this situation. To give an example, at the very early stage of COVID-19 we observed that most papers, instead of using deep learning, many researchers had to adopt the traditional paradigm of training machine learning classifiers with hand-engineered features. Conventional ML-based CAD schemes can yield moderate performance with small amounts of data (e.g., ~100 images). Second, with at least thousands or millions of parameters, deep neural networks are considered black boxes despite their promising results at the experimental level. In other words, there is no well-established, universally agreed theoretical guarantee so far to explain/interpret such complex models.

For conventional CAD schemes, a common developing procedure mainly consists of four steps: (1) lesion or regions of interest (ROI) detection/segmentation, (2) quantitative feature computation, (3) optimal feature selection and reduction, and (4) machine learning classifiers training and validation. For example, Sahiner and colleagues developed a CAD scheme to perform mass classification on digital mammograms [10]. The ROIs containing the target masses were first segmented from the background using a modified active contour algorithm. Then a large number

of image features were applied to quantify the lesion characteristics in size, morphology, margin geometry, texture, etc. Hence, the raw pixel data were converted into a vector of representative features. Finally, a LDA (linear discrimination analysis) based classification model was applied on the feature vector to identify the mass malignancy. Each step mentioned above is a hot research topic accompanied by unique challenges. The classic ML-based pipelines for developing CAD schemes are shown in Figure 1-1.

1) ROI or lesion segmentation

Segmenting the set of pixels or voxels of lesions, organs, and other substructures from background regions, is a hot research topic in in medical image analysis [8]. The accuracy of segmentation results can greatly influence the results of subsequent steps. For example, in paper [11], the authors utilized bilateral mammographic density asymmetry and proposed a new density segmentation method to segment dense breast regions from mammograms. The proposed approach significantly improved the AUC of near-term breast cancer risk prediction from 0.633 ± 0.043 to 0.830 ± 0.033.

Although various methods have been proposed over the years, ROI or lesion segmentation still faces many challenges. One challenge lies in the accurate boundary delineation when lesions or ROIs exhibit high degree of resemblance to the background areas in an image. Another challenge is the segmentation of small-sized lesions (e.g., MRI multiple sclerosis) and organs (e.g., pancreas from abdominal CT scans). Besides, most metrics used to evaluate the segmentation performance are region-based (i.e., segmentation errors are computed in a pixel-wise manner). This can lead to a loss of valuable information regarding structures, shapes, and contours that are important to diagnosis/prognosis in later stages.

2) Quantitative feature extraction

Once lesions or ROIs are determined, quantitative features can be computed to describe their characteristics. This is a crucial step in the whole process of developing ML-based CAD schemes, since the input of ML classifiers are a set of features instead of the image data itself. In other words, the accuracy of ML classifiers strongly depends on the quality of the extracted features. Desirable features should simultaneously reduce the within-class pattern variability (i.e., variance between objects belonging to the same class) and enhance the inter-class variability (i.e., variance between objects from different classes). Various types of features have been designed, including topologic features (Euler number, symmetry, etc.), geometric features from contours (area, perimeter, compactness, circularity ratio, etc.), contour-based features (diameter, minor axis, eccentricity, etc.), statistical image features (mean, standard deviation, skewness, kurtosis, etc.), texture-based image features (co-occurrence matrices, discrete wavelet transforms, power spectrum features, etc.), to name a few.

Despite the effectiveness of the above features, they may not fully reflect the complexity of different types of lesions or ROIs. It would be desirable if additional features can be computed to provide complementary information. In recent years, radiomics has emerged as a new research field that aims at transforming radiological images into minable, much larger amounts (200+) of quantitative features [12]. It is believed that such a large feature pool, if mined appropriately, can improve diagnostic or prognostic outcomes [13]. In paper [14], the authors utilized the radiomics approach to perform a comprehensive quantification for tumor phenotypes in two cancer types (lung, head-and-neck cancer) from CT images. A total of 440 features were computed to quantify tumor image intensity, shape, and texture. This radiomics approach revealed a large number of features that have prognostic power but were not identified before.

3) Optimal feature selection and reduction

Although the radiomics approach can capture additional information that is important for diagnosis and prognosis, too many features may cause machine learning classifiers to suffer severe overfitting issues. For a dataset of finite size (e.g., 150 cases), the classification accuracy initially increases as more features are added and gradually achieves the maximum; however, after the peaking point, adding more features will not improve but harm the classification accuracy [15]. This implies that when many features exist, large amounts of training data are required to achieve the best classification performance. Such phenomenon is called “curse of dimensionality”, first introduced by Bellman [16]. A thumb of rule is that 5~10 cases are generally needed per feature to avoid overfitting issues. Since it is not easy to obtain large medical image datasets in practice, keeping the number of features at a proper level seems to be a reasonable choice.

Feature selection and reduction are two commonly used approaches. For feature selection, a subset of features is picked to maximize the between-class variance and minimize the within-class variance [17]. It is solved as an optimization problem in the sense that feature selection algorithms utilize specific search strategies to search the space of possible feature subsets and select the optimal subset with respect to an objective function [18]. Feature reduction refers to mapping features from the original high-dimensional space to a low-dimensional space [19]. Unlike feature selection that keeps features unchanged, feature reduction changes the original features. Commonly used methods include principal component analysis (PCA), linear discriminant analysis (LDA), etc.

4) Machine learning classifiers training and validation

Depending on whether labels of the training dataset are present, machine learning can be roughly divided into supervised, unsupervised, and semi-supervised learning. Such division also applies to deep learning. In **supervised** learning, all training images are labeled, and the model is optimized using the image-label pairs. For each testing image, the optimized model will generate a likelihood score to predict its class label [20]. For **unsupervised** learning, the model will analyze and learn the underlying patterns or hidden data structures without labels. If only a small portion of training data is labeled, the model learns input-output relationship from the labeled data, and the model will be strengthened by learning semantic and fine-grained features from the unlabeled data. This type of learning approach is defined as **semi-supervised** learning [21]. Due to its good performance, supervised learning is widely used in developing ML-based CAD schemes. More details of unsupervised and semi-supervised learning will be provided in **subsection 1.2.2** "DL-Based CAD Schemes".

In the supervised setting, classifier construction has two stages: training and testing. Two datasets are accordingly assembled to evaluate the classifier's performance, namely the training dataset and testing dataset. Prior to classifier training, each image needs to be converted into a feature vector. At the training stage, the classifier (mathematical model with parameters) takes an image feature vector as input and predicts the label (e.g., benign or malignant) for that image. Then a loss function can be calculated based on the predicted label and actual label. Guided by the loss function, parameters of the classifier are optimized. A good classifier is expected to perform well on the training dataset, but more importantly it should be able to generalize well on the testing dataset, which is unseen by the classifier during training. However, the classifier can easily either overfit or underfit the training dataset (i.e., the model is either too complex or too simple). Both

scenarios will affect the classifier's generalization capability. Cross-validation is widely used to reduce the risk of overfitting or underfitting. Common techniques include K-fold cross-validation, leave-one-out cross-validation (LOOCV), etc.

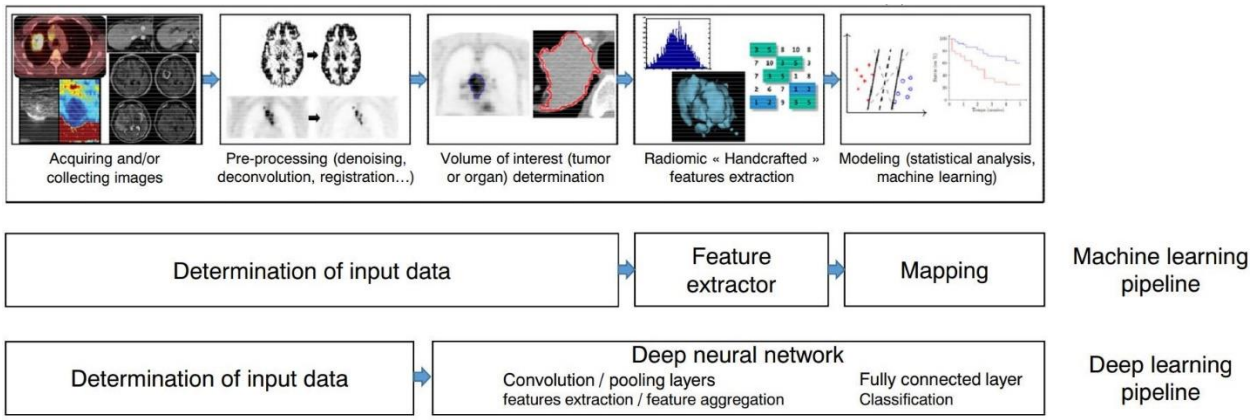


Figure 1-1: Comparison of ML-based and DL-based pipelines for developing CAD schemes (modified figure from paper [22]).

1.2.2 DL-Based CAD Schemes

In 1979, Kunihiko Fukushima invented "Neocognitron", an artificial multi-layered neural network with learning capabilities to mimic the mechanism of the brain's visiual cortex, where two types of neural cells are arranged in cascading order for pattern recognition. This design laid the foundation for the modern artificial intelligence (AI) technology with deep nets to its core.

For deep nets based CAD schemes, hidden patterns inside ROIs are progressively identified and learned by the hierarchical architecture of deep neural networks [20]. During this process, important properties of the input image will be gradually identified and amplified for certain tasks (e.g., classification, detection), while irrelevant features will be attenuated and filtered out. For instance, an ultrasound image depicting suspicious lesions comes with a pixel array [23], and each entry is used as one input feature of the deep learning model. The first several layers of the model

may initially obtain some basic lesion information, such as tumor shape, location, and orientation. The next batch of layers may identity and keep the features consistently related to lesion malignancy (e.g., shape, edge irregularity), while ignoring irrelevant variations (e.g., location). The relevant features will be further processed and assembled by subsequent higher layers in a more abstract manner. When increasing the number of layers, a higher level of feature representations can be achieved. Through the entire procedure, important features hidden inside the raw image are recognized by a general neural network based model in a self-taught manner, and thus the manual feature development is not needed.

As introduced earlier, the learning paradigms for deep learning can be roughly divided into three categories, namely supervised, unsupervised, and semi-supervised learning. Each deep learning paradigm has its own advantages and is commonly used in developing DL-based CAD schemes. Commonly used architectures of these three learning paradigms will be described in the following.

1) Supervised deep learning

Convolutional neural networks (CNNs) are a widely used deep learning architecture in medical image analysis [24]. CNNs are mainly composed of convolutional layers and pooling layers. Figure 1-2 shows a simple CNN in the context of medical image classification task. The CNN directly takes an image as input, and transforms it via convolutional layers, pooling layers, and fully connected layers, and finally outputs a class-based likelihood of that image.

At each convolutional layer l, a bunch of kernels $W = \{W_1, \dots, W_k\}$ are used to extract features from the input image, and biases $b = \{b_1, \dots, b_k\}$ are added, generating new feature maps $W_i^l x_i^l + b_i^l$. Then a non-linear transform, an activation function $\sigma(.)$, is applied resulting in $x_k^{l+1} = \sigma(W_i^l x_i^l + b_i^l)$ as the input of the next layer. After the convolutional layer, a pooling layer is

incorporated to reduce the dimension of feature maps, thus reducing the number of parameters. Average pooling and maximum pooling are two common pooling operations. The above process is repeated for the rest layers. At the end of the network, fully connected layers are usually employed to produce the probability distribution over classes via a *sigmoid* or *softmax* function. The predicted probability distribution gives a label $\hat{y}$ for each input instance so that a loss function $L(\hat{y}, y)$ can be calculated, where y is the real label. Parameters of the network are iteratively optimized by minimizing the loss function.

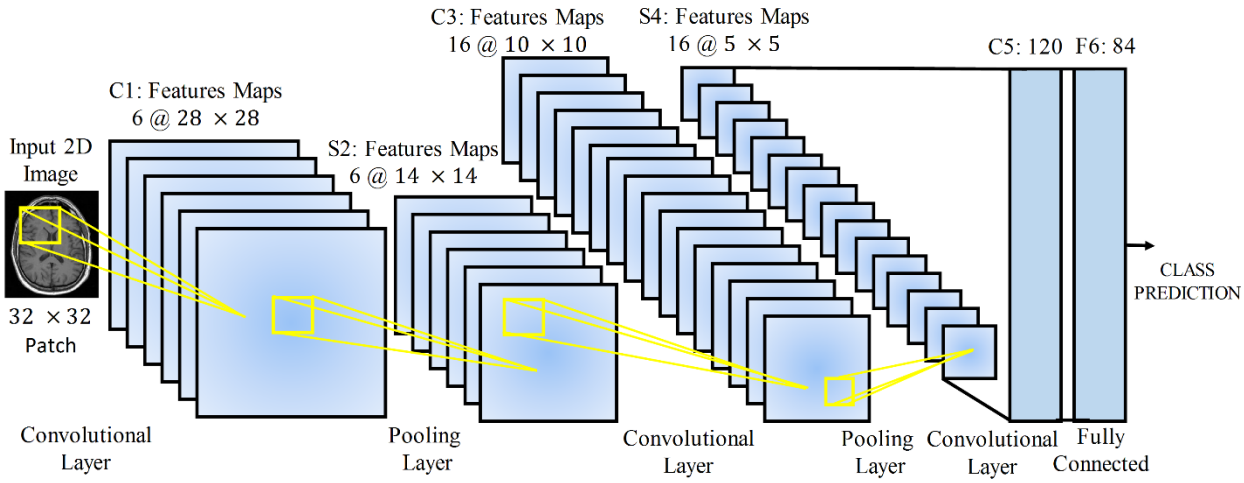


Figure 1-2: A simple CNN for disease classification from MRI images [24].

2) Unsupervised deep learning

The purpose of unsupervised deep learning is to model the underlying distribution or hidden structure of input data. Unlike supervised learning, it can directly deal with unlabeled data. Generative adversarial networks (GANs) are a class of deep nets for generative modeling first proposed by Goodfellow and his colleagues [25]. For this architecture (also known as vanilla

GAN), a framework for estimating generative models is designed to directly draw samples from the desired underlying data distribution without the need to explicitly define a probability distribution. It consists of two models: a generator G and a discriminator D. The generative model G takes as input a random noise vector z sampled from a prior distribution $P_z(z)$, often either a Gaussian or a uniform distribution, and then maps z to data space as $\boldsymbol{G}(z, \theta_g)$, where $\boldsymbol{G}$ is a neural network with parameters θ_g. The fake samples denoted as $\boldsymbol{G}(z)$ or x_g are expected to resemble real samples from the training data $P_r(x)$, and these two types of samples are sent into D. The discriminator, a second neural network parameterized by θ_d, outputs the probability $\boldsymbol{D}(x\ , \theta_d)$ that a sample comes from the training data rather than G. The training procedure is like playing a minimax two-player game. The discriminative network D is optimized to maximize the log likelihood of assigning correct labels to fake samples and real samples, while the generative model G is trained to maximize the log likelihood of D making a mistake. Through the adversarial process, G is desired to gradually estimate the underlying data distribution and generate realistic samples.

In the medical imaging field, the vanilla GAN and its variants are used to generate new medical image samples. This helps alleviate the lack of large, annotated medical datasets for training deep learning models. Figure 1-3 [26] provides a schematic view of using the vanilla GAN to synthesize CT lung nodules.

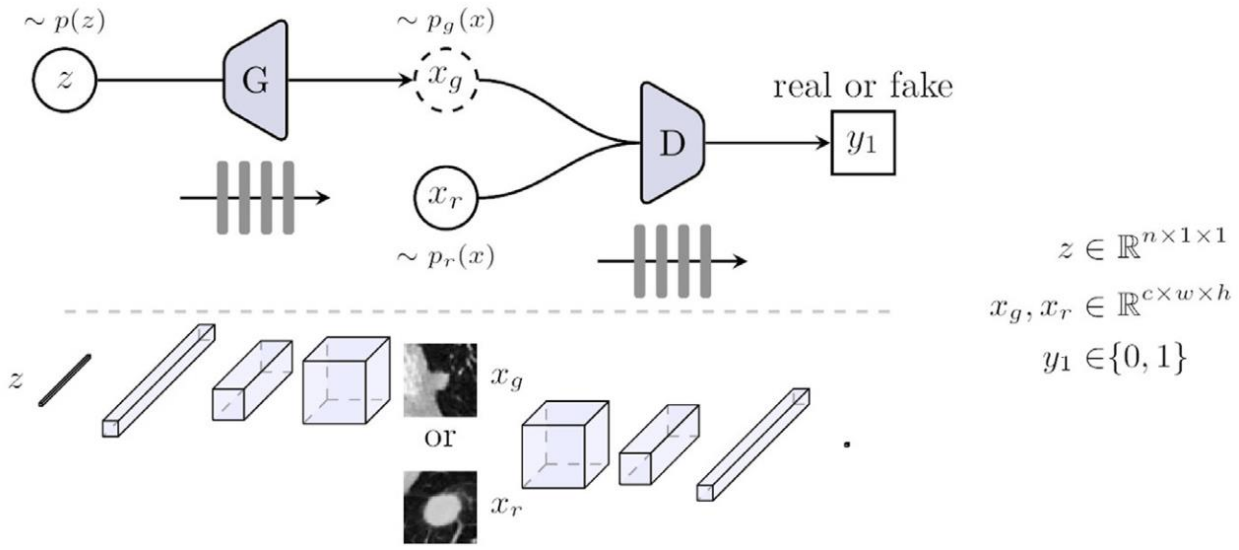


Figure 1-3: Schematic view of the vanilla GAN for synthesis of lung nodule on CT images [26]. "Top of the figure shows the network configuration. The part below shows the input, output and the internal feature representations of the generator G and discriminator D. G transforms a sample z from $p(z)$ into a generated nodule x_g. D is a binary classifier that differentiates the generated and real images of lung nodule formed by x_g and x_r respectively" [26].

3) Semi-supervised deep learning

Semi-supervised learning (SSL) combines labeled and unlabeled data during model training. Especially, SSL applies to the scenario where limited labeled data and large-scale but unlabeled data is available. These two types of data should be relevant, so that the additional information carried by unlabeled data could be useful in compensating the labeled data. It is reasonable to expect that unlabeled data would lead to an average performance boost – probably the more the better for performing tasks with only limited labeled data. In fact, this goal has been explored for several decades, and the 1990s already witnessed a rising interest of applying SSL methods in text classification. The Semi-Supervised Learning book [27] is a good source for readers to grasp the

connection of SSL to classic machine learning algorithms. Interestingly, despite its potential positive value, the authors present empirical findings that unlabeled data sometimes deteriorates the performance. However, this empirical finding seems to be experiencing changes in recent literature of deep learning – an increasing number of works, mostly from the computer vision field, have reported that deep semi-supervised approaches generally perform better than high-quality supervised baselines [28]. Even when varying the amount of labeled and unlabeled data, a consistent performance improvement can still be observed. At the same time, deep semi-supervised learning has been successfully applied in the medical image analysis field to reduce annotation cost and achieve better performance.

Popular SSL methods can be categorized into three groups: (1) consistency regularization based approaches; (2) pseudo labeling based approaches; (3) generative models based approaches. Methods in the first category share one same idea that the prediction for an unlabeled example should not change significantly if some perturbations (e.g., adding noise, data augmentation) are applied. For pseudo labeling [29], an SSL model itself generates pseudo annotations for unlabeled examples; the pseudo-labeled examples are used jointly with labeled examples to train the SSL model. This process is iterated for several times, during which the quality of pseudo labels and the model's performance gets enhanced. For methods in the third category, semi-supervised generative models put more focus on solving target tasks (e.g., classification) than just generating high-fidelity samples.

1.3 Objective

As stated previously, traditional ML-based CAD schemes are limited by the challenges of accurate lesion segmentation and representative feature computation, whereas DL-based CAD schemes are limited by the lack of large, annotated data. To address these limitations, the first part of this dissertation aims at developing new ML-based CAD schemes that can construct a comprehensive, effective radiomics-based feature pool or can avoid the need of lesion segmentation. The second part of this dissertation aims to develop DL-based schemes that are less reliant on large amounts of labeled data via a transferring generative model or by leveraging unlabeled medical images. To this end, state-of-the-art unsupervised and semi-supervised techniques will be adopted and modified to improve the CAD performance.

1.4 Organization of the Dissertation

In this dissertation, we present four novel CAD schemes for the diagnosis of different tumors (i.e., cervical and breast cancer). Two of them were developed with machine learning techniques (Chapter 2, 3) and the other two with deep learning techniques (Chapter 4, 5). In Chapter 2, a cost-effective CT image marker was developed and evaluated to predict lymph node metastasis for locally advanced cervical cancer patients. To this end, a comprehensive radiomics feature pool with 1,763 features was computed and assembled. In Chapter 3, we developed a novel CAD scheme that utilizes global mammographic image features to identify case malignancy without the need of lesion segmentation. In Chapter 4, we investigated the effectiveness of using a transferring GAN model for mammographic breast mass classification. Meanwhile, we quantitatively evaluated the quality of the generated breast mass images by GAN's generator. In Chapter 5, we verified the feasibility of using virtual adversarial training, a semi-supervised approach, to improve the performance of classifying benign and malignant breast masses from mammograms. Last, in

Chapter 6, we summarized the contributions of these four CAD schemes to provide the conclusion of this dissertation. As an important expansion of this dissertation, we discussed how to more effectively utilize state-of-the-art deep learning techniques to solve challenges in medical imaging informatics at both the technical and clinical level.

Chapter 2: Developing a New Radiomics-Based CT Image Marker to Detect Lymph Node Metastasis Among Cervical Cancer Patients

2.1 Introduction

In gynecologic oncology, cervical cancer is the third most prevalent carcinoma, which particularly affects the young females aged from 20 to 39 years [30]. In United States, it may approximately affect 13,800 women and account for 4,290 deaths in 2020 [30]. For the patients with early stage cervical cancer, radical surgery is suggested as the most effective treatment [31]. However, for the women with more advanced diseases, the effectiveness of surgery is limited, and the curative treatment is concurrent chemoradiotherapy [32]. Therefore, it is critically important to categorize the cervical cancer patients for achieving the optimal treatment efficacy.

In clinical practice, the most critical indicator for stratifying the cervical cancer patients is based on the status of the lymph node metastasis [33]. Currently, lymphadenectomy and sentinel node biopsy are the most accurate approaches to determine metastasis, as the sampled lymph node tissues will directly indicate the tumor spread under the following pathological examination. However, both of these two approaches are invasive examinations, which may cause additional pain and complications on patients. Thus, in order to avoid lymphadenectomy or biopsies, a non-invasive imaging examination method is clinically needed to detect the lymph node metastasis for treatment management. Among different types of imaging methods, CT/PET (computed tomography/positron emission tomography) examination is considered the most effective method and has been used in many academic medical centers as the standard procedure. For this method, PET indicates the metabolic activity of the target lymph node, while CT can provide the location and anatomical information. Despite the effectiveness, PET is a nuclear medicine based imaging

modality, which requires much more complicated nuclear imaging agents and operation facility. Therefore, CT/PET examination is a long-time procedure with high cost, which causes heavy financial burden on cancer patients. In addition, the high cost and complicated operation procedure of the CT/PET imaging modality also limits its accessibility to many patients, especially in the rural areas of U.S.A. or other developing countries.

As compared to CT/PET examination, CT imaging alone has the advantages of low operating cost and wide accessibility. However, the CT image interpretation is visually conducted by the radiologists, and the CT image based LN metastasis identification is not as accurate as CT/PET examination [34]. In order to address this clinical challenge, it is crucial to identify new quantitative image markers to improve the accuracy of CT image based LN metastasis detection. For this purpose, we recognize that radiomics is an emerging technology that receives intensive research focus during the last ten years to search for new quantitative image markers [35]. This new technique extracts a large amount of quantitative image features from a variety of imaging modalities (e.g., CT, MRI, etc.), based on which an image marker is generated using a machine learning based prediction model. For instance, one study extracted a total of 440 features to quantify the tumor phenotype from 1,019 lung cancer or head & neck cancer patients. Using radiomics analysis, four features were finally selected as a prognostic signature, which shows superior performance in patient survival prediction as compared to the conventional method (i.e. TNM staging) [14]. The second study utilized a total of 388 quantitative features to analyze the MRI images from 365 glioblastoma patients, and a 24-feature cluster were finally determined to generate an image marker to stratify the patients into three different survival groups (i.e. poor, intermediate, and favorable groups) [36]. Beside the survival analysis, radiomics based image marker has also been successfully utilized in many cancer related applications including the

chemotherapy response prediction [37] or benign/malignant lesions classification [38-41]. However, to the best of our knowledge, no studies have focused on applying radiomics method to develop a CT image marker to identify the LN metastasis for cervical cancer patients.

Thus, in this study, we hypothesize that the clinical information (e.g., related to shape, density, texture, etc.) contained inside the primary cervical tumor are highly associated with underlying cell biology mechanisms (e.g., cycling pathways of tumor cell), which are also closely connected to the lymph node metastasis (LNM). Thus, we can utilize the novel radiomics technology to analyze the primary tumor and uncover this useful information, which will generate a CT image marker for effectively detecting LN metastasis. To test this hypothesis, we assembled a retrospective image dataset, and applied the radiomics concept to create an initial feature pool involving a comprehensive list of 1,763 handcrafted features, which were computed from the primary cervical tumor regions depicted on CT images. Then, an optimal and low-dimensional feature vector was identified and generated by a principal component analysis (PCA) algorithm, which were utilized as the input of a logistic regression model to generate a new CT image marker for detecting LN metastasis. The details of this study [42] are presented as follows.

2.2 Materials and Methods

2.2.1 Image Dataset

In this study, the CT dataset was retrospectively collected from our university' medical center, which consist of 127 locally advanced cervical cancer patients, which were identified based on the following two inclusion criteria: 1) The patients were diagnosed of stage IIb-IVA cervical cancer of all histology types (i.e., squamous cell carcinoma, adenocarcinoma or adenosquamous carcinoma); and 2) they received systematic concurrent chemoradiotherapy in the cancer center of our university. For each patient, we collected the pre-therapy CT images, which were obtained

using GE LightSpeed VCT 64-detector or GE Discovery 600 16-detector CT machines with a standardized image acquisition protocol previously established at our medical center. The protocol is briefly described as follows: (1) X-ray power output is set at 120 kVp and a variable range from 100 to 600mA depending on patient body size. (2) The 100cc contrast agent of Isovue 370 is intravenously injected using a standard power injector with a rate at 2–3cc/sec through a 20-gauge IV needle placed at the antecubital fossa. (3) Two phases of CT scans are performed in each examination. The 1st scan phase begins 60 seconds after start of contrast injection and the 2nd (delay) phase begins 5 minutes after contrast injection. Each phase of scan takes ~4 seconds. (4) CT image scans at 5mm with an axial reconstruction to 1.25 mm, sagittal and coronal reconstructions to 2.5 mm. Among these 127 patients, 52 exhibited lymph node (LN) metastasis and the other 75 did not have LN metastasis.

2.2.2 Image Feature Computation

Before feature computation, the primary cervical tumor was visually segmented by experienced researchers and radiologists. All the tumor segmentation and feature computation were conducted on the images reconstructed from the first phase of the CT scan. Since a tumor is typically depicted on a series of CT slices, some researchers tried to compute radiomics features from multiple CT slices that depict tumor regions [14], which are defined as 3D features. Theoretically, the 3D feature should be more accurate, as it carries more information to characterize the tumor heterogeneity. However, the previous study found no significant performance difference between using 2D and 3D radiomics feature based markers [43], which can be attributed by the following factors. First, the 3D feature must be conducted based on 3D tumor segmentation, and multiple slice segmentation will enlarge the segmentation errors multiple times, which counteracts the advantages of 3D feature computation and reduces the feature

robustness. Second, the 3D tumor accuracy is heavily dependent on the axial scanning resolution of the CT images. The diversified scanning protocol is unavoidable in multiple site studies. Therefore, 3D tumor features may have high generalization errors than 2D features. Meanwhile, 2D features can have the advantages of low computation complexity and wide availability, as most of the image features are based on 2D computation. Accordingly, a total of 1,763 quantitative image features were extracted on each segmented tumor region, which can be divided into the following 3 groups: 1) Shape and density features; 2) Histogram based features; and 3) Multiscale features. All these image features characterize tumor heterogeneity in both time and frequency domains.

1) Shape and Density Feature Group

As illustrated in Table 2-1, this group includes a total of 15 features to describe the shape and density properties of tumor region. The pool of shape features includes one feature to compute the tumor area, 8 features to describe the tumor geometric irregularity, and two features to describe the tumor complexity. These irregularity features are based on shape ratio (e.g., ratio of area to perimeter length), tumor radial length (e.g., Shannon entropy of radial length), or central position shift. Specifically, the central position shift feature estimates the distance between pixel with smallest density (i.e. grayscale intensity) and the gravity center of the tumor region [44].

Tumor complexity is computed by two features: C_0 complexity and fractal dimension. C_0 complexity is computed as follows [45]: First, given a $N \times N$ gray-level image $\{f(j,k), j = 0,1,\dots,N-1, k = 0,1,\dots,N-1\}$, its 2-dimension discrete Fourier transform (DFT) is represented by $\{F(m,n), m = 0,1,\dots,N-1, n = 0,1,\dots,N-1\}$. In DFT domain, we keep all the elements with values higher than the mean square value:

$$\tilde{F}(m,n) = \begin{cases} F(m,n), if\ |F(m,n)|^2 > G \\ 0, \qquad if\ |F(m,n)|^2 \leq G \end{cases} \tag{2-1}$$

where G is the mean square of F(m, n): $G = \frac{1}{N^2}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}|F(m,n)|^2$. Next, we compute the inverse discrete Fourier transform $\tilde{f}(j,k)$ of $\tilde{F}$, and the complexity C_0 is defined as:

$$C_0 = \sum_{j=0}^{N-1}\sum_{k=0}^{N-1}|f(j,k) - \tilde{f}(j,k)|^2 \,/\, \sum_{j=0}^{N-1}\sum_{k=0}^{N-1}|f(j,k)|^2 \tag{2-2}$$

Besides C_0 complexity, fractal dimension [46] is another feature to characterize the tumor complexity [47]. For this feature, the target $N \times N$ image is first down-sampled to a $s \times s$ image, and the ratio of s to n is defined as down sampling scale: r = s/n. Next, the original image is divided into a number of sub-regions, each of which has a size of $s \times s$. For the subregion (i, j), $n_r(i,j) = l - k + 1$ if the maximum and minimum intensity value falls into the lth and kth box. Let $N_r = \sum n_r(i,j)$. Then, the fractal dimension of S is given as follows:

$$FD = \frac{N_r}{\log(1/r)} \tag{2-3}$$

In addition to the shape features, density features characterize the tumor heterogeneity by computing the statistical measures of the CT number (Hounsfield unit) within the tumor area, which includes average pixel value, standard deviation of pixels values (coherence), region contrast and Shannon entropy.

Table 2-1: A list of computed shape and density features.

Class	Description
Shape features	Tumor area, ratio of area to perimeter (roundness), ratio of outer rectangle height to width, ratio of tumor region maximal radius to minimal radius, skewness of radial length of tumor region, normalized mean radial length, standard deviation of radial length, Shannon entropy of radial length, normalized central position shift
Density features	Average pixel value in the tumor area, standard deviation of pixel values in the tumor area, region contrast, Shannon entropy of pixels values in tumor region, fractal dimension, C_0 complexity

2) Histogram Based Feature Group

a) Local Binary Pattern Feature

For this type of the feature, we first extract the eight neighbors for each pixel of the target image, and then compute the average intensity of these 9 pixels (target pixel and its eight neighbors) [48]. Next, each of the eight neighbor pixels is compared with the average intensity, and 1 or 0 will be assigned to the neighbor if its intensity is higher or lower than the average. As illustrated in Figure 2-1, these eight neighbor binary sequences generate a one-byte value ranging from 0 to 255, which will be assigned as the pattern value for the target pixel. After applying this procedure for all the pixels of the target image, a pattern image is created, for which each pixel value is the above pattern value. The histogram of this pattern image is used as a 256-value feature vector in this study.

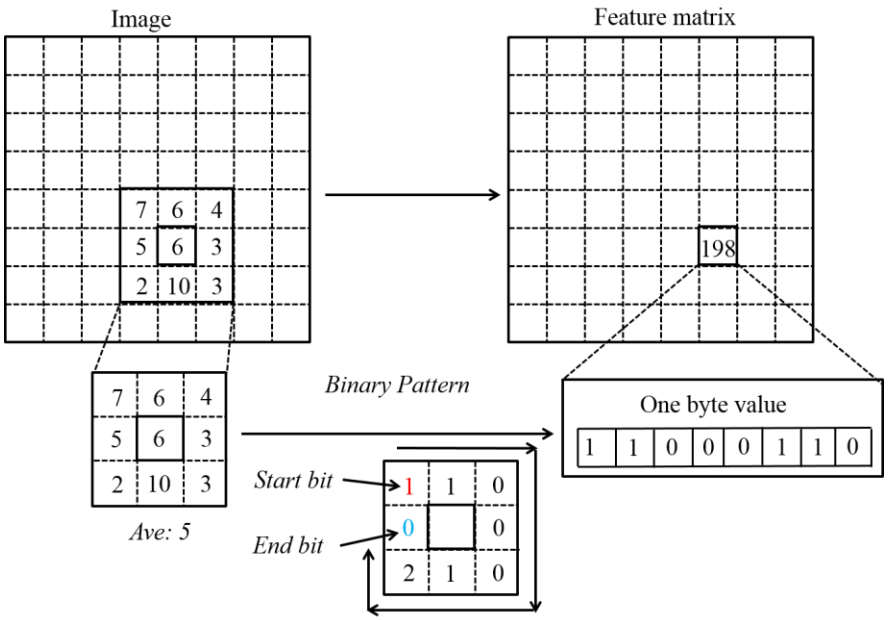


Figure 2-1: The calculation of local binary pattern feature.

b) Edge Histogram Descriptor

The edge histogram descriptor (EHD) is a useful texture signature that characterizes the spatial distribution of image edges [49], which is computed as follows. The original image is first evenly divided into 16 sub-images, each of which is then further partitioned into a certain number of image blocks. For each image block, five edge detectors are applied to detect vertical, horizontal, 45° diagonal, 135° diagonal, and isotropic edges, respectively. The image block is labeled as containing edge with one of the above orientations if the convolution between the block and the corresponding edge detectors is above the preset threshold. For each sub-image, the edges with five different orientations are computed as a histogram. Accordingly, for an image with 16 sub-images, the EHD feature is computed as a feature vector with a dimension of 80.

c) Histogram of Oriented Gradients

The histogram of oriented gradients (HOG) is a widely used feature, and the central idea is that the distribution of intensity gradients can capture the appearance and shape of local objects within an image [50]. The HOG feature extraction can be summarized as follows. First, the values of image pixels should be normalized using gamma transformation or histogram equalization. Next, the gradients at different orientations are computed. Similar to the EHD feature computation, the image is divided into a total of M blocks, while each block is divided into N cells. For each cell, the first-order differential calculation is performed along K different orientations, which are evenly ranging from 0~180 degrees, and a histogram is generated against the gradient direction. Then, the HOG feature vectors are normalized within blocks using L2 norm. After combining the HOG of all the cells together, a final feature with $K \times M \times N$ dimension is generated, where K represents the number of direction angles in each cell, M and N represent the number of blocks and the number of cells in a block, respectively. In this study, the K, M, and N are determined as 9, 20 and 4, which has been verified as an optimal balance between the feature performance and complexity.

3) Multiscale Feature Group

a) Pyramid and Wavelet Based Features

To compute this type of the feature, the original image is first decomposed using Laplacian pyramid or wavelet methods. For Laplacian pyramid method, the procedure is as follows [51]: A Gaussian kernel based low pass filter is applied on the original image, and the difference between the original and low pass filtered images is considered as the level 1 decomposition (L1). Then, the L1 image is down sampled by a scale of 2, on which the level 2 decomposition is generated using the same approach. L3 decomposition is computed similarly, and L1-L3 decompositions are used for the following feature computation.

Meanwhile, in Wavelet decomposition, the Harr wavelet transform [52] is first applied on the original image, and the resulting two-dimensional array of coefficients contains four bands of data, which are labelled as LL (low-low frequency component), HL (high-low frequency component), LH (low-high frequency component) and HH (high-high frequency component) respectively. The LL band captures the low frequency components (smooth variations) that constitute the base of an image, the HH band retains high frequency components (edges details) that can refine the image. The LL band can be further decomposed in the same manner, thereby producing more sub-bands. Similar to pyramid decomposition, L1-L3 wavelet decompositions are conducted and used in this study.

As illustrated in Figure 2-2, each decomposition is evenly divided into four regions namely, A, B, C, D. The center of region E is at the central position of the original image, and it has the same size as regions A-D. Three features are computed on original region and divided regions A-E, including 1D complexity, 2D complexity and FD. For 1D decomposition, all the pixels within the region is reorganized into a 1D array, on which the complexity is computed. Hence, a total of 108 features are generated for all Laplacian pyramid or wavelet decompositions.

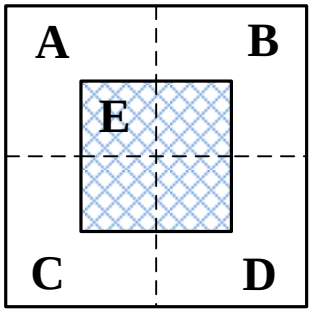


Figure 2-2: Subregion partition of the sub-band LL. The entire band is evenly divided into four regions (A-D). Region E is located in the band center, with the same size as regions A-D.

b) Gabor Features

The frequency domain features contain copious clinically meaningful information for patient stratification [37]. Among different types of frequency domain features, Gabor based wavelet transform are optimal for measuring local spatial frequencies at different scales and orientations [53]. The Gabor transform can be defined as follows [54]:

$$\mathrm{G}(\mathrm{x},\mathrm{y}) = (\frac{1}{2\pi\sigma_x\sigma_y})\exp(-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}) + 2\pi Wx) \quad (2\text{-}4)$$

In this study, we utilize a number of self-similar functions to decompose the original image, which is implemented by rotating and dilating the mother Gabor function $\mathrm{G}(\mathrm{x},\mathrm{y})$ as follows:

$$\mathrm{g}_{mn}(\mathrm{x},\mathrm{y}) = a^{-m}\mathrm{G}(\mathrm{x}',\mathrm{y}') \quad (2\text{-}5)$$

where $\mathrm{x}' = a^{-m}(\mathrm{x}\cos\theta + \mathrm{y}\sin\theta)$, and $\mathrm{y}' = a^{-m}(-\mathrm{x}\cos\theta + \mathrm{y}\sin\theta)$. The parameter $\theta = \frac{n\pi}{n}$ defines the orientation and scale of Gabor filters. For our application, there are 3 scales (m = 3) and 4 orientations (n = 4), resulting in 12 filters. For each filtered image, we calculate the mean and standard deviation. Thus, a 24-dimensional Gabor features vector can be obtained.

Similarly, using Harr wavelet transform [55], we decompose the original image into different scales. At scale L1-L3, we still estimate the mean and standard deviation (STD) on each of the four sub-bands (i.e., LL, LH, HL, HH bands). Therefore, a total of 24 features are computed. Furthermore, we use similar method to generate another 24 features for Daubechies 2 wavelet (DB2) transform [56].

c) GIST Descriptor

GIST characterizes a set of perceptual properties closely related to the dominant target structure including naturalness, openness, roughness, expansion, ruggedness [57]. Specifically, GIST captures the gradient information at different scales and orientations for different parts of an image. For this purpose, a total of 32 Gabor filters at 4 scales and 8 orientations are first applied on the given image to generate 32 feature maps. Each feature map is then divided into 4×4 equal regions. On each region the filter coefficients are averaged as one feature number. As a result, each feature map creates 16 feature values and a total of 512 features are finally computed by the 32 feature maps as the GIST descriptor.

2.2.3 Develop a Machine Learning Based Model to Predict Lymph Node Metastasis

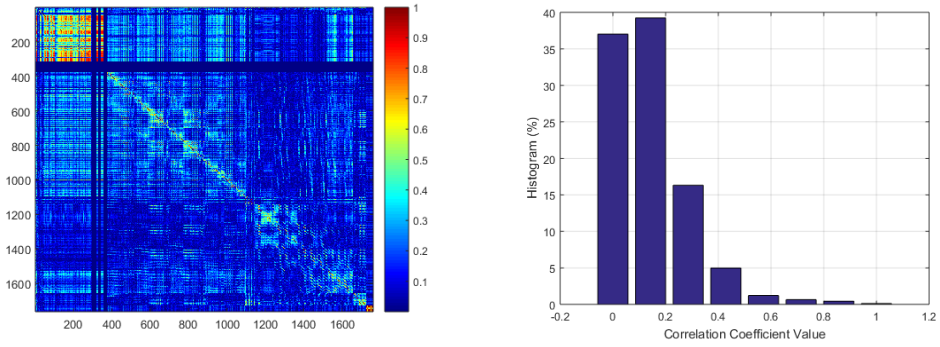
After computing all features discussed in above sections, we developed a multi-feature fusion based machine learning model to predict LN metastasis. Although many different machine learning models can be used to predict risk of LN metastasis, we selected to build two different machine learning models namely: a logistic regression (LR) [58] based and a support vector machine (SVM) [59] based prediction model, which is due to the limited size of our dataset of 127 cases. The model generates a prediction score indicating the estimated risk or likelihood of LN metastasis for each individual patient. Among different types of the machine learning based classification models, logistic regression has relatively simple architecture with strong robustness, which is easy to be optimized with high performance while avoiding the model overfitting. Given that the number of features (i.e., 1763) is much larger than the number of cases and the information redundancy cannot be avoided in the original feature set, dimension reduction is first applied on the initial feature pool to create an uncorrelated optimal feature cluster. In this study, the principle component analysis [60] is adopted to reduce the feature space dimensionality and redundancy.

For this method, let X be the feature matrix, for which each column represents one type of features, and each row represents one sample. Next, the sample co-variance matrix X^TX is computed and defined as M, on which the characteristic transform is performed: $M = W \sum W^T$. All the characteristic vectors and values are arranged with decreasing order in eigenvector matrix W and diagonal eigenvalue matrix $\sum$. Since the eigenvalues decrease rapidly, only a small number of coordinates will be able to accurately approximate the original data matrix. In this study, we only keep the first 18 components. These components will be used to build the prediction model. As a comparison, we also directly apply each class of the original features to build the model for metastasis prediction.

Next, the leave-one-case-out (LOCO) cross-validation technique is adopted to train the logistic regression model and evaluate the model performance. LOCO is essentially a special form of *k*-fold cross-validation, where the number of folds is equal to the number of training instances (e.g., *n*). As compared to the regular K-fold (e.g., K = 5) cross validation, the LOCO strategy has the least bias to optimize the prediction model, because it eliminates the potential case partition bias and increases the size of training samples to the maximum within the available dataset. Therefore, the LOCO is well-recognized as the optimal option to train and test machine learning models when the total dataset is limited [61]. Finally, the trained logistic regression model generates a prediction likelihood score ranging from 0 to 1, in which the higher score indicates the higher likelihood of LN metastasis. The model performance is assessed using the receiver operating characteristic (ROC) curve. All results are tabulated for performance comparison and analysis in which statistical significance (p-value) is also computed.

2.3 Results

Figure 2-3 (a) is the heatmap of all the 1,763 features, which is generated by a total of 127 observations. In this map, each entry represents the colorized correlation coefficient ranging from 0 (Blue) to 1 (red). The map shows that the group of LBP features has less independencies than the others. All these coefficients are categorized into six evenly distributed intervals between 0 and 1, and the results are illustrated as a histogram (Figure 2-3 (b)). The histogram reveals that more than 60% of the correlation coefficients falls into the interval ranging from 0-0.4, implying that our initial feature pool covers comprehensive tumor characteristics with small information redundancy.

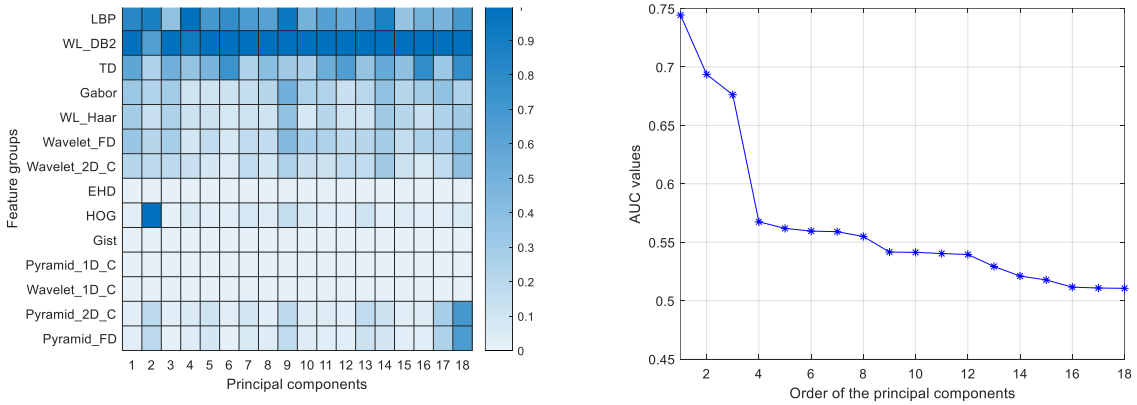


(a) (b)

Figure 2-3: a) Heatmap of all the 1763 features; b) Histogram of the feature correlation coefficients.

After applying the PCA algorithm, a total of 18 principal components (PC) are selected. Given that each component is a linear combination of all the original features, PCA coefficients can directly reflect the impact of original image features on each individual component. To investigate this impact, the coefficient distribution is generated for each component. Accordingly, the coefficients belonging to one feature group are added together, and the summed results are then normalized to create the histogram. These distribution histograms were organized as an impact

heatmap in Figure 2-4(a). In this map, horizontal and vertical direction represent feature group and principal components, respectively. Hence, each column represents the normalized coefficient distribution of one component. The map demonstrates that the Wavelet-DB2 feature group has stronger impact than the other features. For example, in component 3, the Wavelet-DB2 coefficient is 1.0, which is much larger than the second (Shape and Density, 0.514) and third (LBP, 0.368) most important group. Meanwhile, the coefficients of seven feature groups (i.e., HOG, EHD, GIST, Wavelet-1D-Complexity, Pyramid-1D-Complexity, Pyramid-2D-Complexity, and Pyramid-FD) are lower than 0.1 for most of the components, which shows that these features do not significantly contribute to the components generation. In addition, the discriminative power of each individual PC was assessed using area under the ROC curve (AUC), which are sorted in decreasing order ranging from 0.511 to 0.744 (Figure 2-4(b)).



(a) (b)

Figure 2-4: a) Feature coefficients for each principal component; b) The AUC values of the principal components. All the components were sorted with decreasing order.

Using these PCs as input, two different models (i.e., logistic regression and support vector machine) were trained and optimized to predict whether the patient has LN metastasis or not. The model performance was assessed using ROC curve, which was generated by maximum likelihood based curve fitting algorithm (ROCKIT, http://metz-roc.uchicago.edu/, University of Chicago). As demonstrated in Figure 2-5, the SVM model yielded the AUC (i.e., area under the ROC curve) value of 0.841± 0.035, while the performance of the LR model is 0.814 ± 0.037, which is slightly lower than the SVM model without a statistically significant difference (p value 0.149). Thus, both of these two models indicate high discriminative power in LN metastasis prediction. As a comparison, we also utilized the best and second best performed PC to achieve the prediction, while the AUC values of 0.710 ± 0.047, 0.662 ± 0.047 were yielded respectively. Table 2-2a lists the confusion matrix of the prediction results of the SVM prediction model. When setting the operating threshold of 0.5 on the model generated likelihood score, this model grouped 42 and 85 cases as metastasis "positive" and "negative", respectively. Among the "positive" prediction group of 42 cases, a total of 32 were confirmed as positive by CT/PET examinations, achieving a positive prediction value (PPV) of 0.762. Meanwhile, 65 out of the 85 cases in the "negative" class was confirmed by CT/PET examinations, and the corresponding negative prediction value (NPV) of 0.765. Combining both groups together, the overall prediction accuracy is 76.4% (97/127). Similarly, for the LR model (Table 2-2b), the PPV, NPV, and overall predicting accuracy are 0.673 (33/49), 0.756 (59/78), and 72.4% (92/127), respectively.

As a comparison, we also used each group of the features as the input to optimize the regression model for LN metastasis prediction, and all the results were summaried in Table 2-3. It is revealed that the AUC values ranges from 0.532 to 0.740. In five out of the fourteen groups, the AUC values are higher than 0.65, indicating that each group of image features have significantly higher

dicriminative power than random guess (AUC = 0.5) in performing this prediction task. The value of 0.65 is used as a threshold is because that most previously reported cancer prognostic markers have an minimum AUC of 0.6 [62-66]. Specially, the Shape & Density and Wavelet-DB2 feature group achieved the best and second best performance, with AUC values of 0.740±0.044 and 0.737±0.045, respectively. The results indicate that the well trained logistic regression or SVM model yields significantly higher performance (AUC value) than that of the two above single group features (e.g. Shape and Density, Wavelet-DB2). For instance, when comparing the LR model with these two single group features, the significance level of performance difference (p value) were estimated as 0.047 and 0.053, respectively.

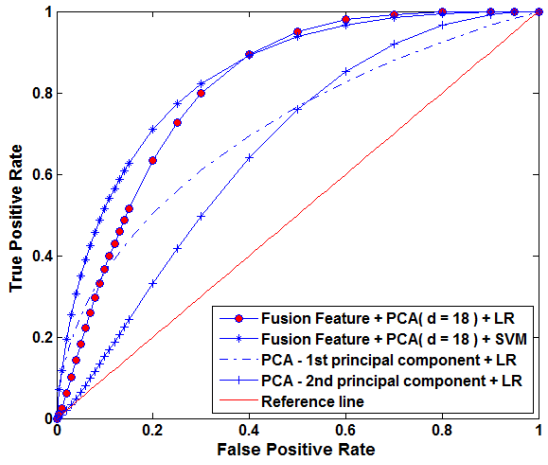


Figure 2-5: ROC curve of LN metastasis prediction results.

Table 2-2: Confusion matrix of LN metastasis prediction.

(a) SVM model

	Predicted: positive	Predicted: negative
Actual: positive	32	20
Actual: negative	10	65
Accuracy	Positive prediction value	Negative prediction value
0.764	0.762	0.765

(b) LR model

	Predicted: positive	Predicted: negative
Actual: positive	33	19
Actual: negative	16	59
Accuracy	Positive prediction value	Negative prediction value
0.724	0.673	0.756

Table 2-3: LN metastasis prediction performance: fusesd feature vs. separate feature groups.

Feature type	**AUC ± STD**
Fused feature (SVM)	**0.841 ± 0.035**
Fused feature (LR)	**0.814 ± 0.037**
Shape and Density	0.740 ± 0.044
Wavelet-DB2	0.737 ± 0.045
Pyramid-FD	0.694 ± 0.049
Wavelet-FD	0.681 ± 0.051
Wavelet-Haar	0.673 ± 0.047
Gabor	0.646 ± 0.054
EHD	0.621 ± 0.053
Pyramid-1D-Complexity	0.607 ± 0.049
Pyramid-2D-Complexity	0.603 ± 0.050
Wavelet-2D-Complexity	0.589 ± 0.050
Wavelet-1D-Complexity	0.555 ± 0.051
HOG	0.539 ± 0.057
Gist	0.533 ± 0.061
LBP	0.514 ± 0.059

2.4 Discussion

In this study, we developed and evaluated a novel and cost-effective image marker to predict LN metastasis for locally advanced cervical cancer patients. This study has several unique characteristics as follows. First, although radiomics features have been utilized to predict patient prognosis of many different cancers [14, 67-69], to the best of our knowledge, this is the first study that applies the CT image based radiomics features to predict the LN metastasis of cervical cancer patients. Currently, the standard update value (SUV) of the PET images is the most important criterion for diagnosis of LN metastasis, as the metastatic tumor activity is mainly indicted by the glucose consumption, which can be quantified by PET examination using FDG tracer (2-fluoro-2-deoxy-D-glucose). However, the model proposed in this study identified and computed a large amount of radiomic features from the primary cervical tumor depicted on the CT images. After applying the PCA algorithm to generate an optimal feature vector, a machine learning based model was optimized for predicting the LN metastasis. Using a retrospectively assembled clinical image dataset, the experimental result demonstrated that the optimized SVM model yielded an AUC value of 0.841 ± 0.035, which implies that the valuable and discriminative metastasis information or features are also contained within CT images. As compared to PET/CT examination, this CT image marker has many potential advantages of low cost, short examination time, simple imaging scanning procedure and wide accessibility.

Second, we computed and assembled a comprehensive feature pool with 1,763 features, which can be grouped into 3 major categories (14 subcategories). Considering the number of feature categories and the quantity of features, we believe that this is by far the most comprehensive feature pool designed specifically for the task of predicting LN metastasis based upon CT images. The results demonstrate that five out of the 14 groups (i.e., Shape and Density, Wavelet-DB2,

Pyramid-FD, Wavelet-FD, and Wavelet-Haar) contain significant discriminative power in classifying LN metastasis cases (i.e., AUC of 0.65 or higher), which implies the effectiveness of the initial feature pool. Although the Shape and Density group feature achieved the highest prediction performance, three of the five best performed feature groups are based on wavelet transform, which demonstrates that the frequency domain information is equally important as compared to the spatial domain. Meanwhile, two FD based feature categories yield AUC values of 0.694 and 0.681, respectively, which may reveal that the tumor pattern complexity should be an effective indicator to manifest the lymph node metastasis. Additionally, four groups of the features were performed at different scales of the segmented tumors, which implies that the clinically meaningful metastasis information may also be depicted in the detailed tumor patterns.

Third, a novel machine learning based image marker was generated by fusing 18 non-redundant radiomics feature components, which were determined by principal component analysis algorithm. As shown in Table 2-3, the fused feature gains 10% AUC value increase (SVM model) as compared to the separate feature groups. The performance improvement indicates that our prediction model successfully fuses the complementary and important metastasis information from separate feature categories. Thus, an accurate and reliable marker can be created to facilitate the diagnosis of LN metastasis. When investigating the contributions of separate feature groups in the final principal components, we observe that high performance feature groups (e.g., Shape and Density, Wavelet-DB2, Pyramid-FD, Wavelet-FD, WL-Haar) correspondingly have relatively large PCA coefficient weights. The performance of Wavelet-DB2 feature alone is only slightly lower than Shape and Density (0.737 vs 0.740), but Wavelet-DB2 exhibits significantly larger coefficients in the final components. This phenomenon shows that these two types of the features are somewhat homogeneous and Wavelet-DB2 features become dominant during the component

synthesis. Another notable observation is the LBP feature has significantly higher coefficients as compared to its relatively low discriminative power, which may be attributed to the complementary information that are missed by other high-performing features. In addition, LBP is used to identify and extract the local tumor texture, this again indicates the importance of local texture change in metastasis prediction.

Although the study results are encouraging, we also recognize that this investigation has several limitations as follows. First, the dataset only consists of 127 patients, which are selected from only one hospital with the limited diversity. We have not investigated whether the model optimized in this study can be directly applied on the CT images acquired from other medical centers, which will be obtained using different models of CT machines with the possibly different in signal-to-noise ratios (SNR). To further verify the performance and robustness of our proposed model, a more comprehensive patient cohort should be established to include the patients from multiple medical centers. Second, we designed a very large feature pool (1,763 features) for the metastasis prediction in this study, but only one linear dimension reduction method (i.e. PCA) was used to identify the effective and non-redundant feature clusters. Some non-linear algorithms, such as sequential forward feature selection (SFFS) [70] and genetic optimization [71], should also be tested and compared with the PCA to further identify the meaningful information within the initial feature pool. Third, this preliminary study only used the conventional prediction models for marker generation. The prediction performance may be further enhanced if we employ more advanced machine learning technologies. For example, the state of the art deep neural networks [20] can also be used as a fixed feature extractor to expand the existing feature pool, and the deep stacked auto-encoder (SAE) [72] can be applied for feature selection. As a result of the progress of applying deep learning in cancer image analysis [73, 74], we believe that the prediction accuracy may be

further improved by combining the radiomics and deep learning based image feature together in future studies.

Fourth, we did not investigate the reproducibility of the computed tumor features, as they may vary due to the change of image quality or noise level under different image acquiring conditions [75]. For example, the adjustment of milliamperage (mAs level) leads to the change of radiation dose, which may impact image contrast-to-noise ratios. Increasing image noise may cause tumor segmentation errors and further affect the computing accuracy of some related features. To overcome this limitation, a phantom study [76] would be a feasible approach to select the most robust features which are insensitive to the change of image acquisition parameters (e.g., mAs and kVp). Finally, although 2D image features are more robust with low generalization errors, 3D features may carry more prognostic information with higher discriminatory power. To validate this hypothesis, more research efforts are needed, which aim to 1) develop accurate segmentation algorithms to reduce the errors in multiple slice tumor segmentation; 2) select the 3D features which are resistive to the change of scanning parameters (e.g., axial scanning resolution); and 3) minimize the variance of the scanning protocols at different clinical sites.

In summary, radiomics and machine learning technologies provide powerful tools to develop new quantitative imaging markers. The scientific rigor of these new markers still needs to be further validated using more independent and comprehensive image datasets in the future.

Chapter 3: Applying a New Quantitative Image Analysis Scheme Based on Global Mammographic Features to Assist Diagnosis of Breast Cancer

3.1 Introduction

Since breast lesions are highly heterogeneous containing overlapped dense fibro-glandular tissues, reading and interpreting mammograms is a difficult task for radiologists [2, 77]. Accordingly, developing CAD schemes of mammograms have attracted extensive research interest in the recent decades, which aims to provide radiologists a "second opinion" supporting tool in reading and interpreting mammograms [78]. Currently, there are two types of CAD schemes namely, computer-aided detection (CADe) schemes and computer-aided diagnosis (CADx) schemes. The former detects suspicious lesions and determines their locations in mammograms [79], while in contrast the latter makes classification between malignant and benign lesions [80]. Although commercialized CADe systems are currently available and used in the clinical practice, it is in controversy of whether using these schemes can actually improve radiologists' performance in detecting breast cancer [81]. Despite the somehow disappointment when using CAD schemes of mammograms in the clinical practice, many researchers believe it is necessary to continue exploring new approaches to improve CAD performance and optimize CAD in the current clinical practice [82].

The performance of CAD schemes for mammograms heavily depends on case difficulty, including the conspicuity of lesions (e.g., fuzziness of lesion boundary) and overlap of dense and heterogeneous fibro-glandular tissues [83]. These difficulties can substantially reduce the accuracy and robustness of lesion segmentation [84], which will affect the performance and reproducibility

of CAD schemes [83, 85]. Previous studies demonstrated that improving lesion segmentation accuracy can produce more reliable image features, which helps achieve better CAD performance in classifying between malignant and benign breast lesions [39, 86]. Nonetheless, accurate lesion segmentation is still challenging for digital mammogram processing due to the high heterogeneity of lesions and overlapping of dense fibro-glandular breast tissues on mammograms [77, 87]. In addition, due to the lack of ground truth, evaluating lesion segmentation accuracy is also very difficult or subjective with large inter-reader variability. Thus, in order to avoid this challenge, different approaches have been proposed to develop new CAD schemes without lesion segmentation including using deep learning algorithms [88]. However, for the deep learning based methods, training a reliable deep learning based CAD schemes usually requires a large and diverse image dataset (e.g., using a dataset with 128,175 retinal images [89]), which is often unavailable in cancer imaging field.

In order to overcome these limitations, our recent studies indicate that quantitative image markers extracted from the whole breast areas of mammograms (namely, the global features) can be used to predict short-term breast cancer risk with significantly higher prediction power [90, 91]. Thus, we hypothesized in this study that it is possible to identify and fuse the global image features computed from the whole breast area depicting on mammograms without lesion segmentation, which enables to generate a new quantitative imaging marker for predicting the likelihood of a testing case being malignant. This new global mammographic image feature-based approach cannot only avoid lesion segmentation, but also reduce the requirement of large training dataset as the conventional deep learning approach [90]. Thus. the objective of this study is to develop a new global mammographic image feature analysis-based CAD scheme and validate our study hypothesis. The experimental details are presented as follows [92].

3.2 Materials and Methods

3.2.1 Image Dataset

From the IRB-approved retrospective study protocols, we have assembled a digital mammography image database as reported in our previous studies (e.g., [79, 85, 86]). From the assembled image database, we selected an image dataset for this study, which consists of fully anonymized digital mammograms acquired from 275 women participants in breast cancer screening. Each case has one suspicious mass-type lesion identified and detected by the radiologists in original mammogram reading and interpretation. All suspicious lesions were biopsied and confirmed by histopathology examinations. Among these cases, 134 were confirmed to be malignant, while the other 141 cases were benign. In addition, cancer was detected only in one breast in this dataset.

All digital mammography screening examinations were performed using Hologic Selenia (Hologic Inc) full-field digital mammography (FFDM) systems. Each mammography screening case has four images including two cranio-caudal (CC) and two medio-lateral oblique (MLO) view images of left and right breasts. Since this study only focused on the cases depicting soft tissue mass type lesions, the mammograms were subsampled to reduce the image size, which is a common practice used in CAD research field including the commercialized CAD schemes [91]. Specifically, as reported in our previous computerized scheme [79], the original FFDM images with a pixel size of 0.07mm were pre-subsampled using the average pixel value computed from a 5×5 scanning window. Thus, the actual pixel size used in the subsampled image is 0.35mm.

As summarized in Table 3-1, the mammographic density information of malignant and benign cases was identified by radiologists according to BIRADS guidelines, which shows no significant difference between two classes of malignant and benign cases using the BIRADS based mammographic density ratings.

Table 3-1: Distribution of mammographic density (BIRADS ratings) for two groups of the cases in dataset.

Characteristic	Malignant cases	Benign cases
Fatty tissue (1)	6	7
Scattered (2)	55	57
Heterogenous (3)	70	73
Extremely dense (4)	3	4

Figure 3-1 illustrates the example images of one malignant case and one benign case, each of which contains four CC and MLO view images of the left and right breasts. Lesions of the example cases exhibit low conspicuity and high fuzziness, making it difficult for accurate lesion segmentation and risk prediction.

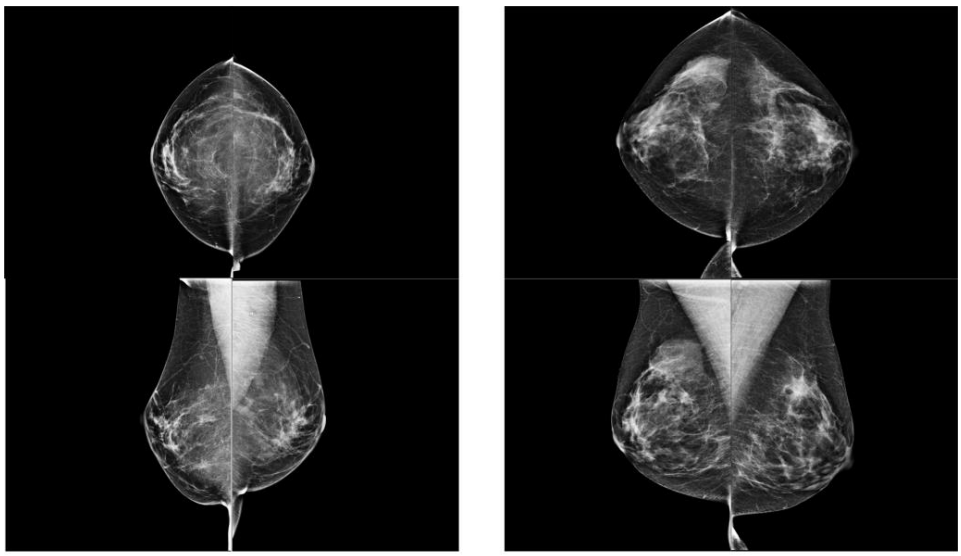


Figure 3-1: Examples of malignant and benign cases with four view display.

a) Malignant case; b) Benign case.

3.2.2 A New CAD Scheme

Our proposed CAD scheme was developed in the following three steps, namely feature computation, feature selection and case classification. We first built an initial feature pool containing four different groups of features. Next, a particle swarm optimization (PSO) algorithm was applied to select optimal features so that redundant features can be removed from the feature pool. Finally, a popular machine learning classifier namely support vector machine (SVM) was used to predict the risk or likelihood of a case being malignant. The CAD scheme was implemented in MATLAB software environment.

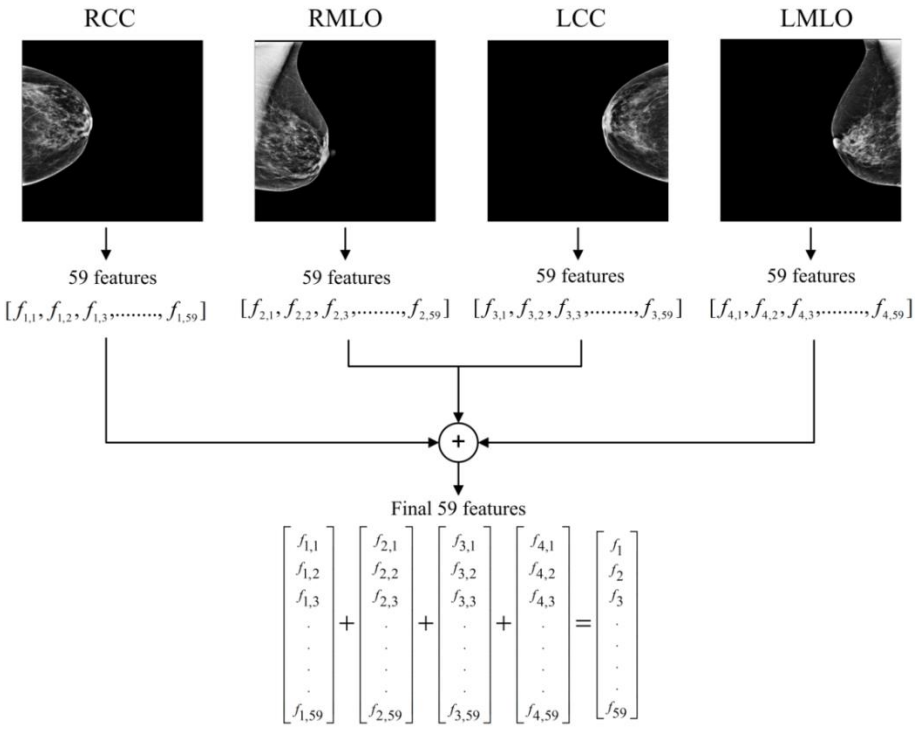


Figure 3-2: Feature extraction considering four-view images of breast area.

1) Feature Computation

As shown in Figure 3-2, we used four images (namely, RCC, LCC, RMLO, and LMLO) of CC and MLO view at the left (L) side and right (R) side. Before building feature pools, our CAD scheme first segmented the whole breast area out in each image by removing all possible artifacts or markers outside the breast areas (assigning all background pixels to 0 as shown in black color in Figure 3-2). The computed image features from the segmented whole breast area can be categorized into four groups. The first group includes 17 statistical image features describing breast area shape and density distribution (as shown in Table 3-2). These statistical features are widely used to quantify pixel value distribution and its heterogeneity in the 2D image [93, 94]. The other three feature groups are block-based Fast Fourier Transform (FFT) features, Discrete Cosine Transform (DCT) features and Wavelet Transform (WT) features, which aim to help detect

and analyze local breast tissue distributions in an image. For this purpose, each target image was first divided into blocks with a scanning window size of 8×8 or 9×9, which is determined based on our previous investigation [95]. Then, FFT, DCT and WT were applied on these blocks to construct three feature matrices, in each of which 14 features were computed (Table 3-2). More details about feature extraction can be found in the supplemental materials. Finally, three groups of FFT, DCT and WT based features (3×14=42 features) and Shape&Density group (17 features) were combined, so that CAD scheme computed 59 initial features from each image.

Next, two feature pools were built with different view images. The first one used all the four images (namely, RCC, LCC, RMLO, and LMLO) of CC and MLO view at the left (L) breast and right (R) breast (as shown in Figure 3-2). In each case, 59 previously mentioned image features were computed separately from each of the four-view images. We can organize each 59 features into a separate vector, so that each case has four feature vectors. A final vector (e.g., feature pool) was generated by adding corresponding or matched features computed from four-view images together. Such feature generating process is demonstrated in Figure 3-2. Similar to computing the first feature vector with four-view images, the second pool includes feature vectors that were computed using only two positive view images (e.g., LCC and LMLO view images of one breast). Two images of negative breast were ignored. Finally, these two feature pools and vectors were applied as input to train two machine learning classifiers embedded with the feature selection algorithm separately.

Table 3-2: List of four feature groups.

Feature Group	Description
Shape&Density	Mean, Std, Convexity, MeanGradient, StdGradient, Skewness, Kurtosis, Energy, Entropy, Max, Min, Median, Range, RMS, MeanDeviation, Uniformity, Correlation
Block-based features (FFT, DCT, and WT groups)	Mean, Std, Max, Min, Median, Range, RMS, Energy, Entropy, Skewness, Kurtosis, MeanDeviation, Uniformity, Correlation

2) Machine Learning Classifier and Performance Assessment

Using the created feature vector, a machine learning classifier is applied to generate the optimal feature cluster and predict the likelihood of the case being malignant. Although many different machine learning classifiers can be used for this purpose [96], SVM classifier uses a constructive machine learning process based on the statistical learning theory to minimize the generalization error [97], which is considered a quite robust classifier applied to the relatively small training datasets and has been used in many biomedical engineering applications [98, 99]. As a result, we adopted SVM classifier in our application, which is built based on the SVM tool box under MATLAB environment.

Based on our dataset, we trained and tested the SVM classifier embedded with the PSO feature selection algorithm to minimize the potential bias during feature selection and lesion classification. In addition, the 10-fold cross validation was performed. Specifically, our dataset was randomly

divided into 10-folds. As illustrated in Figure 3-3, SVM was trained with 9 folds of data and tested with the remaining one-fold of data. The process was repeated 10 times. In each repetition, the training dataset was used to identify an optimal feature vector using PSO algorithm, which is detailed in the supplemental materials. The following objective function is adopted to control the training outcome [18, 95]:

$$F_{obj} = \alpha[1 - Mean(AUC)] + \beta Std(AUC) + \gamma(number\ of\ selected\ features) \quad (3\text{-}1)$$

In the above equation, the parameters α, β, γ are weighted coefficients determined based on the feature distribution in the Euclidean space, and *AUC* is the computed area under a receiver operating characteristics (ROC) curve [100]. When the objective function reaches its minimum, the training process is finished. Then the trained classifier is applied to make prediction for individual case in the testing fold. Thus, through the 10 training and testing iteration cycles, each of 275 cases in our dataset will be independently tested once and receive a classification score indicating the likelihood of the case being malignant. Finally, based on the classification scores of all 275 cases, the performance of the proposed CAD scheme will be evaluated and compared using AUC and other evaluation indices (e.g., classification sensitivity, specificity, positive and negative predictive values).

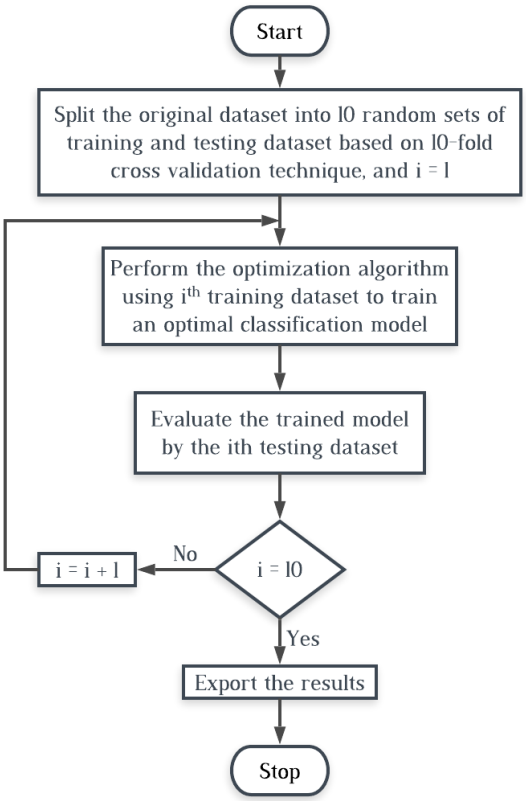


Figure 3-3: Flowchart of proposed 10-fold cross-validation based training and testing method.

3.3 Experiments and Results

3.3.1 Evaluation of Single Features

Figure 3-4 demonstrates the Pearson correlation coefficients of all 59 initially computed features. For the two-view or four-view images of each case, the value of correlation coefficients falls into eight categories as illustrated in the histogram charts of Figure 3-4. The charts show that more than 70% and 40% of the absolute correlation coefficients were smaller than 0.4 and 0.2, respectively, which indicates that the feature pool designed in our study provided a comprehensive view of the cases with a relatively small redundancy.

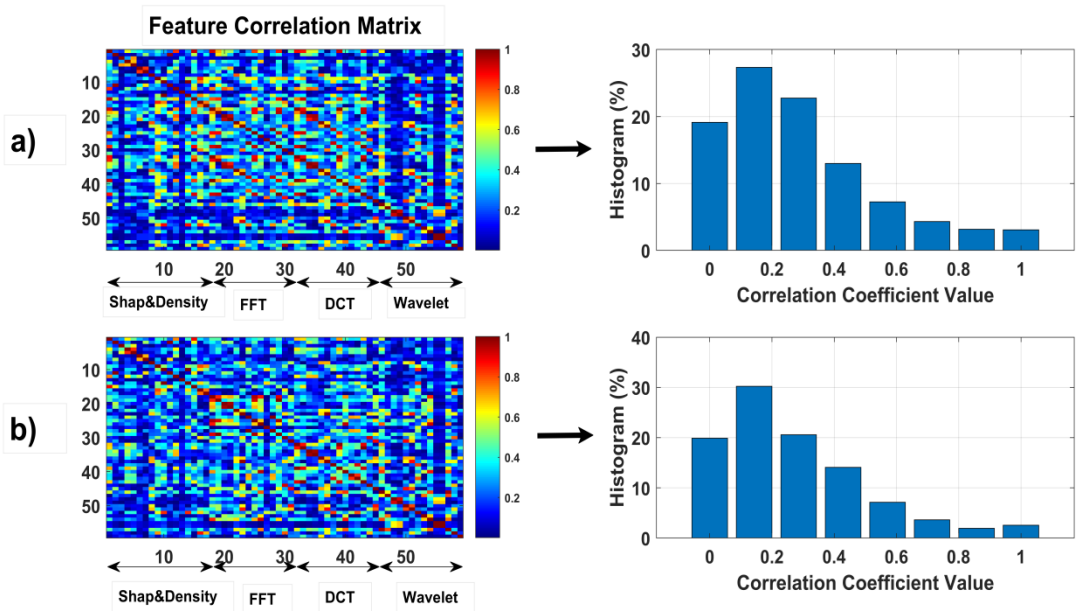


Figure 3-4: Feature correlation coefficients analysis of using a) two-view images and b) four-view images.

Figure 3-5 shows the sorted results of the areas under ROC curves (AUC values) computed from all 59 individual image features, and Table 3-3 summarizes the ten best-performed image features selected from two feature pools including the features computed from the two-view and four-view images, respectively. When using features computed from two-view images of the breast with suspicious lesion detected, the top three best performed features including MeanGradient, MeanDeviation_DCT, and Mean_FFT with the AUC values of 0.678±0.042, 0.668±0.042 and 0.665±0.042, respectively. Similarly, among all features computed from four-view images of two breasts, the top three features are Energy_FFT, Energy_DCT, and Mean_Density with the AUC values of 0.689±0.041, 0.668±0.042 and 0.667±0.042, respectively. Among the top 10 features, there are six common features selected from both two-view and four-view image feature pools, namely MeanGradient, MeanDeviation_DCT, Energy_DCT, StdGradient, Energy_FFT, and Mean_DCT. This indicates that there is a common base to support

classification between malignant and benign cases using the global mammographic image features. However, using two-view and four-view images can also make difference because adding two images of the negative breast may dilate the overall case-based difference between malignant and benign cases. As a result, there is also difference of the top 10 performed image features in two feature pools.

Table 3-3: Ten best performed features for two-view and four-view image prediction.

Top features for two-view image prediction	Top features for four-view image prediction
MeanGradient	Energy_FFT
MeanDeviation_DCT	Energy_DCT
Mean_FFT	Mean_Density
Energy_DCT	RMS
StdGradient	Convexity
Energy_FFT	MeanDeviation_DCT
Mean_Wavelet	Mean_DCT
Mean	MeanGradient
RMS_DCT	StdGradient
Mean_DCT	Entropy_DCT

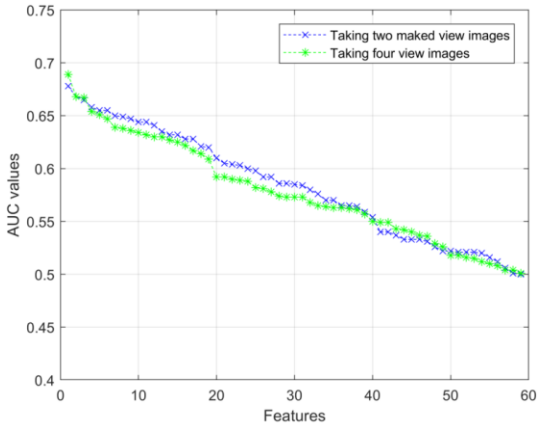
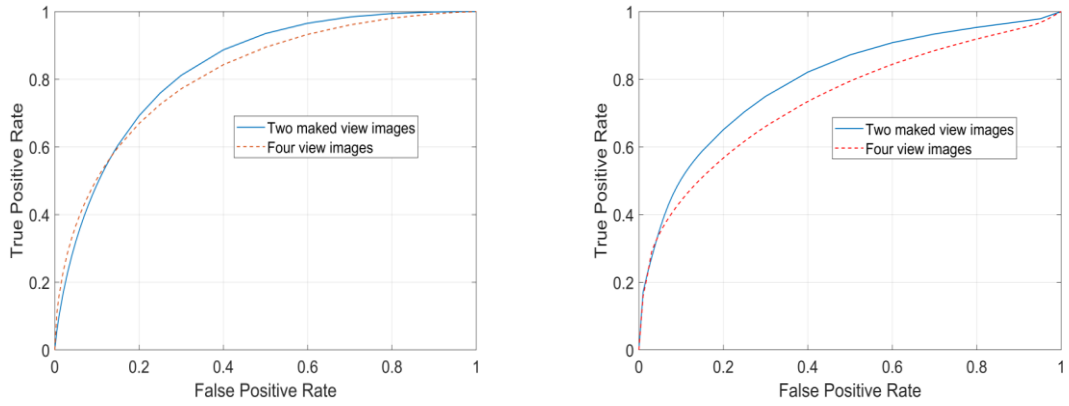


Figure 3-5: The graph of sorting list of AUC values of 59 individual features.

3.3.2 Performance Assessment of the SVM Classifiers

Using classification scores generated by the SVM classifiers on the total 275 cases, we conducted data analysis to assess the SVM classifier's performance with respect to the training dataset and test dataset. When using the training dataset, Figure 3-6 (a) shows two ROC curves for two-view and four-view images, which yields AUC values of 0.81±0.036 and 0.79±0.038 ($p = 0.057$), respectively. On the testing dataset (Figure 3-6 (b)), the AUC values are 0.79±0.07 and 0.75±0.08 for the two-view and four-view images ($p < 0.01$). The comparable AUC values between training and testing results indicated that the SVM classifiers were not over-trained and quite robust.



(a) (b)

Figure 3-6: Comparison of two ROC curves generated using (a) training dataset and (b) testing dataset.

After applying an operating threshold (T = 0.5) to divide the test cases into two predicted classes being malignant or benign, we generated two confusion matrices of testing results, which are presented in Table 3-4. Based on the two confusion matrices, we further computed other assessment indices such as classification sensitivity, specificity, positive and negative predictive values, and odds ratio, which were summarized in Table 3-5. The results show that the SVM classifier optimized with features computed from two-view images yielded better classification performance than the classifier optimized with four-view image features.

Table 3-4: Two confusion matrices of applying two SVMs to classify malignant and benign cases.

	Two-view images		Four-view images	
Actual / Prediction	Malignant	Benign	Malignant	Benign
Malignant	103	27	102	37
Benign	31	114	32	104

Table 3-5: Summary of other assessment indices of two SVM classifiers optimized using two-view and four-view images.

	Sensitivity	Specificity	PPV	NPV	Odds Ratio
Two-view	81%	77%	79%	79%	14.02
Four-view	74%	76%	76%	73%	8.96

3.4 Discussion

Although several imaging modalities have been tested and/or applied as breast cancer screening tools [101], mammography remains the most cost-effective and widely used tool for the population based breast cancer screening. However, a large amount of suspicious breast lesions (particularly the soft tissue masses) can be detected in mammograms and the majority of them are benign. Thus, exploring new approach to develop more effective CAD schemes to assist the classification between malignant and benign cases or lesions depicting on mammograms is crucial to improve the efficacy of the breast cancer screening and diagnosis [82]. In this study, we developed a novel CAD scheme utilizing the global mammographic image features to predict likelihood of the testing cases being malignant without lesion segmentation. As compared to the previously reported research efforts, our investigation has a number of unique characteristics or new observations as follows:

First, instead of computing image features from the segmented lesion and its surrounding area, the new CAD scheme extracts and computes the global image features from the whole breast areas of mammograms in this study. A majority of previous studies of CAD-based breast lesion classification [102] computed image features from the segmented lesions or its neighbors, which may have advantages and disadvantages. The advantages include enabling to compute features that are more focused and/or relevant to the specific lesions, while the disadvantages include the variable or lower accuracy and/or reproducibility of computing the features due to the difficulty and errors in lesion segmentation. In our investigation, when we use the global features extracted from the two-view positive images of one breast with suspicious lesion detected, the trained SVM classifier yielded a comparable performance (i.e., AUC 0.79±0.07) in classification between the malignant and benign cases. Although we cannot directly compare the classification performance

(AUC value) of our new scheme with previously reported CAD schemes due to the use of different image datasets, we believe that our new case-based CAD scheme yielded encouraging and comparable performance as many of previously developed lesion-based schemes (i.e., AUC values ranging from 0.70 to 0.86 as presented in a review table of a previous paper [86, 98]). The new study result indicates that the clinically meaningful information is not only focused on the lesion, but also distributes on the entire breast area of mammogram image. In addition, although CAD schemes without lesion segmentation have been previously developed and reported in the literature (i.e., [85, 88]), these schemes computed image features from a fixed region of interest (ROI) covering the suspicious lesions, which have disadvantages or difficult to adaptively identify the optimal size of the ROIs to cover the lesions with varying size and shape. The approach in this study is different. Thus, to the best of our knowledge, this is the first study that investigate the feasibility of developing a global breast image feature-based CAD scheme to classify between malignant and benign mammographic cases, which avoid difficulty in both segmentation of the lesions and determination of the optimal ROIs, which are the two popular approaches used in previous studies.

Second, we trained and tested two SVM classifiers using two feature pools containing the global images features computed from two-view images of one positive breast and four-view images of two breasts. The testing results show that the SVM classifier yielded AUC of 0.79±0.07 when two-view images of one positive breast were involved in the training and testing process. However, when using the image features computed from four-view images of two breasts to build the SVM classifier, the scheme yielded a reduced performance with an AUC of 0.75±0.08, which implies that the discriminatory information or power may be diluted when adding two negative images of one cancer-free breast. Thus, it should be better to use two-view images of one breast to

train the SVM classifier. Then, CAD scheme can be applied to two-view images of left and right breasts separately. The higher classification score should be selected to represent the likelihood of the testing case being malignant.

Third, unlike many previously developed CAD schemes that focus on computing the morphological and density distribution based features in the spatial domain, we computed image features in both spatial domain (Shape&Density group) and frequency domain (FFT, DCT, Wavelet block-based groups). As shown in Table 3-3, the top 10 performed image features in two feature pools including the features computed from two-view and four-view images contain image features computed in both spatial and frequency domains. For example, among the six common features computed from both two-view and four-view images, two are spatial domain features (i.e., MeanGradient, StdGradient) and four are frequency domain features (i.e., MeanDeviation_DCT, Energy_DCT, Energy_FFT, Mean_DCT). This result shows that the copious lesion pattern information exists in both spatial and frequency domain, which was also indicted in our previous investigation of assessing response of the metastatic tumors to chemotherapy using CT images [103] and verified in this study for classification between malignant and benign mammographic image cases. In addition, we also observed that the top three features are totally different between two-view and four-view image predictions: The top three features computed from two-view images of positive breasts with the suspicious lesion detected are MeanGradient, MeanDeviation_DCT, and Mean_FFT, and the top three features computed from four-view images of two breasts are Energy_FFT, Energy_DCT and Mean_Density. This difference may be due to the nature of the two-view and four-view images. As verified in this study, the normal tissues on the mammogram also contain clinically descriptive information for mass classification. However, the normal and abnormal tissues depict significantly different properties on the mammograms,

thus different types of features are needed to identify and collect the relevantly useful information from both normal and abnormal the tissue structures. Since two view images contain positive masses, the top features should have a balanced capability to collect the discriminative information from both the masses and the normal tissues. For the four view images, given that two images are completely normal, the selected features should have a better capability to extract the discriminant characters from the normal tissues.

Fourth, since identifying optimal and non-redundant images features is one of the most important and challenged tasks in developing the conventional machine learning classifiers including the SVM classifier [97], we in this study investigated advantages of applying a PSO method to select optimal features and guide the process of training SVM classifier. The results demonstrated that this method enabled to identify and combine the useful pattern inside the global mammographic images features while removing the redundancy. Thus, using this optimal feature selection method, we are able to use a relatively small dataset of 275 cases for the SVM model training and optimization, which avoids the large database requirement when developing the deep learning based CAD schemes.

Despite the encouraging results, we recognize that this study has the following limitations. First, the database established in this investigation was only from one institution with a limited number of cases; therefore, a more diversified dataset including the cases from more than one institution would be desirable to further test the reliability and robustness of our proposed scheme. Second, the initial feature pool with 59 features used in this study may not be an optimal feature pool. We should expand the feature pool using a list of functional and diversified features summarized in the literature [104]. Meanwhile, it is also worth investigating different feature selection methods such as Relief [105], recursive feature elimination [106], variable ranking

techniques [107], supervised training [108] or combining them with our PSO-SVM strategy. Third, this preliminary study used the supervised SVM classifiers that have strengths of solving complex problems and adapting well to high dimensional data (or feature vector); however, there may exist necessities to explore other effective classifiers (e.g. linear discriminant analysis (LDA) and artificial neural networks (ANNs)), in particular the fusion of multiple classifiers to loosen the data size requirement [109] and balance the computational efforts to resolve the uncertainty of the model [103]. Last, similar to our previous study, which demonstrated that fusion of the complementary information of global (case-based) and regional (lesion-based) mammographic image features had potential to significantly improve CAD performance in detecting suspicious lesions without increase of false-positive rates [108], we will investigate the optimal fusion method to fuse the classification results of this global feature based scheme with the lesion-based scheme [109] to more accurately classify malignant and benign mammographic cases in the future.

Chapter 4: Breast Mass Classification Using Transferring GAN with a Supervised Loss

4.1 Introduction

During the past two decades, the development of CAD schemes has attracted extensive research interests and efforts, which aim to analyze mammograms accurately and efficiently [39, 110, 111]. Given a small dataset (e.g., several hundred images), a conventional CAD scheme can be developed, which includes the following three steps: 1) design hand-engineered features [42], 2) extract features with or without lesion segmentation [92, 112], and 3) train machine learning classifiers such as SVM. If a relatively larger dataset is accessible, deep learning can be used to achieve better results in breast cancer detection [113] and malignancy identification [114].

During the training of the deep learning models, data augmentation is widely exploited to improve models' performance and robustness [115, 116]. Among various data augmentation techniques, GAN has been commonly used to generate samples of high fidelity [117, 118]. Nonetheless, there exist two challenges when we use GAN toward improving medical image classification. First, in practice, the performance improvement often turns out to be trivial if the generated images are directly included into classifiers' training. For example, a class-conditional GAN was employed to synthesize realistic lesions so that the training dataset was expanded [119]. Despite with a larger dataset, the subsequent classification accuracy was only 0.9% higher than that from classifiers trained with traditional data augmentation. The enhancement does not seem that significant if we consider the vast amounts of efforts invested into GAN's training. Second, it is difficult to make GAN converge with limited data because a stable high-performed GAN

typically requires hundreds of thousands of training images [120]. Using GAN for data augmentation will be trapped in a paradox if there are scanty images.

To address these challenges, we propose a novel approach that adjusts GAN for mammographic mass classification in a limited-data setting. To make the generated images useful to the breast mass classification task, we modified the loss function of the discriminator to make it serve as the classifier for identifying malignant masses from benign masses. The insignificant performance improvement could be attributed to the fact that there lacks a direct link between GAN and the classifier [121, 122]. In other words, the generator emphasizes on generating authentic images, but it does not have any feedback from the classifier to update itself to ensure the classifier will benefit from the generated images. In addition, we adopted transfer learning to stabilize GAN's training and expedite the generator's convergence in a limited dataset [123]. To the best of our knowledge, this is the first application that uses transferring GAN for breast lesion classification with limited data.

4.2 Materials and Methods

4.2.1 Image Dataset

The dataset used in this study consists of 512 benign masses and 512 malignant masses. To obtain our dataset, we selected mammograms with detected suspicious lesion(s) from the previously established full-field digital mammography (FFDM) image database [79, 85, 86]. All suspicious lesions went through biopsy, and the information were all summarized in the histopathology examination reports, including malignancy findings, lesion locations, and lesion sizes. We retrieved 1024 lesions in total, and cropped 1024 regions of interest (ROIs), which were scaled to be of the same size 128 ×128. Besides, to pretrain our model, we established a source

dataset with 25,000 patch images also of size 128×128 that were randomly cropped from mammograms in our database.

4.2.2 GAN with a Supervised Loss for Classification

The vanilla GAN [124] consists of two components: a generator and a discriminator. The generator takes as input noise vectors and generate fake images; the generated images and real images are together sent into the discriminator, which outputs the likelihood of the input image being real or fake. Through adversarial training, the generator can synthesize more and more realistic samples to fool discriminator into thinking the generated images are real while discriminator is optimized to separate fake images from real images. The loss function of the vanilla GAN is

$$L_G = -E_{x\sim P_G}\left[log\left(D(x)\right)\right] \tag{4-1}$$

$$L_D = -E_{x\sim P_{data}}\left[log\left(D(x)\right)\right] - E_{x\sim P_G}\left[log\left(1 - D(x)\right)\right] \tag{4-2}$$

Following the technique of adding supervised loss to the discriminator for semi-supervised learning [125], we modified the vanilla GAN to transform its discriminator for mass classification. Suppose we need a classifier to categorize an image x as one of K classes, so the output of the classifier is a K-dimensional probability vector. However, in the vanilla GAN, the output of discriminator is a scalar. To make the discriminator perform the additional classification task, we changed the number of the discriminator's output units to be K+1. Thus, the objective function of the discriminator is comprised of a supervised loss and an unsupervised loss. The unsupervised loss is the same as that for discriminator in the vanilla GAN. In this study, since the target task is a binary classification problem, the output of discriminator is a 3-dimensional vector. Considering overparameterization [125], we set the discriminator' output to be 2- dimensional.

$$L_D = L_{supervised} + L_{unsupervised} \tag{4-3}$$

$$L_{supervised} = -E_{\boldsymbol{x},y\sim P_{data(x,y)}}[logP_D(y|\boldsymbol{x}, y < K+1)] \tag{4-4}$$

$$L_{unsupervised} = -E_{\boldsymbol{x}\sim P_{data(x)}}\big[log\big(1 - P_D(y = K+1|\boldsymbol{x})\big)\big] \tag{4-5}$$

$$- E_{\boldsymbol{x}\sim P_G}\big[log\big(P_D(y = K+1|\boldsymbol{x})\big)\big]$$

In our experiments, we convert the loss function of the modified GAN into the Wasserstein GAN (WGAN) loss with a gradient penalty (GP) term [126]. The WGAN-GP loss has been shown effective in stabilizing training and alleviating the mode collapse issue. It is worth noting that 1) the WGAN-GP uses logits rather than the log likelihood to compute loss functions, and 2) the training procedure of WGAN-GP is a bit different from vanilla GAN, where discriminator is updated m times, and G updated once. We set m = 2 and the GP coefficient to be 3 in this study.

4.2.3 Training GAN with Transfer Learning

Although the modified GAN with a supervised loss performs well in semi-supervised learning [125], its performance in classifying benign and malignant masses degrades severely in this study, as our dataset is very small containing only 768 images. When the small dataset was used for GAN's training, we observed the mode collapse issue [125, 127], for which the generator kept producing similar mass images with low variety. To circumvent the requirement of a large dataset, we employed transfer learning to train the modified GAN [128]. The underlying mechanism is that mammogram patches in the source resemble masses in the target domain regarding texture, style, etc., so pretraining GAN in the source dataset could help accelerate its convergence in the target dataset.

We adopted the DCGAN architecture in this study [120]. In the pretraining stage, it should be noted that the discriminator outputs a single logit because the mammogram patches are unlabeled. After 5000 epochs' pretraining, the generator and the discriminator were both transferred to the mass dataset for fine-tuning. At the finetuning stage, the generator's structure remained the same, but the last several layers of discriminator were changed to integrate the supervised loss. Specifically, a global pooling layer was appended after the last batch normalization layer, and the output dimension of the dense layer was increased from 1 to 2 so that the discriminator can serve as a classifier. The detailed architecture description of the discriminator is summarized and presented in Table 4-1.

Table 4-1: Structure details of the discriminator network. Block 1, 2, 3, 4 have the same type of layers of the same order except that the number of kernels is different.

Layer Type	**Output Shape**
Input layer	(128, 128, 1)
Block 1: Conv2D Batch Normalization LeakyReLU Dropout	(64, 64, 16)
Block 2	(32, 32, 32)
Block 3	(16, 16, 64)

Block 4	(8, 8, 128)
Global Average Pooling	128
Dense layer	2

4.2.4 Performance Evaluation

To assess the performance, 75% of the total mass ROIs were used for training and the rest 25% for testing. We randomly divided the dataset three times and run the experiments for each division accordingly. The results of the three individual experiments were averaged for performance comparison. At the fine-tuning stage, we evaluated the quality of the generated images using the Fréchet inception distance [129] (FID) that compares the distributions of real images and fake images. A smaller FID score indicates a better quality of the generated images. To evaluate the efficacy of the proposed method, we also implemented two classifiers sharing almost the same network design as the discriminator that were trained with and without conventional data augmentation (e.g., rotation). The classification accuracy and the area under receiver operating characteristic (ROC) curve (AUC) were used to assess the performance of different methods.

4.3 Results

Figure 4-1 demonstrates how the quality of the generated images (i.e., FID score) changes as fine-tuning continues. The score starts with a value higher than 100 but drops quickly at the beginning of fine-tuning. After 25,000 epochs, the FID score fluctuates within 45 and 60, suggesting that the quality of generated images remains stable, and the generator is well-trained.

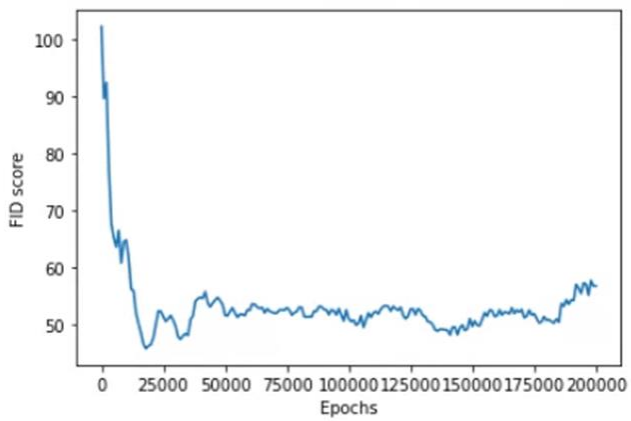


Figure 4-1: The FID score curve of the generated images.

Meanwhile, Figure 4-2 shows the synthesized mass ROIs when the generator is trained at different epochs. These generated ROIs manifest that the dense mass-like objects are mostly located in the middle part, surrounded by normal tissues with less density. As the training epoch increases, we can observe some consecutive smooth changes regarding brightness, boundary details, tissue density, etc., but the overall mass shape and style basically remain the same.

Figure 4-2: Generated images at a) 25,000 b) 50,000 c) 75,000 d) 100,000 e) 125,000 f) 150,000 g) 175,000 h) 200,000 epochs. The FID scores are a) 52.36 b) 53.53 c) 52.53 d) 52.59 e) 51.50 f) 51.07 g) 51.74 h) 56.79 accordingly.

As for the discriminator, we tested its performance every 50 epochs. The accuracy and AUC curves from one of the three experiments are provided in Figure 4-3. The testing accuracy changes drastically at first and then reached two peaks around epochs 70,000 and 125,000. However, the classification performance of the discriminator deteriorates as training continues, which can be attributed by model overfitting. This suggests that the adversarial learning needs to be stopped at appropriate epochs to ensure an optimal performance of the discriminator. Table 4-2 summarizes the average classification accuracy of the best-performed discriminator at three individual

experiments. Compared with the similar classifier trained without data augmentation, our method achieved an average improvement by 6.002% and 3.440% with respect to classification accuracy and AUC. We further compared our approach with the classifier trained with conventional data augmentation, and the testing accuracy and AUC were improved by 1.407% and 1.697%.

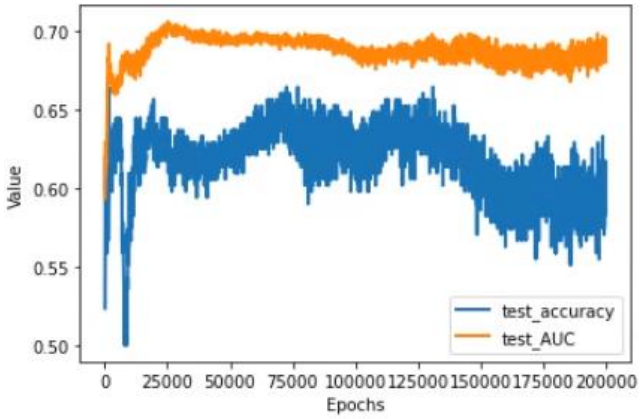


Figure 4-3: Curves of classification accuracy and AUC on the testing dataset.

Table 4-2: Classification performance by different methods on the testing dataset.

Methods	**Accuracy**	**AUC**
Classifier with no augmentation	0.59468	0.66217
Classifier with augmentation	0.64063	0.67960
D of GAN as classifier (proposed)	**0.65470**	**0.69657**

4.4 Discussion

In this study, we initially investigated the effectiveness of using transferring GAN to classify breast masses from mammograms. Our study has two unique characteristics. First, although GAN has been explored as a type of data augmentation method in many previous works, its training is usually separated from the classifier, which makes it difficult for the classifier to have an obvious performance boost [121]. By introducing a supervised loss to the discriminator, we successfully established a connection between the classifier and GAN to ensure that the classifier's training can benefit from the complementary fake images. Some previous studies have reported a modified discriminator with a supervised loss to perform semi-supervised learning tasks on public datasets like MNIST [125]. However, to the best of our knowledge, the effectiveness of this type of discriminator has never been investigated in the context of mammographic mass classification. Our study shows that this transferring GAN is effective in identifying breast mass malignancy, with an average improvement of 1.407% and 1.697% regarding testing accuracy and AUC as compared to classifiers trained with conventional data augmentation (Table 4-2).

Second, and most importantly, it is for the first time that transfer learning is applied to train a GAN for the purpose of classifying benign and malignant masses in a limited-data scenario. Although transfer learning has been widely used in supervised-learning tasks, its impact on advancing GAN's training remains largely undetermined. Only recently, a few studies start investigating the potential of transfer learning, which may be able to efficiently optimize good generators using only hundreds of images [128]. A common goal of these works, however, is restricted to accelerate the generator's convergence and improve the fidelity of generated images, which does not involve the training of a stable generator useful to image classification. In contrast, we integrated transferring GAN into the binary classification task that has a small number of

images (i.e., 768 mass ROIs). The experimental results demonstrate that despite a small dataset, transferring GAN can converge fast once we begin fine-tuning, which can be observed through the FID score curve (Figure 4-1). After the initial stage of fine-tuning (25,000 epochs), the generator was able to generate realistic mass ROIs consistently (Figure 4-2). With the complimentary information from the stable generator, the testing classification accuracy reached peaks near epochs 70,000 and 125,000.

Despite the promising results, we recognize this study has several limitations. First, the ratio of malignant to benign masses cannot reflect the actual cancer incidence rate in the clinical practice. The effectiveness of our method needs to be further validated in common clinical settings where data imbalance is prevalent. Second, we need to evaluate the proposed method's generalizability on different breast mass datasets, since there may exist bias in case selection.

In summary, the novel method presented in this study combines GAN with the transfer learning technique and a supervised loss, which aims to identify benign and malignant breast masses with limited data. This new method may provide a new perspective for researchers in the medical informatics field to effectively integrate GAN into their tasks with small datasets.

Chapter 5: Virtual Adversarial Training for Semi-supervised Breast Mass Classification

5.1 Introduction

Currently, deep learning is the mainstream technique used by most researchers to develop advanced CAD schemes. However, most DL-based CAD schemes are data-hungry – the performance would severely deteriorate when there is no sufficient annotated training data [130, 131]. Due to patient privacy protection and high costs, a large medical image dataset with expert annotations is often difficult to acquire, and data scarcity has become a common issue in clinical practice [8, 9]. Although annotated medical images may be scarce, unlabeled images are sometimes abundant. Numerous studies in computer vision have demonstrated that unlabeled images usually contain valuable information, which is not present in labeled images [132, 133]. If used appropriately, unlabeled images can help achieve better performance but with fewer annotations [134]. Semi-supervised learning (SSL) is one commonly used learning paradigm to extract and exploit the information hidden in unlabeled images [135]. The SSL paradigm applies to scenarios where unlabeled images and labeled ones are relevant (i.e., in the same domain). In the recent literature of medical image analysis, there is a rising interest of applying SSL approaches to leverage unlabeled images to lessen the demand of big, annotated data. For example, in the task of breast mass classification, one study employed a semi-supervised GAN-based model to augment limited amounts of ultrasound images, and this outperformed traditional data augmentation methods [136]. In another study that focuses on segmenting breast mass from ultrasound images, an extended version of a semi-supervised temporal ensembling model was proposed to harness the

underlying knowledge of unlabeled images [137], and this model yielded high segmentation accuracy with a small number of labeled images.

In this study, we adopted "virtual adversarial training" (VAT) [133] to develop a novel semi-supervised mammographic CAD scheme for breast mass classification. Different from other methods, VAT injects additive noise into the training images, aiming to enhance the robustness and generalizability of the model. VAT and VAT-inspired semi-supervised methods [138] have demonstrated great potential in the computer vision field. Nonetheless, few studies investigate the effectiveness of VAT in the medical image analysis field. Only one very recent study [139] shows that VAT is useful for identifying breast cancer from ultrasound images, which are significantly different from mammograms. To the best of our knowledge, this should be one of the pilot studies to investigate the feasibility of employing VAT for the purpose of classifying breast masses from mammograms [140].

5.2 Materials and Methods

5.2.1 Image Dataset

We retrospectively identified a total number of 1024 breast mass images from our existing full-field digital mammography (FFDM) image database [88, 92]. In this assembled dataset, we have 512 malignant and 512 benign cases. For each image, the mass regions were identified and resized to 128×128. The lesion types (benign or malignant) of all the cases were confirmed by biopsy examination. For this dataset, 75% of the whole dataset was used for model training and the rest 25% is for testing. Within the training dataset, a specific ratio of labeled images was used to train models. For example, if the labeled ratio is 20%, the labels of the rest 80% annotated images are hidden on purpose. We used 3 different ratios (20%, 40%, 80%), and the rest training images were considered as unlabeled data.

5.2.2 Introduction of VAT

VAT can be regarded as a more advanced type of data augmentation method, which is able to improve models' generalizability by injecting additive noise to the training data. The central idea behind the adversarial loss is to approximate a virtual adversarial direction r_{vadv} such that the underlying data distribution is perturbed drastically. Suppose each input data point *x* has the corresponding label *y*; the underlying true data distribution is denoted by $q(y|x)$, and the predicted data distribution is $p(y|x,\theta)$, where θ is the model parameters. After applying a perturbation noise r onto the input data points, each perturbed input data point has a new prediction, and the predicted data distribution is denoted by $p(y|x+r,\theta)$. We want to find out the perturbation direction r_{vadv} (i.e., virtual adversarial direction) that can make these two distributions as distant from each other as possible, as defined in equation (5-1).

$$D[q(y|x), p(y|x+r_{vadv},\theta)] \tag{5-1}$$

where $r_{vadv} \coloneqq arg \max_{r;\|r\|_2\leq\varepsilon} D[q(y|x), p(y|x+r,\theta)]$, and $D[q,p]$ measures the divergence between two distributions q and p. A weighting coefficient ε was used to adjust the magnitude of the perturbation [139], which was set as 2.5 in this study.

For unlabeled images, the information of $q(y|x)$ is unknown to users, but its current estimate $p(y|x,\hat{\theta})$ provided by the trained model can be used for approximation, as shown in equation (5-2). This is literally equivalent to inferring "virtual" labels for the unlabeled images. The corresponding divergence value is defined as local distribution smoothness (LDS) value:

$$LDS = D\left[q(y|x,\hat{\theta}), p(y|x+\varepsilon * r_{vadv},\theta)\right] \tag{5-2}$$

In this investigation, the virtual adversarial direction is computationally estimated by the iterative optimization technique described in the paper [133]. Based on the estimated perturbation noise, the corresponding virtual adversarial loss $R_{vadv}(D_{ul}, \theta)$ is determined by averaging the LDS values of all the unlabeled input samples [133]:

$$R_{vadv} = \frac{1}{N}\sum_{x \in S_{ul}} LDS(x, \theta) \tag{5-3}$$

where S_{ul} represents the set containing all the unlabeled samples, and N is the total number of the unlabeled samples. More details of the VAT method can be found in the reference [133].

5.2.3 VAT-Based CAD Scheme for Mammographic Breast Mass Classification

Figure 5-1 shows the pipeline of our VAT-based CAD scheme for mammographic breast mass classification. For labeled mass images, we use the same cross entropy loss l_{ce} as in supervised classification. As discussed in the above section, the virtual adversarial loss can directly work with unlabeled breast mass images, and the complete loss l_c of our VAT-based model is given as follows:

$$l_c = l_{ce} + \alpha * R_{vadv}(D_{ul}, \theta) \tag{5-4}$$

In equation (5-4), the first part is a supervised loss, and the second part is the virtual adversarial loss. The importance of this adversarial loss can be adjusted by changing the value of α. In other words, the adversarial loss of the unlabeled breast mass images serves as a regularization term to provide additional information to facilitate mass classification. By making the model robust to virtual adversarial perturbation, this regularization term can help to improve the generalizability of trained models on unseen data. We set this hyperparameter to be 1 during the experiments.

Two models were used in this study, including a large CNN and a small CNN. Both of these two models were modified from the architectures used in previous study [133]. The large CNN is composed of 19 layers, including 5 convolutional blocks, 2 max-pooling layers (MaxPool), 1 average pooling layer (AdaptiveAvgPool) and 1 linear layer. These five blocks have the same type of layers, but the filter numbers are different. Each convolutional block has 3 layers, namely, convolutional layer, batch normalization layer, and leaky rectified linear unit. The structure details of the large CNN are presented in Table 5-1. In the table, the output shape column provides the number of the filters (channels) and output image size. For example, the output of the third block (256, 64, 64) has a total of 256 feature maps of size 64×64. The small CNN has the same architecture as the large CNN, but only contains half of the filters. Meanwhile, the Adam optimizer was used during model training, with a learning rate of 0.001. The batch size was set to be 32. Using these two types of the losses, the CNN model generates the classification scores, indicating the likelihood that the case belongs to benign or malignant category. The model optimization and validation were performed for 3 times, and the classification accuracies were averaged for performance evaluation.

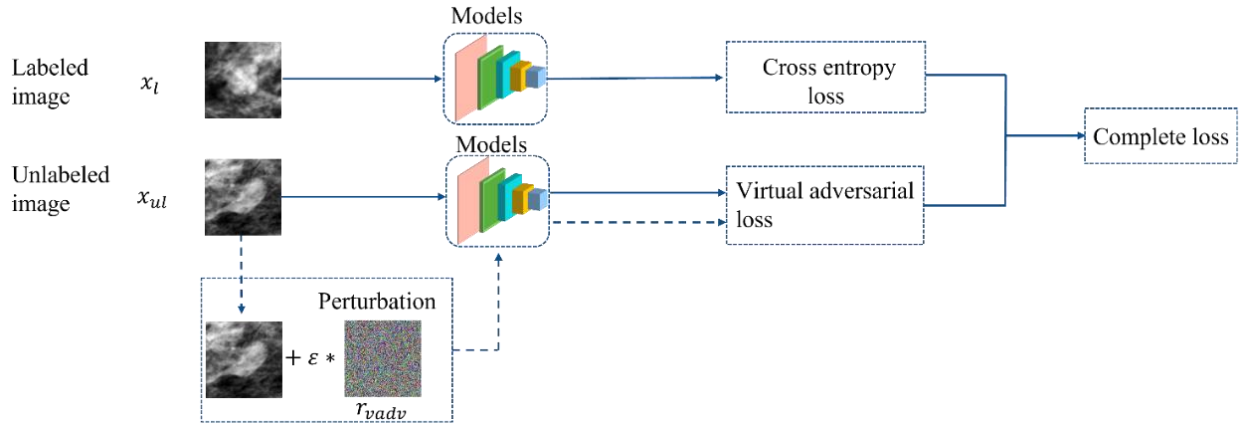


Figure 5-1: Pipeline of VAT-based models for mammographic breast mass classification.

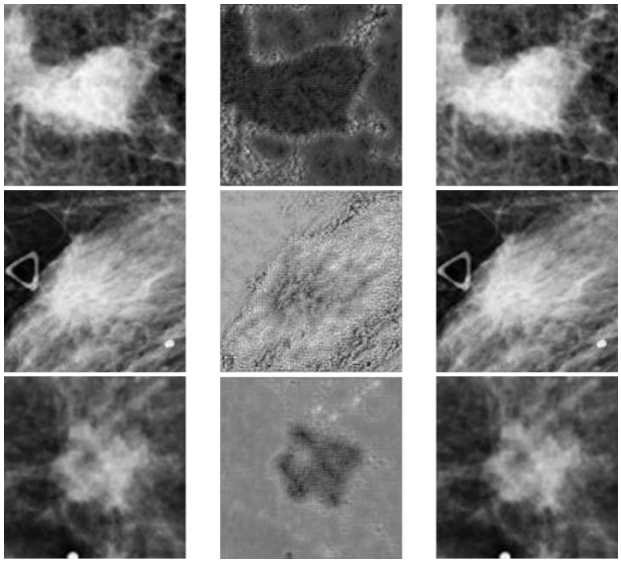
Table 5-1: Structure details of the large CNN used in this study.

Layer Type	Output Shape
Block 1	(128, 128, 128)
Block 2	(128, 128, 128)
MaxPool	(128, 64, 64)
Block 3	(256, 64, 64)
Block 4	(256, 64, 64)
MaxPool	(256, 32, 32)
Block 5	(128, 32, 32)
AdaptiveAvgPool	(128, 1, 1)
Linear	2

5.3 Results

Figure 5-2 demonstrates the effect of virtual adversarial perturbations on mammographic breast mass images. Based on the original image (Figure 5-2a), VAT generates perturbation noises (Figure 5-2b). Different from the random noise, the generated perturbation noise is correlated with the shape of the mass and noticeable tissues. The perturbed image (Figure 5-2c) is the combination of the perturbation noise and original image. The perturbed results are visually similar to the original image, but conventional CNN models may generate opposite predictions. However, our

VAT-based model can help overcome this limitation, increasing the model generalizing capability on unseen data.



(a) (b) (c)

Figure 5-2: a) Examples of original mass images; b) Perturbation noise generated by VAT; c) Perturbed results.

Table 5-2: Classification accuracy of VAT-based models for semi-supervised mass classification.

Models	**Using VAT**	**20% labeled**	**40% labeled**	**80% labeled**
Large CNN	N	0.719±0.026	0.719±0.051	0.750±0.051
	Y	0.729±0.053	**0.740±0.015**	**0.760±0.015**
Small CNN	N	**0.729±0.015**	0.719±0.026	0.729±0.015
	Y	0.719±0.051	0.719±0.026	0.740±0.078

Table 5-2 shows the classification performance of two VAT-based CNN models. Models that do not use VAT are trained only on the labeled data. When 20% of the cases were labeled, the large and small CNN indicate completely different trends: adding VAT loss improves the classification performance (0.729 ± 0.053 vs 0.719 ± 0.026) of the large CNN, but deteriorates the performance of the small CNN (0.719 ± 0.051 vs 0.729 ± 0.015). When the labeled ratio increases to 40%, the VAT based large CNN model outperforms the model without VAT. The small CNN model, however, yielded the same performance no matter VAT is used or not. When the labeled ratio is 80%, the VAT based models consistently achieved better accuracy. Meanwhile, the highest performance of 0.760 ± 0.015 was yielded by the VAT based large CNN with a labeled ratio of 80%.

5.4 Discussion

This study contributes to exploring the potential use of VAT in medical image classification tasks, via evaluating its performance in classifying benign and malignant breast masses from mammograms. We show that when the labeled ratios were 40% and 80%, VAT-based CNNs that exploit useful information from unlabeled images can outperform the CNNs trained without VAT. However, when the labeled ratio was 20%, CNNs trained without VAT achieved better performance. As discussed in **Section 5.2**, the virtual adversarial perturbations for unlabeled images are dependent on the label estimates produced by CNNs; however, when the number of labeled images becomes too low, it may be difficult for CNNs to converge and provide high-quality label estimates for unlabeled samples. Therefore, we believe that the performance deterioration is probably associated with the model's poor ability to produce reliable label estimates for unlabeled samples.

Despite the encouraging results obtained from the VAT-based models, it is still desirable to investigate 1) whether similar results can be obtained on a larger and more diversified breast mass dataset, and 2) whether changing the weight of the virtual adversarial loss can further improve the performance. In addition, this study treats the importance/ influence of each labeled or unlabeled sample equally, which may affect the performance of VAT. We expect to extend VAT by incorporating the importance of training samples in future work.

In summary, our study initially verified the feasibility of using VAT to improve the performance of classifying benign and malignant breast masses from mammograms in a semi-supervised manner. This new method may provide a new perspective regarding how to leverage unlabeled medical images into supervised tasks that have only a small number of labeled samples.

Chapter 6: Conclusions and Future Work

6.1 Conclusions

CAD schemes can assist radiologists to interpret medical images more accurately and efficiently as well as decrease inter-reader variability. Both the traditional ML-based and modern DL-based schemes have been extensively researched over the years, but certain challenges still exist for each type of scheme.

In Chapter 2, to overcome the challenge of lacking effective features for CT LN metastasis prediction, we established the largest radiomics-based feature pool at that time by computing 1,800 features that were categorized into 14 groups. We evaluated the effectiveness of this feature pool to predict LN metastasis for locally advanced cervical cancer patients. Five out of the fourteen groups (i.e., Shape and Density, Wavelet-DB2, Pyramid-FD, Wavelet-FD, and Wavelet-Haar) were identified containing significant discriminative power. The best of our knowledge, this is the first study that applies a comprehensive CT image based radiomic feature pool to predict LN metastasis of cervical cancer patients.

In Chapter 3, to tackle the limitations caused by the difficulty in mammographic lesion segmentation (e.g., low accuracy and/or poor reproducibility of feature computation), we developed a CAD scheme that utilizes global image features for mammogram malignancy prediction. Without the need for lesion segmentation, this new CAD scheme achieved a comparable performance to traditional CAD schemes with lesion segmentation. We demonstrated that the global features extracted from the two-view positive images of one breast have diagnostic power to classify between malignant and benign mammographic cases. This CAD scheme avoids the difficulty in both segmentation of the lesions and determination of the optimal ROIs.

In Chapter 4, to tackle the challenge of lacking sufficient medical data for training deep learning models, we initially investigated the effectiveness of using transferring GAN to classify benign and malignant breast masses from mammograms. This differs from previous studies that focus on utilizing transfer learning techniques to improve the generated images' quality. We show that despite a small dataset, GAN's training can converge fast with the help of transfer learning. Since this transferring GAN outperformed classifiers trained with traditional data augmentation methods, we believe the complementary information from a stable generator can serve as a useful source for performance boost in low-data regimes.

Chapter 5 also aims at alleviating the issue of lacking large amounts of labeled medical images for training deep learning models. We proposed employing a semi-supervised method (virtual adversarial training (VAT)), to leverage and learn useful information underlying in unlabeled data for better classification of breast masses. The experimental results suggest that the VAT-based CAD scheme can effectively utilize meaningful knowledge from unlabeled data to better classify mammographic breast mass images. This new CAD scheme may provide a new angle in terms of how to appropriately leverage unlabeled medical images into supervised tasks with a small number of labeled samples.

Throughout my PhD study, I have published 4 journal papers and 5 conference papers, which are listed as follows.

6.1.1 Journal Papers

1. **Chen, Xuxin**, Ximin Wang, Ke Zhang, Roy Zhang, Kar-Ming Fung, Theresa C. Thai, Kathleen Moore et al. "Recent advances and clinical applications of deep learning in medical image analysis." (To appear in *Medical Image Analysis*).

2. Zhang, Ke, Xianglan Lu, **Xuxin Chen**, Roy Zhang, Kar-Ming Fung, Hong Liu, Bin Zheng, Shibo Li, and Yuchen Qiu. "Using Fourier ptychography microscopy to achieve high-resolution chromosome imaging: an initial evaluation." *Journal of Biomedical Optics* 27, no. 1 (2022): 016504.

3. **Chen, Xuxin**, Wei Liu, Theresa C. Thai, Tara Castellano, Camille C. Gunderson, Kathleen Moore, Robert S. Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. "Developing a new radiomics-based CT image marker to detect lymph node metastasis among cervical cancer patients." *Computer Methods and Programs in Biomedicine* 197 (2020): 105759.

4. **Chen, Xuxin**, Abolfazl Zargari, Alan B. Hollingsworth, Hong Liu, Bin Zheng, and Yuchen Qiu. "Applying a new quantitative image analysis scheme based on global mammographic features to assist diagnosis of breast cancer." *Computer methods and programs in biomedicine* 179 (2019): 104995.

6.1.2 Conference Proceeding Papers

1. **Chen, Xuxin**, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C. Thai, Kathleen Moore, Robert S. Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. "Virtual Adversarial Training for Semi-supervised Breast Mass Classification." (To appear in *Biophotonics and Immune Responses XVI* of SPIE).
2. Ke Zhang, Patrik Gilley, Xianglan Lu, **Xuxin Chen**, Neman Abdoli, Roy Zhang, Kar-Ming Fung, Hong Liu, Bin Zheng, Shibo Li, Yuchen Qiu. "Improving the resolution of chromosome imaging by high numerical aperture Fourier ptychography microscopy." (To appear in *Biophotonics and Immune Responses XVI* of SPIE)
3. **Chen, Xuxin**, Theresa C. Thai, Camille C. Gunderson, Katheleen Moore, Robert S. Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. "Development of a transferring GAN based CAD scheme for breast mass classification: an initial study." In *Biophotonics and Immune Responses XVI*, vol. 11643, p. 116430H. International Society for Optics and Photonics, 2021.
4. **Chen, Xuxin**, Roy Zhang, Kar-Ming Fung, Hong Liu, Bin Zheng, and Yuchen Qiu. "Utilizing a transfer model to classify epithelium and stroma on digital histopathological images for ovarian cancer patients." In *Biophotonics and Immune Responses XV*, vol. 11241, p. 112410F. International Society for Optics and Photonics, 2020.
5. Liu, Wei, Shiyu Pei, **Xuxin Chen**, Theresa C. Thai, Tara Castellano, Camille C. Gunderson, Kathleen Moore et al. "Developing a low-cost image marker to identify lymph node metastasis for cervical cancer patients: an initial study." In *Biophotonics and Immune Responses XV*, vol. 11241, p. 112410Y. International Society for Optics and Photonics, 2020.

6.2 Future Studies

6.2.1 Toward Better Combinations of Deep Learning and Medical Image Analysis

1) On the Task-Specific Perspective

The progress of medical image analysis using deep learning follows a lagging but similar timeline to computer vision. However, due to the difference between medical images and natural images, a direct use of methods from computer vision may not yield satisfactory results. In practice, developing a high-performing DL-based CAD scheme often involves several different medical image analysis tasks. In order to achieve good performance, challenges unique to medical image analysis tasks need to be addressed [135]. For the **classification** task, the key to success lies in extracting highly discriminative features with respect to certain classes. This is relatively easy for domains with large inter-class variance (e.g., accuracies on many public chest X-ray datasets often exceed 90%), but it can be difficult for domains with high inter-class similarity. For example, the performance of mammogram classification is not so good overall (e.g., 70~80% accuracies are commonly seen on private datasets), since discriminative features for breast tumors are difficult to capture in the presence of overlapping, heterogeneous fibroglandular tissues [141]. The notion of *fine-grained visual classification (FGVC)* [142], which aims at identifying subtle differences between visually similar objects, might be suited for learning distinctive features given high inter-class similarity. But note that, benchmark FVGC datasets were purposely collected to make all the image samples unanimously exhibit high inter-class similarity. As a result, approaches developed and evaluated on these datasets may not be readily applicable to medical datasets, where only a certain fraction rather than all the images exhibit high inter-class similarity. Nonetheless, we believe FVGC methods, if modified appropriately, will be valuable to learning feature representations with high discriminative power in medical image classification. Other possible

ways to enhance features' discrimination power include the use of attention modules, local and global features, domain knowledge, etc.

Medical object **detection** is more complicated than classification as can be seen from the process of bounding box prediction. Naturally, detection faces the challenges inherent to classification. Meanwhile, there exist additional challenges, especially the detection of small-scale objects (e.g., small lung nodules) and class imbalance. One-stage detectors typically perform comparably well as two-stage detectors in detecting large objects but struggles more in detecting small objects. Existing studies show that using multi-scale features can greatly alleviate this issue both in one-stage and two-stage detectors. A simple yet effective approach is *featurized image pyramids* [143], where features are extracted from multiple scales of the same image independently. This method can help enlarge small objects to achieve better performance but is computationally expensive and slow. Nonetheless, it is suitable to medical detection tasks with no requirement of fast speed. Another useful but much faster approach is *feature pyramids*, which utilizes multi-scale feature maps from different convolutional layers. Although there exist various ways to build feature pyramids, a rule of thumb is that it is necessary to fuse strong, high-level semantics with high-resolution feature maps. This plays an important role in detecting small objects, as shown by FPN [144].

Class imbalance arises from the fact that detectors need to evaluate a huge number of candidate regions, but only a few contain objects of interest. In other words, class balance is severely skewed toward negative examples (e.g., background regions), most of which are easy negatives. The presence of large amounts of easy negatives can overwhelm the training process, leading to bad detection results. Two-stage detectors can handle this class imbalance issue much better than one-stage detectors, because most negative proposals are filtered out at the region proposal stage. In

terms of one-stage detectors, recent studies show that abandoning the dominant use of anchor boxes can largely alleviate class imbalance [145]. However, most approaches adopted in medical object detection are still anchor-based. In the near future, we expect to see more explorations of anchor-free, one-stage detectors in medical object detection.

Medical image **segmentation** combines challenges in classification and detection. Just like detection, class imbalance is a common issue across 2D and 3D medical segmentation tasks. Another similar challenge is the segmentation of small-sized lesions (e.g., MRI multiple sclerosis) and organs (e.g., pancreas from abdominal CT scans). Also, these two challenges often appear intertwined. These issues have been largely alleviated by adapting metrics/losses to evaluate the segmentation performance, such as Dice coefficient [146], generalized Dice [147], the integration of focal loss [148], etc. However, these metrics are region-based (i.e., segmentation errors are computed in a pixel-wise manner). This can lead to a loss of valuable information regarding structures, shapes, and contours that are important to diagnosis/prognosis in later stages. Therefore, we believe it is necessary to develop non-region-based metrics that can provide complementary information to region-based metrics for better segmentation performance. Currently only a few studies exist in this direction [149]. We expect to see more in the future.

In addition, strategies such as incorporating local and global context, attention mechanisms, multi-scale features, and anatomical cues are generally beneficial to increasing segmentation accuracy for both large and small objects. Here we want to emphasize the great potential of Transformers due to their strong capability of modeling long-range dependencies. Although long-range dependencies are helpful to achieving accurate segmentation, a majority of CNN-based methods do not explicitly focus on this aspect. There are roughly two types of dependencies, namely intra-slice dependency (pixel relationships within a CT or MRI slice) and inter-slice

dependency (pixel relationships between CT or MRI slices) [150]. Recent studies show that Transformer-based approaches are powerful in both cases [151]. Applications of Transformers for medical image segmentation especially 3D are still in the initial stage, and more works in this trial are likely to emerge soon.

2) On the Perspective of Different Learning Paradigms

Although deep learning has brought about huge successes across different tasks in the context of radiological image anlaysis, the further performance improvement is majorly hurdled by the requirement for large amounts of annotated datasets. Supervised transfer learning can greatly alleviate this issue, by initializing the model's weights (for the target task) with the weights of the model that is pre-trained on relevant/irrelevant datasets (e.g. ImageNet). Besides the widely used transferring learning, we identify two possible directions: (1) utilizing GAN model to enlarge the labeled dataset; (2) utilizing semi-supervised learning models to exploit the information underlying vast unlabeled medical images.

GAN has shown great promise in medical image synthesis and semi-supervised learning; but one challenge is how to build a strong connection between GAN's generator and the target task (e.g., classifier, detector, segmentor). The lack of such connection may cause a subtle performance boost as compared to the conventional data augmentation (e.g., rotation, rescale, and flip) [119]. The connection between the generator and classifier can be strengthened by utilizing semi-supervised GAN, in which the discriminator was modified to serve as a classifier [152]. Several training strategies can also be employed: identifying a "bad" generator that can significantly contribute to good semi-supervised classification [153]; jointly optimizing the triple components of a generator, a discriminator, and a classifier [154]. It is meaningful to explore new ways that can effectively set up connections between the generator and a specific medical image task for a

better performance. Additionally, GAN usually needs at least thousands of training examples to converge, which limits its applicability on small medical datasets. This challenge can be partially addressed by using classic data augmentation for adversarial learning [118]. Further, if there exist relatively large amounts of medical images that share structural, textural, and semantic similarities with the target dataset, pre-training generators and/or discriminators may facilitate faster convergence and better performance [155]. Meanwhile, some recent novel augmentation mechanisms, such as the differentiable augmentation [156] and adaptive discriminator augmentation [157] have enabled GAN to effectively generate high-fidelity images under data-limited conditions, but they have not been applied to any medical image analysis tasks. We anticipate that these new methods can also demonstrate promising performance in future studies of the medical image analysis field.

Recent **semi-supervised** methods such as FixMatch [158] heavily rely on advanced data augmentation strategies to achieve good performance. To facilitate the applications of semi-supervised learning in medical image analysis, it is necessary to develop appropriate augmentation policies in a *dataset-driven* and/or *task-driven* manner. Being "dataset-driven" means finding the optimal augmentation policy for a specific dataset of interest. In the past, this was not easy to achieve due to the extremely very large size of the parameter search space (e.g., 10^{34} possible augmentation policies as shown in paper [159]). Recently, automated data augmentation strategies like *RandAugment* [159] have been proposed to significantly reduce the search space. However, the concept of automated augmentation remains largely unexplored in medical image analysis. Being "task-driven" means finding suitable augmentation strategies for a specific task (e.g., MRI prostrate segmentation) that have several datasets. This could be regarded as the extension of

dataset-driven augmentation and thus is more challenging, but it can help algorithms developed on one dataset generalize better to other dataset(s) of the same task.

Another issue is the potential performance degradation caused by violation of the underlying assumption of semi-supervised learning – labeled and unlabeled data are from the same distribution. Indeed, distribution mismatch is a common problem when semi-supervised methods are applied for medical image analysis. Consider the following example: in the task of segmenting COVID-19 lung infections from CT slices, say you have a set of labeled CT volumes containing a relatively balanced number of infected and non-infected slices, while the unlabeled CT volumes available may contain no or just a few infected slices. Or the unlabeled CT images contain not only COVID-19 infections but also some other disease class(es) (e.g., tuberculosis) that are absent from the labeled images. What will happen if the distribution of unlabeled data mismatches with the distribution of labeled data? Exiting studies suggest this will cause the performance of semi-supervised methods to degrade drastically, sometimes even worse than that of a simple supervised baseline [160, 161]. Therefore, it is necessary to adapt semi-supervised algorithms to be tolerant of the distribution mismatch between labeled and unlabeled medical data. As a related field, "domain adaption" may provide insights for achieving this goal.

3) Finding Better Architectures and Pipelines

The continuing success of deep learning in medical image analysis originates from not only different learning paradigms (unsupervised, semi-supervised) but also, maybe to a larger extent, the architectures/models proposed over time. Looking back, we find non-trivial improvements are closely related to the progress of "architectures", and examples include AlexNet [162], residual connections [163], skip connections [164], self attention [165], etc. "Given this progression history, it is certainly possible that a better neural architecture can by itself overcome many of the

current limitations", as pointed out by Yuille and colleagues [166]. We disucuss two aspects that may be helpful to finding better architectures. First, biologically and cognitively inspired mechanisms will continue playing an important role in architecture designing. Deep learning neural networks were originally inspired by the architecture of cerebal context. In recent years the concept of attention, which was inspired by primates' visual attention mechanisms, has been successfully used in NLP and computer vision to make models focus on important parts of input data, leading to superior performance. A preeminent example are the family of Transformers based on self attention [165]. Transformer-based architectures are better at capturing global/long-range dependencies between input and output sequences than mainstream models based on CNNs. Also, inductive biases inherent to CNNs (e.g., translation equivariance and locality) are much less in Transformers [167]. Aside from the attention mechanisms, many other biological or cognitive mechanisms, such as dynamic hierarchies in human language, one-shot learning of new objects and concepts without gradient descent, etc [168], may provide inspirations for designing more powerful achitectures. Second, automatic architecture engineering may shed light on developing better architectures. Currently employed architecutres mostly come from human experts, and the designing process is iterative and prone to errors. Partially for this reason, models used for medical image anlaysis are primarily adapted from models developed in computer vision. To avoid the need of manual designing, reserachers have proposed to automate architecutre engineering, and one related field is *neural architecture search* (NAS) [169]. However, most exisiting studies of NAS are confined within image classification [170], and truly revolutionary models that can bring fundamental changes have not come out of this process [166]. Nonetheless, NAS is still a direction worthy exploration.

At a broader level, piplelines with automated configuration capabilities would be desirable. Although architecture engineering still faces many difficulties, developing automatic pipelines, which are capable of automatically configuring its subcomponents (e.g., choosing and adapting an appropriate architecture among the exisiting ones) to achieve better performance, will be beneficial to radiolgical image analysis. At present, deep leanring based pipelines typically involve several interdepedent subcomponents such as image preprocessing and post-processing, adapting and training a network architecture, selecting appropriate losses, data augmentation methods, etc. But the design choices are often too many for experimenters to manually figure out an optimal pipeline. Moreover, a high-performing pipeline configured for a dataset (e.g., CT images from one hospital) of a specific task may perform badly on anther dataset (e.g., CT images from a different hospital) of the same task. Therefore, pipelines that can automatically configure their subcomponents are needed to speed up empirical design. Examples falling in this scope include NiftyNet [171], a modular pipeline for different medical applications, and nnU-Net [172] specifically for medical image segmentation. We expect more research will be coming out of this track.

4) Incorporating Domain Knowledge

Domain knowledge, which is an important aspect but sometimes overlooked, can provide insights for developing high-performing deep learning algorithms in medical image analysis. As mentioned previously, most models used in medical vision are adapted from models developed for natural images; however, medical images are generally more difficult to handle due to unique challenges (e.g., high inter-class similarity, limited size of labeled data, label noise). Domain knowledge, if used appropriately, helps alleviate these issues with less time and computation costs. It is relatively easy for researchers with strong deep learning background to utilize *weak* domain knowledge, such as anatomical information in MRI and CT images [173, 174], multi-instance data

from the same patient [175], patient metadata [176], radiomic features, and text reports accompanying images [177]. On the other hand, we observe it can be more difficult to effectively incorporate *strong* domain knowledge that radiologists are familiar with. One example is breast cancer identification from mammograms. For each patient, four mammograms are available, including two cranio-caudal (CC) and two medio-lateral oblique (MLO) view images of left (L) and right (R) breasts. In clinical practice, the bilateral difference (e.g., LCC vs. RCC) and unilateral correspondence (e.g., LCC and LMLO) serve as important cues for radiologists to detect suspicious regions and determine malignancy. Currently there exist few methods that can reliably and accurately to utilize this expert knowledge. Therefore, more research efforts are needed to maximize the use of strong domain knowledge.

6.2.2 Toward Large-Scale Applications of Deep Learning in Clinical Settings

Deep learning, despite being intensively used for analyzing medical images in academia and industrial research institutions, has not made that significant impact as we expected in clinical practice. This is clearly reflected in the early stages of fighting against COVID-19, the first global pandemic falling in the era of deep learning. Due to its widespread medical, social, and economic consequences, this pandemic, to a large extent, can be regarded as a big test for examining the current status of deep learning algorithms in clinical translation. Soon after the outbreak, researchers around the world applied deep learning techniques to analyze mainly chest X-rays and CT images from patients with suspected infection, aiming at accurate and efficient diagnosis/prognosis of the disease. To this end, numerous deep learning and machine learning based approaches were developed. However, after systematically reviewing over 200 prediction models from 169 studies that were published up to 1 July 2020, Wynants and colleagues [178] concluded that all these models were of high or unclear risk of bias, and thus none of them were

suitable for clinical use – either moderate or excellent performance was reported by each model; however, the optimistic results were highly biased due to model overfitting, inappropriate evaluation, use of improper data sources, etc. Similar conclusion was drawn in another review paper [179] – after reviewing 62 studies that were selected from 415 studies the authors concluded that, because of methodological flaws and/or underlying biases, none of the deep learning and machine learning models identified were clinically applicable to the diagnosis/prognosis of COVID-19.

Going beyond the example of COVID-19, the high-risk bias of deep learning approaches is indeed a recurring concern across different medical image analysis tasks and applications [180], which has severely limited deep learning's potential in clinical radiology. Although quantifying the underlying bias is difficult, it can be reduced if handled appropriately. In the following we summarize three major aspects that could lead to biased results and provide our recommendations.

1) Image Datasets

Medical image datasets with increasingly larger size (e.g., usually at least several hundred images) have been or are being developed to facilitate training and testing new algorithms. One notable example is the yearly MICCAI challenges where benchmark datasets for different diseases (e.g., cancer) are released, greatly promoting the progress of medical vision. However, we need to be cautious about the potential biases caused by using a single public dataset alone – as the whole community strive for achieving state of the art performance, community-wide overfitting is likely to exist on this dataset [179]. This problem has been recognized by many researchers, so it is common to see several public datasets and/or private dataset(s) are used to test a new algorithm's performance more comprehensively. In this way the community-wide bias is reduced but not to the extent of large-scale clinical applications.

The community-wide bias can be further lowered by incorporating additional data to train and test models. One direct way to introduce new data, of course is data curation, i.e., continually creating large, diverse datasets via collective work with experts. Different from this track, we recommend a less direct but effective way – integrating scattered private datasets as ethical and law regulations permit. The medical image analysis community might have the overall impression that large, representative, labeled data seems always lacking. This is only partially true, though. Due to time and cost constraints, it is true that many established public datasets have limited size and variety. On the other hand, rich medical image sources (labeled and unlabeled) of different sizes and difficulty levels already exist but inconveniently "in the form of isolated islands" [181]. Because of factors such as privacy protection and political intricacy, most existing data sources are kept private and scattered in different institutions across different countries. Thus, it would be desirable to exploit the unified potential of private datasets and even personal data without comprising patients' privacy. A promising approach to achieving this goal is *federated learning* [182], which allows models to securely access sensitive data. Federated learning can train deep learning algorithms collaboratively on multi-institutional data without exchanging data among participating institutions [183]. Although this technology is accompanied by new challenges, it facilitates learning less biased, more generalizable, more robust, and better-performed algorithms that would better meet the needs of clinical applications.

2) Performance Evaluation

Most research papers in medical image analysis report models' performance via commonly used metrics, for example, accuracy and AUC for classification tasks, and Dice coefficient for segmentation tasks. While these metrics can easily quantify the technical performance of presented approaches, they often fail to reflect clinical applicability. Ultimately, clinicians care about whether the use of algorithms would bring about a beneficial change in patient care, rather than the performance gains reported in papers [184]. Therefore, aside from applying necessary metrics, we believe it is important for research teams to collaborate with clinicians for algorithms appraisal.

We simply mention two possible directions as to establishing collaborative evaluation. First, involve clinicians into viewpoints sharing of open clinical questions, paper writing, and even the peer review process of conferences and journals. For example, the *Machine Learning for Healthcare* (MLHC) conference provides a research track and clinical track for members from separated communities to exchange insights. Second, measure if the performance and/or efficiency of clinicians can be improved with the assistance of deep learning algorithms. Utilizing model results as a "second opinion" to facilitate clinicians' final interpretation has been explored in some studies. For instance, in the task of predicting breast cancer from mammograms, evaluated the complementary role of deep learning model. They found that the model could correctly identify many cancer cases missed by radiologists. Furthermore, in the "double-reading process" (standard practice in UK), the model significantly reduced the second reader's workload while maintaining a comparable performance to the consensus opinion.

3) Reproducibility

The quick progress of computer vision is closely related to the research culture that advocates reproducibility. In medical image analysis, more and more researchers choose to make their code publicly available, and this greatly helps avoid duplication of effort. More importantly, good reproducibility can help deep learning algorithms gain more trust and confidence from a wide population (e.g., researchers, clinicians), which is beneficial to large-scale clinical applications. To make the results more reproducible, we suggest paying extra attention to describing data selection in papers. It is not uncommon to see that different studies select different subsets of samples from the same public dataset. This could increase the difficulty of reproducing results stated in the paper. In a case study on lung nodule classification, Baltatzis and colleagues [185] demonstrated that specific choices of data turn out to be favorable to proving the proposed models' superiority. Advanced models with bells and whistles may underperform simple baselines if data samples are changed. Therefore, it is necessary to clearly state the data selection process to make the results more reproducible and convincing.

In conclusion, deep learning is a fast-developing technology, and has produced promising potential in broad medical image analysis fields including disease classification, segmentation, detection. Despite of significant research progress, we are still facing many technical challenges or pitfalls [179] to develop deep learning based CAD schemes that can achieve high scientific rigor. Therefore, more research efforts are needed to overcome these pitfalls before the deep learning based CAD schemes can be commonly accepted by clinicians.

References

1. Kruger, R.P., et al., *Automated Radiographic Diagnosis via Feature Extraction and Classification of Cardiac Size and Shape Descriptors.* IEEE Transactions on Biomedical Engineering, 1972. **BME-19**(3): p. 174-186.
2. Ko, J.M., et al., *Prospective assessment of computer-aided detection in interpretation of screening mammography.* AJR Am J Roentgenol, 2006. **187**(6): p. 1483-91.
3. Litjens, G.J., et al., *Clinical evaluation of a computer-aided diagnosis system for determining cancer aggressiveness in prostate MRI.* European radiology, 2015. **25**(11): p. 3187-3199.
4. McKinney, S.M., et al., *International evaluation of an AI system for breast cancer screening.* Nature, 2020. **577**(7788): p. 89-94.
5. Meyers, P.H., et al., *Automated computer analysis of radiographic images.* Radiology, 1964. **83**(6): p. 1029-1034.
6. Sezaki, N. and K. Ukena, *Automatic Computation of the Cardiothoracic Ratio with Application to Mass Screening.* IEEE Transactions on Biomedical Engineering, 1973. **BME-20**(4): p. 248-253.
7. Doi, K., et al., *Computer-aided diagnosis in radiology: potential and pitfalls.* European Journal of Radiology, 1999. **31**(2): p. 97-109.
8. Litjens, G., et al., *A survey on deep learning in medical image analysis.* Medical image analysis, 2017. **42**: p. 60-88.
9. Shen, D., G. Wu, and H.-I. Suk, *Deep learning in medical image analysis.* Annual review of biomedical engineering, 2017. **19**: p. 221-248.

10. Sahiner, B., et al., *Computer-aided characterization of mammographic masses: accuracy of mass segmentation and its effects on characterization.* IEEE Transactions on Medical Imaging, 2001. **20**(12): p. 1275-1284.

11. Yan, S., et al., *Applying a new bilateral mammographic density segmentation method to improve accuracy of breast cancer risk prediction.* International journal of computer assisted radiology and surgery, 2017. **12**(10): p. 1819-1828.

12. Lambin, P., et al., *Radiomics: extracting more information from medical images using advanced feature analysis.* European journal of cancer, 2012. **48**(4): p. 441-446.

13. Kumar, V., et al., *Radiomics: the process and the challenges.* Magnetic resonance imaging, 2012. **30**(9): p. 1234-1248.

14. Aerts, H.J., et al., *Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.* Nat Commun, 2014. **5**(4006).

15. Hughes, G., *On the mean accuracy of statistical pattern recognizers.* IEEE transactions on information theory, 1968. **14**(1): p. 55-63.

16. Bellman, R.E., *Adaptive control processes*. 2015: Princeton university press.

17. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* Journal of machine learning research, 2003. **3**(Mar): p. 1157-1182.

18. Kennedy, J. and R. Eberhart. *Particle swarm optimization*. in *Proceedings of ICNN'95 - International Conference on Neural Networks*. 1995.

19. Van Der Maaten, L., E. Postma, and J. Van den Herik, *Dimensionality reduction: a comparative.* J Mach Learn Res, 2009. **10**(66-71): p. 13.

20. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-444.

21. van Engelen, J.E. and H.H. Hoos, *A survey on semi-supervised learning.* Machine Learning, 2020. **109**(2): p. 373-440.

22. Visvikis, D., et al., *Artificial intelligence, machine (deep) learning and radio (geno) mics: definitions and nuclear medicine imaging applications.* European journal of nuclear medicine and molecular imaging, 2019. **46**(13): p. 2630-2637.

23. Tanaka, H., et al., *Computer-aided diagnosis system for breast ultrasound images using deep learning.* Physics in Medicine & Biology, 2019. **64**(23): p. 235013.

24. Anwar, S.M., et al., *Medical image analysis using convolutional neural networks: a review.* Journal of medical systems, 2018. **42**(11): p. 226.

25. Goodfellow, I., et al., *Generative adversarial nets.* Advances in neural information processing systems, 2014. **27**.

26. Yi, X., E. Walia, and P. Babyn, *Generative adversarial network in medical imaging: A review*. Medical image analysis, 2019. **58**: p. 101552.

27. Chapelle, O., B. Scholkopf, and A. Zien, *Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]*. IEEE Transactions on Neural Networks, 2009. **20**(3): p. 542-542.

28. Ouali, Y., C. Hudelot, and M. Tami, *An overview of deep semi-supervised learning.* arXiv preprint arXiv:2006.05278, 2020.

29. Lee, D.-H. *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks*. in *Workshop on challenges in representation learning, ICML*. 2013.

30. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2020.* CA: A Cancer Journal for Clinicians, 2020. **70**(1): p. 7-30.

31. Selman, T.J., et al., *Diagnostic accuracy of tests for lymph node status in primary cervical cancer: a systematic review and meta-analysis.* Cmaj, 2008. **178**(7): p. 855-62.

32. Rose, P.G., et al., *Concurrent cisplatin-based radiotherapy and chemotherapy for locally advanced cervical cancer.* N Engl J Med, 1999. **340**(15): p. 1144-53.

33. Bernardini, M.Q. and A. Covens, *Imaging of lymph node metastases in cervical cancer.* Cmaj, 2008. **178**(7): p. 867-9.

34. Elit, L.M., et al., *Effect of Positron Emission Tomography Imaging in Women With Locally Advanced Cervical Cancer: A Randomized Clinical Trial.* JAMA Netw Open, 2018. **1**(5).

35. Aerts, H.J., *The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review.* JAMA Oncol, 2016. **2**(12): p. 1636-1642.

36. Itakura, H., et al., *Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities.* Sci Transl Med, 2015. **7**(303).

37. Zargari, A., et al., *Prediction of chemotherapy response in ovarian cancer patients using a new clustered quantitative image marker.* Phys Med Biol, 2018. **63**(15): p. 1361-6560.

38. Zhang, H.X., et al., *A pilot study of radiomics technology based on X-ray mammography in patients with triple-negative breast cancer.* J Xray Sci Technol, 2019. **27**(3): p. 485-492.

39. Danala, G., et al., *Classification of breast masses using a computer-aided diagnosis scheme of contrast enhanced digital mammograms.* Annals of biomedical engineering, 2018. **46**(9): p. 1419-1431.

40. Sun, Z.-Q., et al., *Radiomics study for differentiating gastric cancer from gastric stromal tumor based on contrast-enhanced CT images.* Journal of X-ray Science and Technology, 2019. **27**(6): p. 1021-1031.

41. Gong, J., et al., *Fusion of quantitative imaging features and serum biomarkers to improve performance of computer-aided diagnosis scheme for lung cancer: A preliminary study.* Med Phys, 2018. **45**(12): p. 5472-5481.

42. Chen, X., et al., *Developing a new radiomics-based CT image marker to detect lymph node metastasis among cervical cancer patients.* Computer Methods and Programs in Biomedicine, 2020. **197**: p. 105759.

43. Shen, C., et al., *2D and 3D CT radiomics features prognostic performance comparison in non-small cell lung cancer.* Translational oncology, 2017. **10**(6): p. 886-894.

44. Zheng, B., et al., *A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment.* Med Phys, 2006. **33**(1): p. 111-7.

45. Cao, Y., et al., *Quantitative analysis of brain optical images with 2D C0 complexity measure.* Journal of Neuroscience Methods, 2007. **159**(1): p. 181-186.

46. Sarkar, N. and B.B. Chaudhuri, *An efficient differential box-counting approach to compute fractal dimension of image.* IEEE Transactions on Systems, Man, and Cybernetics, 1994. **24**(1): p. 115-120.

47. Park, S.C., X.-H. Wang, and B. Zheng, *Assessment of performance improvement in content-based medical image retrieval schemes using fractal dimension.* Academic Radiology, 2009. **16**(10): p. 1171-1178.

48. Ojala, T., M. Pietikäinen, and D. Harwood, *A comparative study of texture measures with classification based on featured distributions.* Pattern Recognition, 1996. **29**(1): p. 51-59.

49. Manjunath, B.S., et al., *Color and texture descriptors.* Ieee Transactions on Circuits and Systems for Video Technology, 2001. **11**(6): p. 703-715.

50. Dalal, N. and B. Triggs. *Histograms of oriented gradients for human detection*. in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. 2005.

51. Burt, P. and E. Adelson, *The Laplacian Pyramid as a Compact Image Code.* IEEE Transactions on Communications, 1983. **31**(4): p. 532-540.

52. Walnut, D.F., *An Introduction to Wavelet Analysis*. 1st ed. 2004, New York: Springer.

53. Chengjun, L. and H. Wechsler, *Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition.* Ieee Transactions on Image Processing, 2002. **11**(4): p. 467-476.

54. Manjunath, B.S. and W.Y. Ma, *Texture features for browsing and retrieval of image data.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 1996. **18**(8): p. 837-842.

55. Porwik, P. and A. Lisowska, *The Haar-wavelet transform in digital image processing: its status and achievements.* Machine graphics and vision, 2004. **13**(1/2): p. 79-98.

56. Lina, J.-M. and M. Mayrand, *Complex Daubechies Wavelets.* Applied and Computational Harmonic Analysis, 1995. **2**(3): p. 219-229.

57. Oliva, A. and A. Torralba, *Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope.* International Journal of Computer Vision, 2001. **42**(3): p. 145-175.

58. Witten, I.H., et al., *Data Mining: Practical Machine Learning Tools and Techniques*. 4th ed. 2017, Singapore: Elsevier.

59. Vapnik, V.N., ed. *Statistical learning theory*. 1998, Wiley: New York.

60. Jolliffe, I.T., *Principal Component Analysis*. 4th ed. 2002, New York: Springer.

61. Aghaei, F., et al., *Applying a new quantitative global breast MRI feature analysis scheme to assess tumor response to chemotherapy.* J Magn Reson Imaging, 2016. **44**(5): p. 1099-1106.

62. Yu, K.-H., et al., *Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features.* Nature Communications, 2016. **7**(1): p. 12474.

63. Du, M., et al., *Plasma exosomal miRNAs-based prognosis in metastatic kidney cancer.* Oncotarget, 2017. **8**(38): p. 63703-63714.

64. Zhang, Y., et al., *Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer.* Sci Rep, 2017. **7**: p. 46349.

65. Zhang, S., et al., *Improvement in prediction of prostate cancer prognosis with somatic mutational signatures.* J Cancer, 2017. **8**(16): p. 3261-3267.

66. Chen, H.-C., et al., *Assessment of performance of survival prediction models for cancer prognosis.* BMC medical research methodology, 2012. **12**: p. 102-102.

67. Wang, S., et al., *Preoperative computed tomography-guided disease-free survival prediction in gastric cancer: a multicenter radiomics study.* Med Phys, 2020.

68. Taghavi, M., et al., *Machine learning-based analysis of CT radiomics model for prediction of colorectal metachronous liver metastases.* Abdominal Radiology, 2020.

69. Zhang, B., et al., *Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma.* Cancer Letters, 2017. **403**: p. 21-27.

70. Tan, M., J. Pu, and B. Zheng, *Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model.* Int J Comput Assist Radiol Surg, 2014. **9**(6): p. 1005-20.

71. Zheng, B., et al., *Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm.* Academic Radiology, 1999. **6**(6): p. 327-332.

72. Wang, J., et al., *Discrimination of Breast Cancer with Microcalcifications on Mammography by Deep Learning.* Scientific Reports, 2016. **6**(1): p. 27327.

73. Du, Y., et al., *Classification of Tumor Epithelium and Stroma by Exploiting Image Features Learned by Deep Convolutional Neural Networks.* Ann Biomed Eng, 2018. **46**(12): p. 1988-1999.

74. Zhao, X., et al., *Deep CNN models for pulmonary nodule classification: Model modification, model integration, and transfer learning.* J Xray Sci Technol, 2019. **27**(4): p. 615-629.

75. Berenguer, R., et al., *Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters.* Radiology, 2018. **288**(2): p. 407-415.

76. Mackin, D., et al., *Measuring Computed Tomography Scanner Variability of Radiomics Features.* Invest Radiol, 2015. **50**(11): p. 757-65.

77. Schell, M.J., et al., *Evidence-based target recall rates for screening mammography.* Radiology, 2007. **243**(3): p. 681-9.

78. Ko, J.M., et al., *Prospective Assessment of Computer-Aided Detection in Interpretation of Screening Mammography.* American Journal of Roentgenology, 2006. **187**(6): p. 1483-1491.

79. Zheng, B., et al., *Computer-aided detection of breast masses depicted on full-field digital mammograms: a performance assessment.* The British Journal of Radiology, 2012. **85**(1014): p. e153-e161.

80. Wang, X., et al., *An Interactive System for Computer-Aided Diagnosis of Breast Masses.* Journal of Digital Imaging, 2012. **25**(5): p. 570-579.

81. Fenton, J.J., et al., *Influence of Computer-Aided Detection on Performance of Screening Mammography.* New England Journal of Medicine, 2007. **356**(14): p. 1399-1409.

82. Nishikawa, R.M. and D. Gur, *CADe for Early Detection of Breast Cancer—Current Status and Why We Need to Continue to Explore New Approaches.* Academic Radiology, 2014. **21**(10): p. 1320-1321.

83. Zheng, B., et al., *Computer-Aided Detection: The Effect of Training Databases on Detection of Subtle Breast Masses.* Academic Radiology, 2010. **17**(11): p. 1401-1408.

84. Oliver, A., et al., *A review of automatic mass detection and segmentation in mammographic images.* Medical Image Analysis, 2010. **14**(2): p. 87-110.

85. Gundreddy, R.R., et al., *Assessment of performance and reproducibility of applying a content-based image retrieval scheme for classification of breast lesions.* Medical Physics, 2015. **42**(7): p. 4241-4249.

86. Wang, Y.Z., et al., *Computer-aided classification of mammographic masses using visually sensitive image features.* Journal of X-Ray Science and Technology, 2017. **25**(1): p. 171-186.

87. Carney, P.A., et al., *Individual and Combined Effects of Age, Breast Density, and Hormone Replacement Therapy Use on the Accuracy of Screening Mammography.* Annals of Internal Medicine, 2003. **138**(3): p. 168-175.

88. Qiu, Y., et al., *A new approach to develop computer-aided diagnosis scheme of breast mass classification using deep learning technology.* Journal of X-ray science and technology, 2017. **25**(5): p. 751-763.

89. Gulshan, V., et al., *Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus PhotographsAccuracy of a Deep Learning Algorithm for Detection of Diabetic RetinopathyAccuracy of a Deep Learning Algorithm for Detection of Diabetic Retinopathy.* JAMA, 2016. **316**(22): p. 2402-2410.

90. Yan, S., et al., *Improving Performance of Breast Cancer Risk Prediction by Incorporating Optical Density Image Feature Analysis: An Assessment.* Academic Radiology, 2017.

91. Heidari, M., et al. *Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm*. Physics in medicine and biology, 2018. **63**, 035020 DOI: 10.1088/1361-6560/aaa1ca.

92. Chen, X., et al., *Applying a new quantitative image analysis scheme based on global mammographic features to assist diagnosis of breast cancer.* Computer methods and programs in biomedicine, 2019. **179**: p. 104995.

93. Wang, X., et al., *Computerized prediction of risk for developing breast cancer based on bilateral mammographic breast tissue asymmetry*. Medical engineering & physics, 2011. **33**(8): p. 934-942.

94. Danala, G., et al., *Applying Quantitative CT Image Feature Analysis to Predict Response of Ovarian Cancer Patients to Chemotherapy.* Academic radiology, 2017. **24**(10): p. 1233-1239.

95. Zargari, A., et al., *Prediction of chemotherapy response in ovarian cancer patients using a new clustered quantitative image marker.* Physics in Medicine & Biology, 2018. **63**(15): p. 155020.

96. Kourou, K., et al., *Machine learning applications in cancer prognosis and prediction.* Computational and structural biotechnology journal, 2014. **13**: p. 8-17.

97. Vapnik, V.N., *An overview of statistical learning theory.* IEEE Trans Neural Netw, 1999. **10**(5): p. 988-99.

98. Tan, M., J. Pu, and B. Zheng, *Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model.* International journal of computer assisted radiology and surgery, 2014. **9**(6): p. 1005-1020.

99. Lederman, D., et al., *Improving Breast Cancer Risk Stratification Using Resonance-Frequency Electrical Impedance Spectroscopy Through Fusion of Multiple Classifiers.* Annals of Biomedical Engineering, 2011. **39**(3): p. 931-945.

100. Metz, C.E., B.A. Herman, and J.H. Shen, *Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data.* Statistics in Medicine, 1998. **17**(9): p. 1033-1053.

101. Sun, H., et al., *Performance evaluation of breast cancer diagnosis with mammography, ultrasonography and magnetic resonance imaging.* J Xray Sci Technol, 2018. **26**(5): p. 805-813.

102. Giger, M.L., N. Karssemeijer, and J.A. Schnabel, *Breast Image Analysis for Risk Assessment, Detection, Diagnosis, and Treatment of Cancer*, in *Annual Review of Biomedical Engineering, Vol 15*, M.L. Yarmush, Editor. 2013, Annual Reviews: Palo Alto. p. 327-357.

103. Brzakovic, D., X.M. Luo, and P. Brzakovic, *An approach to automated detection of tumors in mammograms.* IEEE Transactions on Medical Imaging, 1990. **9**(3): p. 233-241.

104. Cheng, H.D., et al., *Approaches for automated detection and classification of masses in mammograms.* Pattern Recognition, 2006. **39**(4): p. 646-668.

105. Robnik-Šikonja, M. and I. Kononenko, *Theoretical and Empirical Analysis of ReliefF and RReliefF.* Machine Learning, 2003. **53**(1): p. 23-69.

106. Guyon, I., et al., *Gene Selection for Cancer Classification using Support Vector Machines.* Machine Learning, 2002. **46**(1): p. 389-422.

107. Chandrashekar, G. and F. Sahin, *A survey on feature selection methods.* Computers & Electrical Engineering, 2014. **40**(1): p. 16-28.

108. Battiti, R., *Using mutual information for selecting features in supervised neural net learning.* IEEE Transactions on Neural Networks, 1994. **5**(4): p. 537-550.

109. Constantinidis, A.S., M.C. Fairhurst, and A.F.R. Rahman, *A new multi-expert decision combination algorithm and its application to the detection of circumscribed masses in digital mammograms.* Pattern Recognition, 2001. **34**(8): p. 1527-1537.

110. Zheng, B., Y.-H. Chang, and D. Gur, *Computerized detection of masses in digitized mammograms using single-image segmentation and a multilayer topographic feature analysis.* Academic radiology, 1995. **2**(11): p. 959-966.

111. Qiu, Y., et al. *An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology*. in *Medical Imaging 2016: Computer-Aided Diagnosis*. 2016. International Society for Optics and Photonics.

112. Heidari, M., et al., *Development and assessment of a new global mammographic image feature analysis scheme to predict likelihood of malignant cases.* IEEE transactions on medical imaging, 2019. **39**(4): p. 1235-1244.

113. Shen, L., et al., *Deep learning to improve breast cancer detection on screening mammography.* Scientific reports, 2019. **9**(1): p. 1-12.

114. Dhungel, N., G. Carneiro, and A.P. Bradley. *The automated learning of deep features for breast mass classification from mammograms*. in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2016. Springer.

115. Perez, L. and J. Wang, *The effectiveness of data augmentation in image classification using deep learning.* arXiv preprint arXiv:1712.04621, 2017.

116. Shorten, C. and T.M. Khoshgoftaar, *A survey on image data augmentation for deep learning.* Journal of Big Data, 2019. **6**(1): p. 1-48.

117. Korkinof, D., et al., *High-resolution mammogram synthesis using progressive generative adversarial networks.* arXiv preprint arXiv:1807.03401, 2018.

118. Frid-Adar, M., et al., *GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification.* Neurocomputing, 2018. **321**: p. 321-331.

119. Wu, E., et al., *Conditional infilling GANs for data augmentation in mammogram classification*, in *Image analysis for moving organ, breast, and thoracic images*. 2018, Springer. p. 98-106.

120. Radford, A., L. Metz, and S. Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks.* arXiv preprint arXiv:1511.06434, 2015.

121. Saxena, D. and J. Cao, *Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions.* arXiv preprint arXiv:2005.00065, 2020.

122. Li, C., et al., *Triple generative adversarial nets.* arXiv preprint arXiv:1703.02291, 2017.

123. Chen, X., et al. *Development of a transferring GAN based CAD scheme for breast mass classification: an initial study*. in *Biophotonics and Immune Responses XVI*. 2021. International Society for Optics and Photonics.

124. Goodfellow, I.J., et al., *Generative adversarial networks.* arXiv preprint arXiv:1406.2661, 2014.

125. Salimans, T., et al., *Improved techniques for training gans.* arXiv preprint arXiv:1606.03498, 2016.

126. Gulrajani, I., et al., *Improved training of wasserstein gans.* arXiv preprint arXiv:1704.00028, 2017.

127. Che, T., et al., *Mode regularized generative adversarial networks.* arXiv preprint arXiv:1612.02136, 2016.

128. Wang, Y., et al. *Transferring gans: generating images from limited data*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

129. Heusel, M., et al., *Gans trained by a two time-scale update rule converge to a local nash equilibrium.* arXiv preprint arXiv:1706.08500, 2017.

130. Sun, C., et al. *Revisiting unreasonable effectiveness of data in deep learning era*. in *Proceedings of the IEEE international conference on computer vision*. 2017.

131. Samala, R.K., et al., *Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets.* IEEE transactions on medical imaging, 2018. **38**(3): p. 686-696.

132. Tarvainen, A. and H. Valpola. *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*. in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017.

133. Miyato, T., et al., *Virtual adversarial training: a regularization method for supervised and semi-supervised learning.* IEEE transactions on pattern analysis and machine intelligence, 2018. **41**(8): p. 1979-1993.

134. Xie, Q., et al. *Unsupervised Data Augmentation for Consistency Training*. in *Advances in Neural Information Processing Systems*. 2020.

135. Chen, X., et al., *Recent advances and clinical applications of deep learning in medical image analysis.* arXiv preprint arXiv:2105.13381, 2021.

136. Pang, T., et al., *Semi-supervised GAN-based Radiomics Model for Data Augmentation in Breast Ultrasound Mass Classification.* Computer Methods and Programs in Biomedicine, 2021. **203**: p. 106018.

137. Cao, X., et al., *Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation.* IEEE Transactions on Medical Imaging, 2020. **40**(1): p. 431-443.

138. Berthelot, D., et al. *MixMatch: A Holistic Approach to Semi-Supervised Learning*. in *Advances in Neural Information Processing Systems*. 2019.

139. Wang, X., et al., *Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification.* Medical image analysis, 2021. **70**: p. 102010.

140. Chen, X., et al., *Virtual Adversarial Training for Semi-supervised Breast Mass Classification.* arXiv preprint arXiv:2201.10675, 2022.

141. Geras, K.J., R.M. Mann, and L. Moy, *Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives.* Radiology, 2019. **293**(2): p. 246-259.

142. Yang, Z., et al. *Learning to navigate for fine-grained classification*. in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

143. Liu, L., et al., *Deep Learning for Generic Object Detection: A Survey.* International Journal of Computer Vision, 2020. **128**(2): p. 261-318.

144. Lin, T., et al. *Feature Pyramid Networks for Object Detection*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

145. Duan, K., et al. *CenterNet: Keypoint Triplets for Object Detection*. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.

146. Milletari, F., N. Navab, and S. Ahmadi. *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation*. in *2016 Fourth International Conference on 3D Vision (3DV)*. 2016.

147. Sudre, C.H., et al., *Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations*, in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. 2017, Springer. p. 240-248.

148. Abraham, N. and N.M. Khan. *A novel focal tversky loss function with improved attention u-net for lesion segmentation*. in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019. IEEE.

149. Kervadec, H., et al. *Boundary loss for highly unbalanced segmentation*. in *International conference on medical imaging with deep learning*. 2019. PMLR.

150. Li, M., et al., *SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network.* IEEE transactions on medical imaging, 2020. **39**(7): p. 2289-2301.

151. Chen, J., et al., *Transunet: Transformers make strong encoders for medical image segmentation.* arXiv preprint arXiv:2102.04306, 2021.

152. Salimans, T., et al., *Improved techniques for training GANs*, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, Curran Associates Inc.: Barcelona, Spain. p. 2234–2242.

153. Dai, Z., et al., *Good semi-supervised learning that requires a bad GAN*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 6513–6523.

154. Li, C., et al., *Triple generative adversarial nets*, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 4091–4101.

155. Rubin, M., et al., *TOP-GAN: Stain-free cancer cell classification using deep learning with a small training set.* Medical Image Analysis, 2019. **57**: p. 176-185.

156. Zhao, S., et al. *Differentiable Augmentation for Data-Efficient GAN Training*. in *Advances in Neural Information Processing Systems*. 2020.

157. Karras, T., et al. *Training Generative Adversarial Networks with Limited Data*. in *Advances in Neural Information Processing Systems*. 2020.

158. Sohn, K., et al. *FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence*. in *Advances in Neural Information Processing Systems*. 2020.

159. Cubuk, E.D., et al. *Randaugment: Practical automated data augmentation with a reduced search space*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020.

160. Oliver, A., et al., *Realistic evaluation of deep semi-supervised learning algorithms.* arXiv preprint arXiv:1804.09170, 2018.

161. Guo, L.-Z., et al. *Safe deep semi-supervised learning for unseen-class unlabeled data*. in *International Conference on Machine Learning*. 2020. PMLR.

162. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems, 2012. **25**: p. 1097-1105.

163. He, K., et al. *Deep Residual Learning for Image Recognition*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

164. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.

165. Vaswani, A., et al., *Attention is all you need,* in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 6000–6010.

166. Yuille, A.L. and C. Liu, *Deep nets: What have they ever done for vision?* International Journal of Computer Vision, 2021. **129**(3): p. 781-802.

167. Dosovitskiy, A., et al., *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929, 2020.

168. Marblestone, A.H., G. Wayne, and K.P. Kording, *Toward an integration of deep learning and neuroscience.* Frontiers in computational neuroscience, 2016. **10**: p. 94.

169. Zoph, B. and Q.V. Le, *Neural architecture search with reinforcement learning.* arXiv preprint arXiv:1611.01578, 2016.

170. Elsken, T., J.H. Metzen, and F. Hutter, *Neural architecture search: A survey.* The Journal of Machine Learning Research, 2019. **20**(1): p. 1997-2017.

171. Gibson, E., et al., *NiftyNet: a deep-learning platform for medical imaging.* Computer methods and programs in biomedicine, 2018. **158**: p. 113-122.

172. Isensee, F., et al., *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nature Methods, 2021. **18**(2): p. 203-211.

173. Zhou, Z., et al., *Models Genesis.* Medical Image Analysis, 2021. **67**: p. 101840.

174. Zhou, Z., et al. *Models Genesis: Generic Autodidactic Models for 3D Medical Image Analysis*. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. 2019. Cham: Springer International Publishing.

175. Azizi, S., et al. *Big Self-Supervised Models Advance Medical Image Classification*. in *arXiv preprint arXiv:2101.05224*. 2021.

176. Vu, Y.N.T., et al. *MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation*. in *arXiv preprint arXiv:2102.10663*. 2021.

177. Zhang, Y., et al. *Contrastive Learning of Medical Visual Representations from Paired Images and Text*. in *arXiv preprint arXiv:2010.00747*. 2020.

178. Wynants, L., et al., *Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal.* BMJ, 2020. **369**: p. m1328.

179. Roberts, M., et al., *Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans.* Nature Machine Intelligence, 2021. **3**(3): p. 199-217.

180. Nagendran, M., et al., *Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies.* BMJ, 2020. **368**: p. m689.

181. Yang, Q., et al., *Federated machine learning: Concept and applications.* ACM Transactions on Intelligent Systems and Technology (TIST), 2019. **10**(2): p. 1-19.

182. Li, T., et al., *Federated learning: Challenges, methods, and future directions.* IEEE Signal Processing Magazine, 2020. **37**(3): p. 50-60.

183. Rieke, N., et al., *The future of digital health with federated learning.* NPJ digital medicine, 2020. **3**(1): p. 1-7.

184. Kelly, C.J., et al., *Key challenges for delivering clinical impact with artificial intelligence.* BMC medicine, 2019. **17**(1): p. 1-9.

185. Baltatzis, V., et al. *The Pitfalls of Sample Selection: A Case Study on Lung Nodule Classification*. in *Predictive Intelligence in Medicine*. 2021. Cham: Springer International Publishing.