2017-01-01

Deep Learning Method Vs. Hand-Crafted Features For Lung Cancer Diagnosis And Breast Cancer Risk Analysis

Wenqing Sun
University of Texas at El Paso, wenqingsun12@gmail.com

DEEP LEARNING METHOD VS. HAND-CRAFTED FEATURES FOR LUNG CANCER DIAGNOSIS AND BREAST CANCER RISK ANALYSIS

WENQING SUN

Doctoral Program in Electrical and Computer Engineering

APPROVED:

Wei Qian, Ph.D., Chair

Bill (Tzu-Liang) Tseng, Ph.D., Co-chair

Thompson Sarkodi-Gyan, Ph.D.

Naijun Sha, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School

DEEP LEARNING METHOD VS. HAND-CRAFTED FEATURES FOR LUNG CANCER DIAGNOSIS AND BREAST CANCER RISK ANALYSIS

by

WENQING SUN, M.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Electrical and Computer Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

May 2017

Acknowledgements

To my parents and family, thank you for your ultimate sacrifices, shedding light on me even on the darkest days, and all the efforts your made to see me succeed.

&

To my committee members, thank you for your support and providing me all these resources and opportunities.

&

To all of my friends and 15-year-old myself: do what you can't, good is perfect, one step at a time.

Abstract

Breast cancer and lung cancer are two major leading causes of cancer deaths, and researchers have been developing computer aided diagnosis (CAD) system to automatically diagnose them for decades. In recent studies, we found that the techniques in CAD system can also be used for breast cancer risk analysis, like feature design and machine learning. Also we noticed that with the development of deep learning methods, the performance of CAD system can be improved by using computer automatically generated features. To explore these possibilities, we conducted a series of studies: the first two studies focused on transferring the original CAD system techniques to breast cancer risk analysis models; and the next two studies compared the performance of our proposed schemes using deep learning methods and traditional methods on breast cancer risk analysis and lung cancer diagnosis.

Table of Contents

List of Tables

List of Figures

List of Illustrations

Chapter 1: Introduction

1.1 Breast cancer and lung cancer

Breast Cancer is the most common cancer and second leading cause of cancer deaths of women (R. a Smith et al. 2015). Scientific evidence has shown that early cancer detection is important to enhance the survival rates of the patients through more effective patient management and treatment (R. A. Smith et al. 2015)(Madigan et al. 1995). Since the majority of breast cancers are detected in women with no known risk factors defined in the existing epidemiology models (Amir et al. 2010)(Boyd et al. 2005), a uniform mammography screening program in the general population is currently applied and considered important (Madigan et al. 1995). Lung cancer is the leading cause of cancer death for both men and women worldwide. The American Cancer Society (ACS) reported that the early detection of lung cancer, stage 1, could significantly increase the survival rate from 2% to 47% compared to the detection at stage 5. However, only 15% of early stage lung cancers are detected. Computer-aided diagnosis (CADx) system has the potential to aid radiologists as a second reader and attracted much attention in the last few decades. Computed tomography (CT) is typically used for lung cancer screening and diagnosis in clinic. A single CT examination can generate up to 700 axial images creating a challenging task for image interpretation.

However, due to the large variability in the depiction of breast abnormalities, the overlapping dense fibroglandular tissue on the projection images and the low cancer prevalence in the screening environment, both detection sensitivity and specificity of screening mammography are relatively low (Pinsky and Helvie 2015).

A number of recently reported studies made the debate related to the efficacy (risk-benefit and cost-benefit) of population-based screening mammography more controversial (Hendrick and Helvie 2011)(Jørgensen 2012).

1.2 Computer aided analysis system

To help radiologists improve detection and diagnosis performances in reading and interpreting screening mammograms, a great amount of research has been conducted to develop CAD systems or schemes including our work on developing a variety of two-dimensional computerized image analysis algorithms optimized to enhance the performance of the traditional CAD systems during the last two decades (e.g., (Glide-Hurst, Duric, and Littrup 2007)(W Qian, Li, and Clarke 1999)(X. Sun, Qian, and Song 2004)). Currently, a number of commercialized CAD schemes, including the one originally developed in our group and then being licensed to Carestream Health Inc., are widely used in the clinical practice to assist radiologists in reading and interpreting mammograms to date. Illustration 1.1 shows a typical CAD structure for breast cancer risk analysis, which is used in our previous studies.

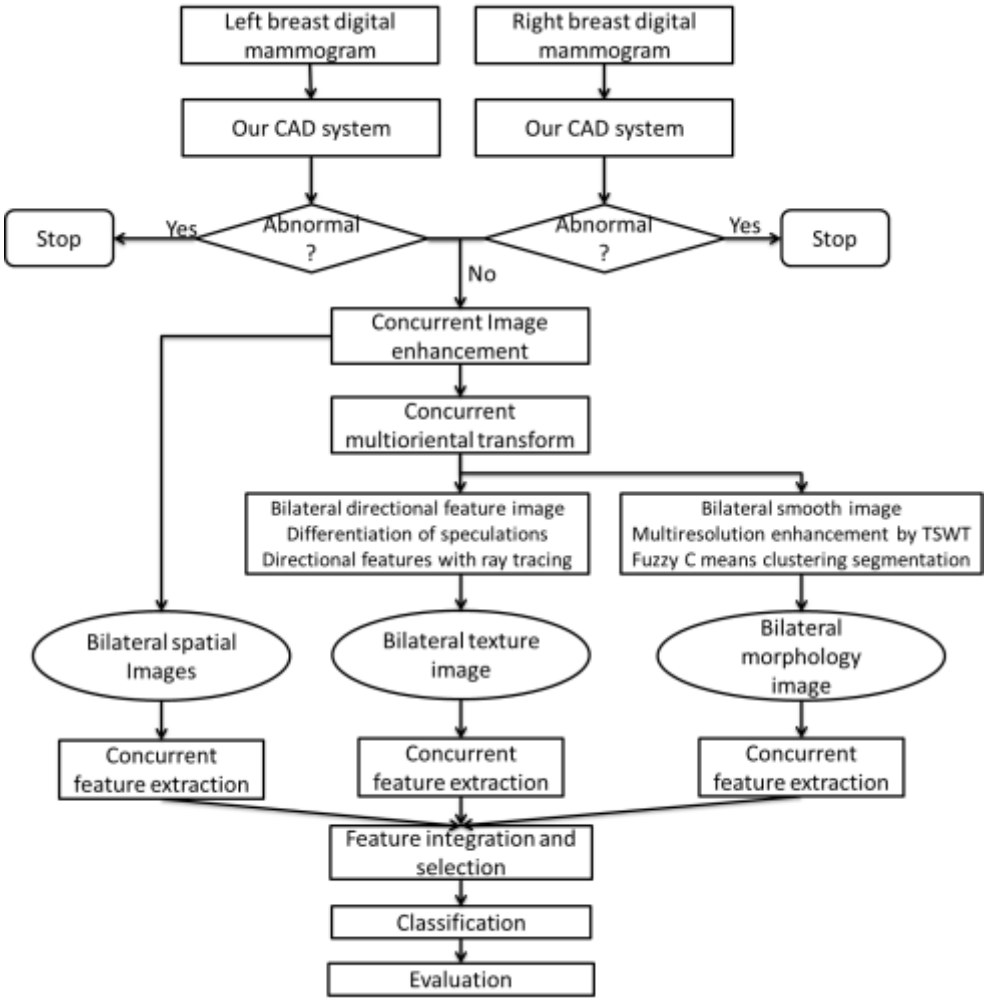

Illustration 1.1: A typical CAD structure for breast cancer risk analysis

A typical CAD system can be divided into several different modules: preprocessing module, segmentation module, feature extraction module, classification module. The first module is to generate the best resolution image and remove noise for further processing; the second module is to identify and segment the region of interest (ROI) from the original image; the third module is to extract the computational features from the ROIs or specific areas; the last module is to select the features and build a classifier based on the selected features. The details of each module will be presented in the following sections.

1.3 Hand-crafted features

Designing and extracting features is the core part of a CAD system. We can generally divide the features into three feature groups or feature domains: morphological features, density features and texture features. The morphological features identify the shape information of the suspicious regions, the density describes the distribution of the pixel wise density distribution, the texture features are the most complicated feature type which describe the texture characters of the regions. Below we gave a brief introduction of texture features for five groups of typical texture features: grey level co-occurrence matrix (GLCM) feature, local binary pattern (LBP) feature, scale-invariant feature transform (SIFT) feature, steerable feature, and wavelet feature.

The first group is GLCM feature. In grey level co-occurrence matrix, the number of rows and columns is equal to the number of gray levels in original image, and each element represent the relative frequency of two pixels with given intensity separated by a pixel distance. Then the mean and standard deviation were computed for every matrix(W. Sun, Tseng, Qian, et al. 2015).

The second group is LBP feature. It was first proposed as a texture descriptor for 2D images (Ojala, Pietikäinen, and Mäenpää 2002). From the equally divided blocks of the original image, the extracted LBP features can describe the local and global textures. We computed the LBP features by comparing each pixel with the 8 neighbor pixels, and it returns the 8 digits

binary code for each pixel. Then the local binary pattern feature histogram was calculated from the coded image.

SIFT features are the third feature group. This scale invariant feature transform is invariant to uniform scaling, orientation, and it is also partially invariant to affine distortion and illumination changes (Barley and Town 2014). Because of this property, it is a classical algorithm in object matching and action recognition. When creating descriptors, the sub-region was considered in our algorithm instead of every pixel, thus it is more adaptive to the image noises and subtle distortions.

Steerable features are in group four. The steerable filters are linear combination of a set of basis filters with arbitrary orientations (Freeman and Adelson 1991). The feature group five contains wavelet features, and they provide the spatial and frequency representation of the images. In wavelet transform, high-pass and low-pass filters are used in two dimensions. For each scale, one input image can generate eight sub-band images: HH, HL, LH, and LL, where H represent high frequency band and L represent low frequency band (W Qian, Li, and Clarke 1999).

1.4 Deep learning methods

Inspired by the human brain's architecture, deep learning algorithms have been attracting more and more researchers' attentions in the past ten years (Hinton, Osindero, and Teh 2006). Compared to traditional machine learning algorithms with shallow architectures, the deep learning algorithms are organized in a deep structure with several levels of composition of non-linear operations in the learned functions. Like the architecture depth of the brain, a given input percept represented at multiple levels of abstraction in the algorithm, and each level corresponds to a different area of cortex (Bengio 2009). This architecture allows the algorithm to automatically learn features at multiple levels of abstraction so that it can generate the complex functions linking the input to the categories directly from raw data without using human-crafted features (i.e. manually designed features). The features extracted higher level of the hierarchies is

composed by the weighted combination of the features of lower levels, and each level contains hundreds or thousands of neurons. In computer vision area, feature designing is the most challenging and time consuming part, and the ability to automatically extracting features from raw data is extremely attractive especially in the age of big data.

The primary task for big data analytics is discriminative analysis. In recent years with the development of cloud storage and the explosion of big data, the sizes of available digital image collections have been increasing rapidly. These images were generated from a variety of sources such as social networks, cloud services, global positioning satellites, clinical imaging systems, military surveillance, and security systems (Najafabadi et al. 2015). To efficiently classify these massive image collections, automatically extracting semantic information from the images with deep structured algorithms is the most popular and promising method. Their advantages in constructing high level complicated representations from multiple domains of the original image captured the attention from big companies like Facebook, Google, IBM to invest millions of dollars every year for deep learning research, and recently has begun topping the companies benefit (Najafabadi et al. 2015)(Lohr 2012). Learning from the deep, layered and hierarchical models of data, numerous study results show that deep learning algorithms can outperform the traditional machine learning models in many different tasks, including speech recognition [5][6][7], computer vision (Hinton, Osindero, and Teh 2006)(Bengio et al. 2007) (Krizhevsky, Sutskever, and Hinton 2012), and natural language processing (Mikolov et al. 2011)(Socher, Huang, and Pennington 2011)(Bordes et al. 2012). Compared to its early stage, deep learning applications have already extended from simple image classifications like handwritten numbers recognition to more complicated classification tasks. In ImageNet LSVRC-2012 contest, the winner group successfully designed the deep learning algorithm and classified 1.2 million high-resolution images into 1000 different classes at the error rate of 15.3%, compared to 26.2% reported by the second-best group (Krizhevsky, Sutskever, and Hinton 2012). Dean et al. (Dean et al. 2012) used large-scale software infrastructure based deep learning models made further success on a visual object recognition task with 16 million images and 21k categories. In another

contest, deep learning algorithms beat other algorithms and won MICCAI 2013 Grand Challenge and ICPR 2012 Contest on Mitosis Detection (Cireşan et al. 2013). In 2015, Shen et al. (Shen et al. 2015b) diagnosed lung cancer on LIDC database using a multi-scale two layer convolutional neural network, and the reported accuracy was 86.84%. Kumar et al. (D. Kumar, Wong, and Clausi 2015) tested their algorithm using deep features extracted from autoencoder on 157 cases from the same dataset, and reached the accuracy of 75.01%.

Chapter 2: Breast cancer risk analysis

2.1 RATIONALE:

Breast cancer risk assessment has been studied for more than 30 years, and it has been reported that the promotion of breast cancer screening has significantly reduced mortality rates of breast cancer patients along with the advancement of cancer treatment methods (R. A. Smith, Duffy, and Tabar 2012). But the low efficacy and/or potential harms of the current population-based screening mammography cannot be overlooked, including a high false-positive recall rates and over-diagnosis (Pace and Keating 2014). False positive recalls raised due to the suspicions detected on the mammograms often lead to the further tests (i.e., additional imaging or biopsy procedures), which bring the woman the unnecessary mental, physical and financial burden. Over-diagnosis is the detection of a tumor through screening that would have become clinically apparent but would not have shortened a woman's life (Nelson et al. 2009), which is a major concern in clinic. Thus, the analysis of cancer risk has become increasingly important for establishing an optimal personalized screening recommendations (Wolfe 1976). To predict breast cancer at its early stage, many researchers have developed and tested different risk stratification models to predict cancer risk, including Gail model, Claus model, BRCAPRO model, Jonker model, etc. (M H Gail et al. 1989) (Claus, Risch, and Thompson 1991) (Parmigiani, Berry, and Aguilar 1998) (Jonker et al. 2003). These models show that several factors are associated with an increasing risk for developing breast cancer, such as woman's age, BRCA1 and BRCA2 gene mutations, breast density, body mass index (BMI), age at first birth, and family cancer history. In addition, numerous studies have shown the association between the breast density and cancer risk. The higher breast density is not only related to the decreased sensitivity of mammograms because of a masking effect (Pinsky and Helvie 2015), but is also the strongest independent breast cancer risk indicator with the exception of women's age and relative gene mutations that apply to a very small fraction of the population (Amir et al. 2010).

However, the measurements of breast density in the risk models mentioned above are based on visual assessment of mammographic density into four American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) density categories by the radiologists (Boyd et al. 2010). These subjective ratings are time-consuming and often inaccurate or inconsistent due to the large intra- and inter-observer variability (Berg et al. 2000). Thus, many research groups have developed computerized algorithms to segment and quantify breast tissue density based on a combination of computational features including the statistic texture features of the pixel values, the mathematical morphology and density based features, and the other higher order momentum based measurements computed from the original and/or processed images (X. Wang et al. 2011). Furthermore, several research groups have also tested the feasibility of utilizing the computed mammographic density information to develop computerized breast cancer risk prediction or analysis schemes (J. Wei et al. 2011).

Recently, American Cancer Society (ACS) issued a new breast cancer screening guideline (Oeffinger et al. 2015), which recommended biennial screening for women 55 years old. However, the guideline also keeps the option for women who want to continue annual screening. This creates a new dilemma of how an individual woman can optimally decide that she should screen annually or biennially. In order to address this dilemma, developing an effective cancer risk model is very important. Although a number of epidemiology-based breast cancer risk prediction models (Amir et al. 2010) (Irwig, Houssami, and van Vliet 2004) have been developed and tested, these models do not have clinically acceptable discriminatory power (or positive predictive value) to determine whether an individual woman should be mammography screened next year (Mitchell H. Gail and Mai 2010). As a result, developing new risk stratification model based on the quantitative mammographic image feature analysis, which can more accurately predict the risk of women having image-detectable cancer in a near-term (i.e., < 2 to 5 years) after a negative screening, has been attracting research interest (W. Sun, Tseng, Qian, et al. 2015) (Tan et al. 2013). Some researchers concluded that the biggest difficulty of cancer risk analysis using mammographic image feature analysis based models is to

accurately measure the percentage of the volume of dense tissue that is a three-dimensional (3D) object on the basis of two-dimensional (2D) mammograms (Kopans 2008). Since the mammography is a projection imaging technique, the mammogram is an overlapping of all tissues on a flat plane. Moreover, for a 3D object, the 2D projections from different angles are different, creating variations that affect the precision of measurement of dense tissue. For example, shell or bar shaped object may have huge projection area difference depending on the orientation of the object. If the features from one view are in common with the other view, the prediction results are more reliable. Combining the information from two mammographic views improves the possibility of detecting high-risk cases and abnormal regions, since the dense or suspicious regions may be partly or fully obscured in one projection by overlapping glandular tissues (including lobules and ducts) (Samulski and Karssemeijer 2011). Also, with the information from the additional perspective, superimposition of normal or non-dense structures simulating a suspicious region or dense region can be recognized, thus reducing the chance of false positive. Figure 2.1 shows three examples of the differences of dense tissues on the craniocaudual (CC) view and mediolateral oblique (MLO) view projections.

Figure 2.1: Three pairs of CC and MLO view mammograms showing the difference of the true volume, shape and texture characteristics of dense tissue seen from different views. a, b) CC and MLO view of a woman's breast. CC view shows a large dense area, but the obvious dense area is barely evident in MLO. c, d) Another pair of CC and MLO view mammograms, the shapes and orientations of the dense region are different. e, f) The third example of a pair of CC and MLO mammograms depicting different texture characteristics.

The idea of combining ipsilateral view mammograms has been used for computerized breast cancer detection and diagnosis, and the results were superior than using single view mammograms. In a recent paper, Maurice *et al*. (Samulski and Karssemeijer 2011) computed absolute differences of mass likelihood of each view, and used this as a similarity feature to detect the breast cancer. In the mass detection paper of Jun *et al*. (J. Wei et al. 2009), for each object, they used two scores from single view system (one score for each view) and one score from fused dual view system, and the final classifier was trained by these three scores. In another similar study by Xuejun *et al*. (X. Sun, Qian, and Song 2004), they combined the single view features and concurrent features together to train the classifier.

As part of our continuing research effort in this new near-term breast cancer risk assessment field, we hypothesize in this study that combing image features independently computed from the CC and MLO view mammograms together can achieve better performance of computerized risk analysis compared to using the single view images only. This hypothesis is based on several underlying scientific evidences and validated preliminary experimental observations. First, most early studies of tissue pattern and cancer risk are based on radiologist's subjective assessment of density patterns by using both CC and MLO projections (Nicholson et al. 2006). By comparing both MLO view and CC view mammograms, the radiologists can have some intuitive and non-quantitative information of the 3D distribution of the dense tissues. Second, the texture characteristics can be different on these two projections, as occlusion of the dense region by glandular tissue is very common. Third, most CC view mammogram images miss some volume of the breasts, MLO view can be used as a supplement (Kopans 2008). Fourth, computerized quantitative measurements are more reliable and consistent in evaluating mammographic tissue density by eliminating inter-observer variability (J. Wei et al. 2011).

This study is an extension of our previous studies using single view mammogram for each breast (W. Sun et al. 2014). We developed and extracted a group of features from both CC view and MLO view mammograms, and analyzed and compared different methods of how to efficiently use these features to stratify the women into high risk and low risk groups. The

feature fusion method and similarity test were developed and applied to each pair of mammograms, and thus our proposed scheme can adaptively choose the best strategy to analyze each case. The details of methodologies, experimental designs and results are described in the flowing sections.

2.2 MATERIALS

The dataset used in this study includes full-field digital mammography (FFDM) images acquired from 392 women. All these cases were randomly selected from our pre-established database, and for every woman at least two routine screening examinations preceding diagnosis were available. In each examination there are four FFDM images representing both CC and MLO views from left and right breast. The participants aged between 44 and 71 years old, with a mean age of 56.5 years and median age of 55 years. All these "prior" FFDM screening examinations were interpreted as "negative" or "definitely benign" (without recall) by radiologists in the original screening practice. The images we used for analysis in this study were all "prior" FFDM image. If more than two screening examinations existed, we preferentially used the most recent screening mammograms up to three years prior to the date of latest examinations or cancer being detected. Based on the detection and diagnostic status change in the latest "current" FFDM examinations and the verified pathology reports (if the cases were recalled and biopsied), we split the cases into three subgroups. Of these 392 cases, 190 were from benign or negative cases in "current" FFDM examinations and marked as group 1. The second group included the 104 cases that were diagnosed as cancer which were detected and verified 6-24 months after the "prior" FFDM examinations, but no later than the "current" examinations. The cancer was detected less than 12 months after the negative screening in 50 out of 104 cases, and for the other 54 cases, the cancer was found after 12 months. The remaining 98 cases were group 3 and they were recalled in the "current" examinations due to suspicious findings in mammography. Community radiologists at each local site classified breast density for each screening mammogram as part of routine clinical practice. Among these 392 cases, 34.9%

(137/392) and 55.9% (219/392), were rated as scattered fibro-glandular (BIRADS 2) and heterogeneously dense (BIRADS 3), respectively. Questionnaires were also provided to all the participants to obtain information of common risk factors such as woman's age, body mass index (BMI), family medical and cancer history.

2.3 PREPROCESSING AND DENSE REGION EXTRACTION

The first step of our scheme is preprocessing aiming to reduce image noises and artifacts, and thus enhance the original image. For this purpose, a hybrid processing scheme including the adaptive tree-structured nonlinear filtering (TSF) method, directional wavelet transform (DWT) and tree structured wavelet transform (TSWT), was applied to every mammogram. The detailed procedures have been reported in our previous publications (W. Sun et al. 2014) (Wei Qian et al. 1994) (W Qian et al. 2000).

For both CC and MLO view mammograms, the air-tissue interface of the breast was identified by the threshold based on the gray level intensity histogram (Keller et al. 2012) and then eliminated from breast area. For MLO view mammograms, the chest wall (including pectoral muscles) was detected by Hough transform and also removed from the image. Then, the breast area was segmented from the mammogram.

From our previous studies we found that some features extracted from dense regions can achieve better performance in risk prediction than the features extracted from the whole breast area (W. Sun, Tseng, Qian, et al. 2015) (W. Sun, Tseng, Zheng, et al. 2015). So that the sub breast regions with relatively homogenous gray level intensity were estimated before feature extraction. For each mammogram, we used region grow method to segment the dense region. Since the seed point was located by identifying the brightest point in each mammogram, the segmented region was the dense region of breast.

2.4 HAND-CRAFTED FEATURE EXTRACTION

2.4.1 Single view feature extraction

Computational features were used to describe the density and texture information of mammograms, and based on where these features extracted from, we grouped them into two major categories: single view features and dual view features. The single view features were computed in every single mammogram, and the dual view features were either fused from two single view features or extracted from a pair of CC and MLO view mammogram directly.

There are 466 single view features were extracted from both dense region and whole breast to characterize the heterogeneity/inhomogeneity density and texture patterns of the breast, and to estimate the cancer risk accordingly. These features can be grouped into six groups: the first group is density based features, the second and third groups are statistic texture features, the fourth and fifth groups are mathematical based texture features, and the sixth and last group of features are shape based texture features. The details of the features are listed below:

Single view feature group one (S1) includes 16 features, they are: mean, standard deviation, coarseness, homogeneity, inertia, energy, entropy, skewness, kurtosis, smoothness, mean gradient, the number of pixels with the maximum gray value in the histogram and the total number of pixels of the whole calculated region, the number of pixels with a gray value larger than the average value of the histogram and the total number of pixels of the breast, the average difference of two adjacent values in the histogram, the average value of the histogram; the standard deviation of the histogram, and the uniformity of intensity in the histogram. All these features are density based features, which closely relate to the mammographic density and its distribution of the breast image depicted on each mammogram.

The second group (S2) of features are gray-level run length statistics (RLS) texture features (Tang 1998). These RLS features consist of short run emphasis, long run emphasis, low gray-level run emphasis, high gray-level run emphasis, short run low gray-level emphasis, short run high gray-level emphasis, long run low gray-level emphasis, long run high gray-level emphasis, and run percentage. The long runs of one pixel values indicate coarser textures, and

the short runs indicate fine textures. To achieve the best performance of these features, all the original images were downgraded to 256 gray level images(Tan et al. 2013). Four run length matrices were computed along 0, 45, 90, and 135 degrees, and 36 features were calculated altogether.

In the third single view feature group (S3), we used statistical texture features that were extracted from the Gray Level Co-occurrence Matrices (GLCMs) (Haralick, Shanmugam, and Dinstein 1973). GLCM is a second order statistical measurement, and it measures the intensity of its neighbor at certain distance and orientation. The matrices were created by calculating how often a pixel with gray level value *i* occurs horizontally adjacent to a pixel with value *j*. Mean, variance, contrast, correlation, angular second moment, entropy and homogeneity features were then extracted from these matrices.

The fourth feature group (S4) contains steerable features (Freeman and Adelson 1991) (Greenspan et al. 1994). To construct the power map of steerable feature analysis, Gaussian pyramids need to be generated from each mammogram. Then by taking the difference between two consecutive levels Gaussian pyramids, a Laplacian pyramid can therefore be formed with each layer representing the information at different level of details (Barley and Town 2014). Then the power map can be generated by calculating the convolution with each oriented filter at different scales. In our study, we used four oriented filters, they are: $m_1(x,y)=e^{i(\pi/2)x}$, $m_2(x,y)=e^{i(\pi\sqrt{2}/4)(x+y)}$, $m_3(x,y)=e^{i(\pi/2)y}$, $m_4(x,y)=e^{i(\pi\sqrt{2}/4)(y-x)}$. The real and imaginary parts of the sinusoids are treated as two filters, and all the filters were used at three different levels. The mean and standard deviation were calculated for each map, which gives a 48-element features vector.

The fifth group of features was computed based on Gabor filters. A Gabor filter can be seen as a sinusoidal plane of a particular frequency and orientation with the formula $g(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi jWx\right]$, while j is $\sqrt{-1}$ and W is the frequency of the modulated sinusoid. Applying the filter to the image, the possibility wave is transformed

into a probability matrix, and the resulting coefficients represent the probability of the intensity of the wave (C. Wei, Li, and Li 2007). We tested 5 different frequencies at four directions, and for each Gabor image we computed the following features: mean, contrast, angular second moment, inverse difference moment, entropy, and correlation.

The features in the sixth group were extracted from directional texture images. In the preprocessing step, the multi-orientation signal decomposition will generate three images: density image, texture image and morphological image (W. Sun et al. 2014) (Wei Qian et al. 1994). From the texture image, the background detail signals, including weak random directional lines, were eliminated. Then, we used a ray-tracing algorithm to detect strong directional signals, and the directional features were calculated from these isolated lines as described in our previous publications (W. Sun, Tseng, Qian, et al. 2015). We designed and developed five features from the processed images: total number of directional tissues (strong isolated lines); principle orientations of directional tissues determined by gradient operators; average normalized directional tissue length; number of directional tissues with a length exceeding a given threshold; and the directional angular distribution that was computed as entropy.

All the features mentioned above were calculated on both whole breast area and dense area of the mammogram, in addition, two extra features were calculated on the dense region itself. The first feature is circularity of the dense region which describes the shape of dense region. The second feature is the center location differences of the dense region and the whole breast region, which describes the relative location of the dense region.

2.4.2 Dual view feature fusion and similarity features

To combine the information from both CC and MLO view mammograms, two different types of similarity measurements of dual view mammograms were investigated in this study. The first type of feature is called fused features, and they can adaptively fuse the CC and MLO view image features together. They are computed by fusing the corresponding CC view and MLO view features using the following equations:

$$DF1_i = \frac{\max(CC_i, MLO_i)}{(CC_i - MLO_i) \times \alpha + 1} \quad (1)$$

$$DF2_i = \min(CC_i, MLO_i) \times \left[(CC_i - MLO_i) \times \beta + 1 \right] \quad (2)$$

where CC_i and MLO_i represent the corresponding ith feature extracted from CC and MLO view, α and β are the tuning parameters and we can decide to what degree we want to compromise the difference of the two images. The reason we designed these two equations is based on the fact that different features have different sensitivity of the two view differences. For example, the maximum values of the dense region areas on CC view and MLO view are better at representing the true volume of the dense region, while the maximum value of average RLS on these two views cannot effectively represent the overall texture distortions. These two equations are relatively moderate in handling the differences of the two views compared to using the maximum and minimum values directly, since the added penalty coefficients can decrease the maximum value and increase minimum value when the difference is large.

The second type of feature is called similarity features, and they measure the overall similarity of each pair of mammogram. Instead of using the equations to combine the corresponding single view features, similarity features measure the difference from CC view and MLO view mammograms directly. These features quantify the texture and dense region differences. The details are listed below:

(1) Overall dual view feature differences:

First, we calculated the statistical distance between the CC and MLO view observations. For the CC and MLO view p-dimensional feature vectors $\vec{f}_{CC} = [f_{CC1}, f_{CC2}, ..., f_{CCp}]$ and $\vec{f}_{MLO} = [f_{MLO1}, f_{MLO2}, ..., f_{MLOp}]$, the statistical distance is $d(\vec{f}_{CC}, \vec{f}_{MLO}) = \sqrt{(\vec{f}_{CC} - \vec{f}_{MLO})' \vec{S}^{-1} (\vec{f}_{CC} - \vec{f}_{MLO})}$, where $\vec{S}$ is the covariance matrix. This measurement considers the variance from different sources, and is able to eliminate the variation effect.

(2) Orientation correlation differences:

Another measurement of similarity is the difference of the textural orientation in the two views. The Gabor filters were used to generate the Gabor image, and the correlation of the Gabor features was calculated. First, we calculated the correlations of corresponding Gabor features in CC and MLO views at every frequency and length. The formulation is defined as:

$$r_{orientation}(l,\theta) = \frac{\sum_{i\in CC, j\in MLO} (p_i - \overline{p_{CC}})(p_j - \overline{p_{MLO}})}{\sqrt{\sum_{i\in CC} (p_i - \overline{p_{CC}})^2 \sum_{j\in MLO} (p_j - \overline{p_{MLO}})^2}},$$

where p_i and p_j are the pixel values of CC and MLO view Gabor filtered image at length l and angle θ. Then we measured the overall correlation difference by using: $D_{orientation} = 24 / \sum_{l=1}^{4} \sum_{\theta=0}^{5\pi/6} |r(l,\theta)|$, the greater values represent larger orientation differences.

(3) Histogram difference:

Instead of comparing the difference of each pixel, we measured the difference of density distribution on histograms:

$D_{histagram} = \sum_{hist=1}^{256} |n_{CC}(hist) - n_{MLO}(hist)|$, where n_{CC} and n_{MLO} represent the number of pixels at gray value $hist$.

(4) Dense area difference:

The area of dense region is an important measurement for cancer risk, however, because of the overlapping of projections, the dense region shown in one view is different from the other. Capturing this difference can help us increase the confidence levels of risk predictions, and in this study we computed the difference of the dense region areas by using:

$D_{area} = |A_{CC} - A_{MLO}| / \max(A_{CC}, A_{MLO})$, where A_{CC} and A_{MLO} are the areas of dense regions of CC view and MLO view mammogram.

2.5 FEATURE ANALYSIS

2.5.1 Feature selection and classification

In this study, we introduced and adopted a computational method can integrate the feature selection and classification procedures at the same time, which shows advantages in handling high dimensional data. In most existing CAD systems, support vector machine (SVM) is considered as a popular and powerful classification method. However, since the standard SVM decision rule analyzes all the variables without discrimination, redundant variables may affect the classification results negatively (Hastie et al. 2005). This disadvantage can be minimized by adding a feature selection module before applying the data to SVM. However, most feature selectors may not achieve expected performance when the number of instances is smaller than or comparable to the number of image features. In our study, due to (1) the relatively higher correlation between single view features extracted from the MLO and CC view mammograms, and (2) the larger feature pool (set) size than our data sample size, the adding Elastic net (L1 + L1 norm) penalty to SVM is an optimal solution to overcome these restrictions. Thus, in this study, we used an elastic net SVM (EnSVM), so that the feature selection and classification can be achieved simultaneously with penalization method (Ma and Huang 2008) (Ye and Chen 2011).

In brief, given a dataset $\{(\vec{x}_i, y_i)_{i=1}^n\}$ with n cases, where $\vec{x}_i$ represents predictable variables and $y_i = \pm 1$ shows its corresponding label. SVM identifies a hyperplane that maximizes the margin of data with different labels by a linear boundary $f(x) = \sum_{j=1}^{p} w_j x_j + b$, where j is the number of features. The optimal pair $(\vec{w}, b)$ can be found by solving $\min_{w,b} \sum_{i=1}^{n} [1 - y_i f(x_i)]_+ + penalty_{\lambda}(w)$, where $i \in (1, n)$ (L. Wang, Zhu, and Zou 2006). Traditionally, there are two types of regularizations, L1-norm penalty and L2-norm penalty. L1 penalty allows the feature selection, and L2 penalty helps groups of correlated variables be selected together (Zou and Hastie 2005). But L2 penalty tends to give similar fitted coefficients to highly correlated features (i.e. grouping effect). An elastic net was proposed as a new

regularization and variable selection method which combines L1 and L2 penalties (Zou and Hastie 2005). The elastic net shows many advantages when handling grouping effects and the situation in which the feature size is greater than sample size. The optimization function thus becomes: $\min_{w,b} \sum_{i=1}^{n} [1 - y_i f(x_i)]_+ + \frac{\lambda_2}{2} \|\beta\|_2^2 + \lambda_1 \|\beta\|_1$, where λ_1 and λ_2 are tuning parameters for L1 penalty $\|b\|_1$ and L2 penalty $\|b\|_2^2$ (L. Wang, Zhu, and Zou 2006). An optimized r package penalizedSVM was used in this study (Becker et al. 2009).

2.5.2 Experiment design and evaluation

All the features from both CC view and MLO view mammograms were sent to our proposed similarity based dual view classification scheme which calculated the similarity scores of every pair of mammograms before further classification (Illustration 2.1). The correlation of feature vectors from two corresponding views varies from case to case: for some cases the CC view mammograms are visually similar to MLO view mammograms, while other cases are very different. However, the highly correlated features will interrupt the performance of classification results, while using the features only from single view will lose the additional information. To solve this problem, we calculated the similarity score for every pair of mammograms, which is the summation of the four normalized similarity features proposed in Section 2.4. According to the similarity scores, all the cases were stratified into high similarity group and low similarity group. For the low similarity cases (i.e., cases with low similarity scores), three EnSVM classifiers were used: two classifiers to compute CC view and MLO view risk scores, and another one for dual view risk scores that used fused features from two views. Then, we combined these three risk scores with a similarity score, and applied Artificial Neural Network (ANN) to generate the final risk score. For the high similarity cases, we predicted the cancer risk scores only using single view features. Since from our previous experiment results, we found that using CC view based image features outperformed using MLO view (W. Sun et al. 2014), we applied the features extracted from CC view mammograms to the classifier. For all the schemes

tested in this study, we added the following genomic biomarkers to the final classifiers: women's age, body mass index (BMI), family history of breast cancer. The flowchart is shown in Illustration 2.1.

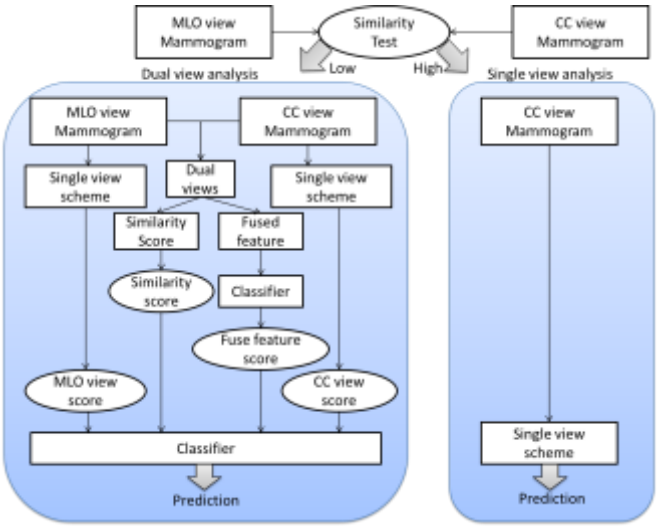

Illustration 2.1: Flowchart of our proposed similarity based dual view scheme.

To test the efficiency of our proposed algorithm, we compared it with several other schemes. The first two schemes are CC single view scheme (M1) and MLO single view scheme (M2), the third scheme combined and mixed features from two single views together (M3), the fourth scheme used the fusion function to combine the features from dual views (M4), and the last scheme is the similarity based dual view scheme we developed in this study (M5). Based on the follow-ups of the participants after the "prior" negative screening, we divided the women diagnosed as positive into three groups: diagnosed as cancer in less than 12 months after the negative screening; diagnosed as cancer in less than 12 months after the negative screening; detected cancer only after the "current" screening (screen detected cancer). We compared the risk score responses and prediction accuracies of each participating group.

For the evaluation purpose, a 10-fold cross-validation method was applied to train and test the classifier in this study. In each testing case, the final classifier generates a risk classification score ranging from 0 to 1. Then, we applied area under the receiver operating characteristic curve (AUC) to assess the discrimination and calibration performance of our new near-term

breast cancer risk prediction model. The AUCs were computed by ROCKIT program (ROCKIT, http://www-radiology.uchicago.edu/krl/). The positive and negative predictive value were also assessed and reported. To evaluate the performance of our scheme on different group of features, we tested on each of the three data groups. All the tests are two-sided, strong family wise error rates were controlled by Tukey multiple-comparisons test and the adjusted p value less than 0.05 were considered as statistically significant.

2.6 DEEP LEARNING METHOD

2.6.1 Deep learning for breast cancer

Breast cancer is one of leading death cause all over the world and predicting breast cancer risk has been studied more than 30 years (R. A. Smith, Duffy, and Tabar 2012). Breast density has been treated as one of the most important and effective image based breast cancer risk measurements, and higher breast density usually means higher breast cancer risk(Amir et al. 2010). However, the measurements of breast density are based on visual assessment of mammographic density by radiologists, and the density is rated and grouped into four American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) density categories (Nelson et al. 2009). But these subjective ratings are inaccurate or inconsistent due to the large intra- and inter-observer variability (Kopans 2008). From our previous research we also noticed that except density information, other image based features, like texture features, are also important to analyze breast cancer risk (W. Sun, Tseng, Qian, et al. 2015). So how to efficiently design the reliable features is a great challenge for all the breast cancer risk researchers (Wei Qian, Sun, and Zheng 2015)(W. Sun et al. 2014)(W. Sun, Tseng, Zheng, et al. 2015).

Recent three years, convolutional neural network (CNN) (Yann LeCun, Kavukcuoglu, and Farabet 2010)(Simard, Steinkraus, and Platt 2003) has been showing dramatic power in improving the performance of the state of the art computer vision tasks, including from segmentation to classification. Instead of designing and extracting the computational features from images, CNN can use the image itself as input and learn the features from the training

images itself. We observed this opportunity in improving the performance of near-term breast cancer risk prediction, but like most deep learning algorithms, CNN requires a large number of training data for tuning parameters (Bengio 2009). To collect data for breast cancer risk analysis research, we need the participants have at least two screening mammograms: the "prior" screening with negative diagnosis result, and "current" screening or medical report after 6 to 24 months after "prior" screening. In clinic, the majority women remain negative in the second screening, and to build a balanced dataset, we generally use same amount of negative and positive data. For these reasons, it is extremely hard to generate a large enough dataset for CNN. In this study, we developed a new scheme to use CNN on a limited data set for risk analysis, the details and results are listed below.

2.6.2 Deep learning scheme

From these 420 cases, we get 840 mammograms altogether. The first step is to remove the mammogram labels and background, then we used computer to find the largest containing rectangular in each breast image. The size of these rectangular varies according to the size and shape of the original breast image. From each of the rectangular, we selected 100 52 pixel by 52 pixel region of interests (ROIs). To make these ROIs as much evenly distributed as possible, we meshed each rectangular with a 10 by 10 grid, and the center of each grid is the center of a ROI. If the rectangular is big, there would be some gap among the ROIs; if the rectangular is small, there would be some overlaps. After this step, 840000 ROIs were extracted and used as input for next step. Figure 2.2 showed an example of the ROIs extracted from one mammogram.

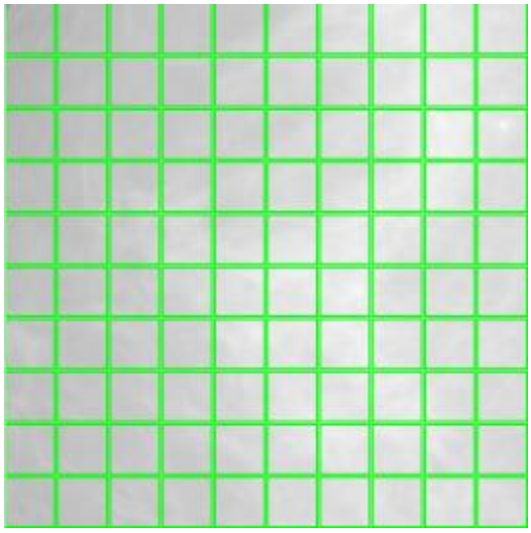

Figure 2.2: ROI preparation for deep learning on breast cancer risk analysis

Our deep neural network has eight convolutional neural network layers and one fully connected layer. The input image starts with the convolutional layer: a bank of 5 by 5 kernels is convolved with the original ROI in a moving window fashion. These kernels are randomly generated from uniform distribution with zero mean of unit norm. The second layer is the max-pooling layer, the bank of convolved images output the maximum activations in every non-overlapped square region. In convolution layers of our proposed scheme, every odd number layers are convolutional layers, and every even number layers are max-pooling layers. The first two convolutional layers used 5 by 5 kernels, and the last two layers used 3 by 3 kernels. Also, the first and last convolutional layers have six output maps, and the two middle layers have twelve output maps. The fully connected layer transforms the output of the previous layer into a feature vector, and they are fully connected with the two classes (low risk and high risk classes) with a softmax function. The convolutional architecture is shown in Illustration 2.2.

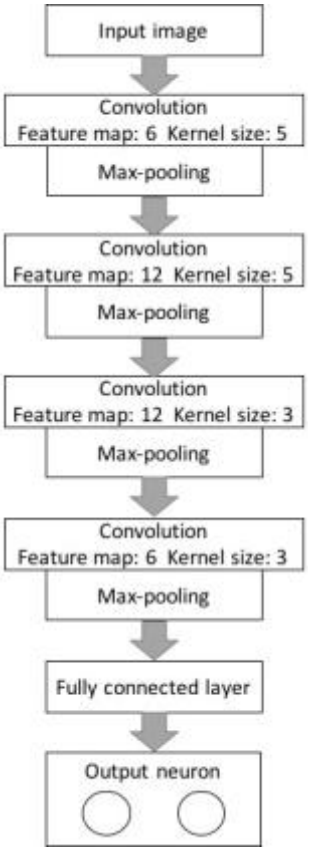

Illustration 2.2: The deep convolutional neural network for breast cancer risk analysis

To develop and analyze the best way to use CNN in predicting the near-term breast cancer, we conducted two experiments for comparison. In the first experiment, we used the largest rectangular extracted from the breast area as the input to CNN, and all the rectangular are downsampled to the size of 52 by 52. So the 840 samples will be the used for the training and testing. In the second experiment, we used all ROIs extracted from every rectangular to train and test the CNN. The total sample size is 84000. Since we divided one rectangular into many ROIs, we also tested and compared the CNN performance on different number of ROIs extracted from each rectangular at three different levels: 16, 64, 100. The 10-fold cross validation method is used for all experiments.

2.7 RESULTS

2.7.1 Results on ipsilateral view scheme

The region segmentation results showed that the dense regions depicted on the CC and MLO view mammograms are often not identical. As a result, many feature measurements are different in the CC and MLO view projections. To visualize the quantitative differences of these dense regions shown in CC and MLO views, we chose three typical features, size, variance of GLCM matrix, and the relative center differences of the dense regions, to represent the dual view dense region differences. The scatterplot of the three features of dense regions is shown in Figure 2.3.

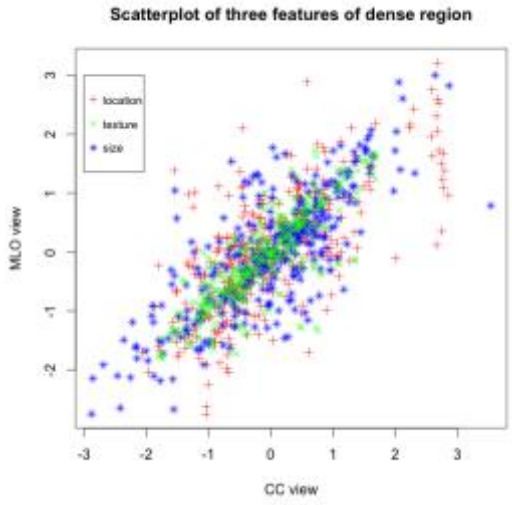

Figure 2.3: the scatterplot of size, texture characters and the relative location of the dense regions shown on CC and MLO view. (If no dense region was identified on the mammogram, the whole breast region will be used instead.) To display three measurements in one figure, we standardized the data and one unit indicates the one σ.

Another visualization of these similarity measurements is shown in Figure 2.4. Four groups of ROIs were shown in Figure 2.4, and each group is a typical example of mammogram images with remarkable differences on each of the four similarity features (i.e., overall dual view feature differences, orientation correlation differences, histogram differences, and dense area differences). In each group, there are two pairs of ROI images from CC and MLO views, and one

pair with high similarity and the other pair with low similarity. The correlation of our proposed four similarity features are shown in Table 2.1.

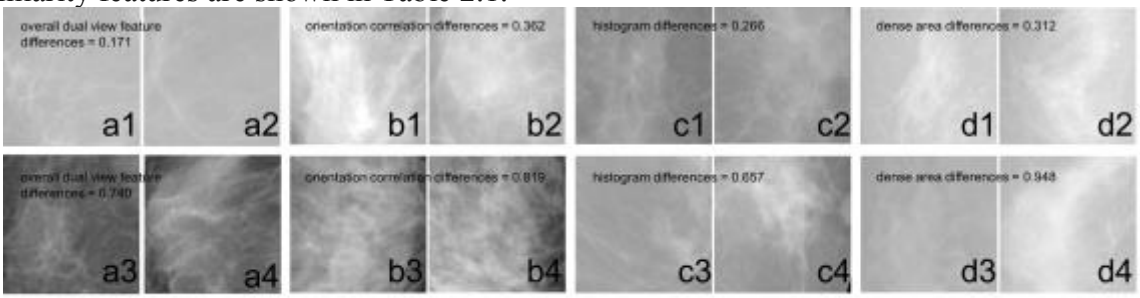

Figure 2.4: Four groups of ROI pairs with each of the four similarity features, group a to d are the examples of overall dual view feature differences, orientation correlation differences, histogram differences, and dense area differences, respectively. In each image group, the first pair of ROIs (sub image 1-2) were extracted from corresponding CC view and MLO view mammograms of the same breast, and they have high similarity feature response; the other pair of ROIs (sub image 3-4) were extracted from another pair of mammograms with low similarity feature response. For example, the overall dual view feature differences of a1 and a2 is relatively low, however, the a3 and a4 ROI pair is relatively higher.

Table 2.1: The correlation of each similarity feature

	Dual view feature differences	Orientation correlation differences	Histogram differences	Dense area differences
Overall dual view feature differences	1	0.73	0.62	0.54
Orientation correlation differences		1	0.61	0.47
Histogram differences			1	0.44
Dense area differences				1

Combining all the single view features from every feature group and the extra two shape and location based features together, we get 466 features for each mammogram. Since the number of features is bigger than the number of samples, many traditional feature selection methods will not work or will give poor results. In our proposed scheme, we used EnSVM for feature selection and classification. For comparison purpose, we also used random forest and Elastic Net as feature selection methods combined with ANN and SVM classifiers. Grid search method was used to optimize the parameters. The experiment results are shown in Table 2.2.

SFFS was used on each group individually, and the final selected features are the combination of selected features from each group

Table 2.2: Comparison of different feature selectors and classifiers using dual view similarity based scheme.

Feature selector	Classifier	AUC	Accuracy	# of selected features						
				group 1	group 2	group 3	group 4	group 5	group 6	total
EN	ANN	0.643±0.063	0.600	9	11	9	15	14	4	64
	SVM	0.661±0.056	0.638							
	EN	0.654±0.046	0.607							
RF	ANN	0.708±0.031	0.653	5	5	5	13	17	2	47
	SVM	0.711±0.039	0.656							
	EN	0.622±0.062	0.589							
SFFS*	ANN	0.683±0.050	0.635	3	2	3	2	3	1	14
	SVM	0.694±0.034	0.658							
	EN	0.619±0.069	0.571							
EnSVM		0.737±0.052	0.694	7	8	6	13	13	4	51

Table 2.3 compares similarity scores, density distributions and predicted risk scores of single view schemes using CC view and MLO view mammograms respectively. All the features and classifiers are the same as dual view scheme, but no similarity features and fused features were used. Based on the prediction results of CC and MLO view mammograms from each case, we got three different types of result: TP-TP, TP-FP, and FP-FP. From the table we can see the percentage of correctly predicted positive cases (TP-TP pairs) is 52.0% (105 out of 202) when analyzing the single view mammogram individually, which is significantly lower than 60.4% (122 out of 202) (Table 2.4) using our proposed dual view scheme. The prediction disagreements mainly happened in less dense cases, BIRADS 2 and 3, which have lower similarity scores.

Table 2.3: The result analysis on CC and MLO single view mammogram risk analysis schemes

# of cases in each BIRADS group					
	BIRADS 1	BIRADS 2	BIRADS 3	BIRADS 4	Total
TP-TP	2	34	65	4	105
TP-FP	0	12	7	1	20
FP-FP	1	16	18	2	37
Average similarity score					
	BIRADS 1	BIRADS 2	BIRADS 3	BIRADS 4	Total
TP-TP	0.81	0.86	0.90	0.94	0.89
TP-FP	N/A	0.62	0.67	0.91	0.65
FP-FP	0.89	0.85	0.88	0.94	0.87
Average predicted risk score difference					
	BIRADS 1	BIRADS 2	BIRADS 3	BIRADS 4	Total
TP-TP	0.14	0.09	0.08	0.05	0.08
TP-FP	N/A	0.32	0.26	0.11	0.29
FP-FP	0.07	0.13	0.11	0.09	0.12

Table 2.4 shows comparison of our proposed scheme with the other four different schemes. The parameters α and β are set to 0.2 and 0.6 using grid search method. The Tukey's pairwise comparison diagram was shown in Figure 2.5, and the confidence intervals were considered for the pairwise differences. The lines were drawn under a group of methods if no pair of treatments in that group is significantly different, and two lines were observed. No significant differences were observed among method 2, 3, 4, and method 1, 4, but method 5 is significantly different from all the other methods in terms of AUC.

Table 2.4: Prediction performance comparisons of different schemes.

Index	Method	TP	FP	TN	FN	PPV	NPV	AUC
M1	CC view	117	44	146	85	0.727	0.632	0.714
M2	MLO view	113	55	135	89	0.673	0.603	0.689
M3	CC+MLO	116	42	148	86	0.734	0.632	0.708
M4	CC+MLO+fusion	119	47	143	83	0.717	0.633	0.711
M5	Similarity based dual view	122	40	150	80	0.753	0.652	0.737

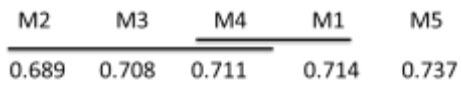

Figure 2.5: Tukey's pairwise comparison diagram

From Figure 2.6 we noticed that the average accuracy for women having interval cancer diagnosed in less than 12 months was the highest (75.7%), while the women with the screen detected cancer has the lowest accuracy (66.9%). The predicted risk scores have the similar trend as accuracies. All the adjusted p values (Tukey multiplicity correction) are less than 0.05.

Figure 2.7 shows and compares ROC curves using three different schemes, which were all generated by the ROCKIT program with all 392 cases. Two lower curves are generated by the schemes using either CC view or MLO view mammograms only, and the top curve was generated by the similarity based dual view mammogram scheme.

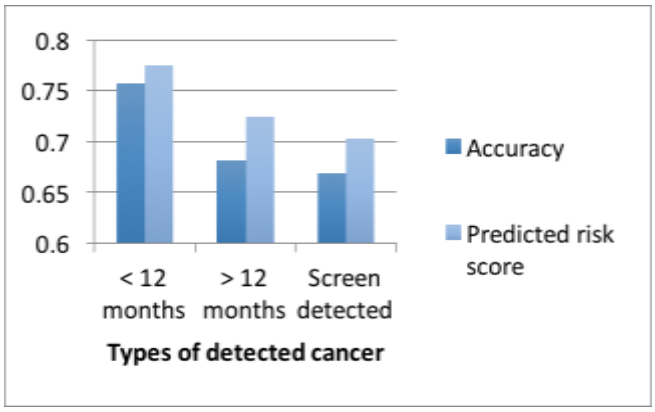

Figure 2.6: Accuracies and predicted risk scores of each participant cancer group: women who diagnosed with interval cancers with time <12 and >12 months after the negative prior screening, as well as the women who have screen-detected cancers.

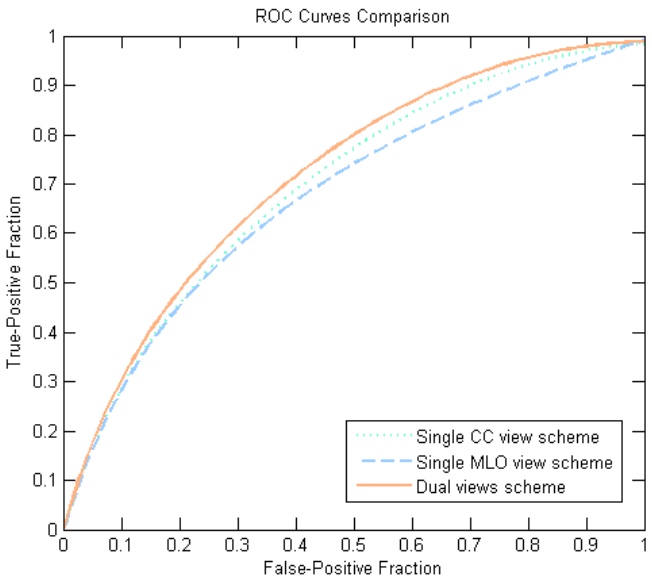

Figure 2.7: ROC plots of using CC view, MLO view, and similarity based dual view mammogram schemes.

2.7.2 Results on deep learning scheme

For the first experiment, the CNN didn't converge; for the second experiment, the averaged accuracy is 0.6707 for all the 84000 ROIs (ROI based accuracy) using 10-fold cross validation. Since every breast image has 100 ROIs, and not every ROI has the same predicted risk label, so we set a percentage threshold t. If more than t ROIs from one image were predicted as high risk, we regarded this case as high risk case; otherwise, it was treated as low risk case. Based on different threshold t, we plotted the ROC curve in Figure 2.8, and the threshold maximize the area of the containing rectangular under the curve is 0.5200. Using this threshold (t=0.5200), the case based accuracy is 0.6972 and the ROI based accuracy is 0.6707, while the area under the curve (AUC) is 0.7173 and 0.6982 respectively. We also compared the different number of ROIs at three different levels: 16, 64, and 100 ROIs for each breast image. We compared the case based accuracy and ROI based accuracy, and the results are shown in the Table 2.5.

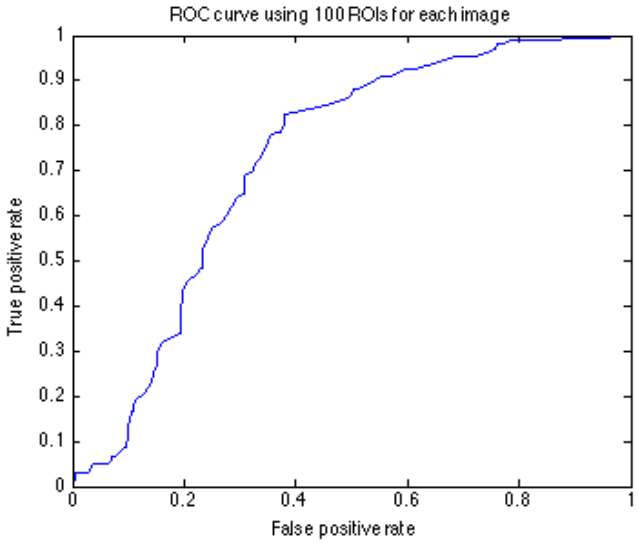

Figure 2.8 ROC curve using 100 ROIs for each breast image

Table 2.5 Prediction accuracies using different amount of ROIs

	ROI amount in each breast image		
	16 (4 by 4)	64 (8 by 8)	100 (10 by 10)
ROI based accuracy	0.6590	0.6664	0.6707
Case based accuracy	0.6134	0.6523	0.6972

Figure 2.9 shows the some of the examples of feature maps from our CNN structure. These feature maps were generated by computer itself, and it is more representative than human-designed features. Intuitively, they are the combination of density, texture and orientations.

Figure 2.9. Some examples of feature maps generated from our CNN algorithm

Chapter 3: Lung cancer diagnosis

3.1 RATIONALE

Healthcare industry is facing the opportunities and challenges of big data nowadays. Reports said that at this rate of growth, healthcare data in United States will soon reach the zettabyte (1021 gigabytes) scale (Cottle et al. 2013). If these big data can be effectively synthesized and analyzed, those relations, patterns and trends can be revealed, thus the doctors can provide more thorough and insightful diagnoses and treatments, and potentially resulting in higher quality care at lower costs (Raghupathi and Raghupathi 2014). With the development of precision medicine and radiomics, massive radiomics data collected from multiple modalities in a mineable form are gradually becoming available to build descriptive and predictive models (V. Kumar et al. 2012). Deep structured algorithms have the potential to generate valuable features and reveal the quantitative predictive or prognostic associations between raw data and medical outcomes.

In the last three decades, many researchers have been developing computer aided diagnosis (CADx) algorithms or systems optimized to enhance the performance of radiologists reading and interpreting medical images (Giger, Chan, and Boone 2008)(Wei Qian, Sun, and Zheng 2015). Most of the previous researches are based on manually designed computational features, and these extracted features are sent to linear classifiers to distinguish the benign cases and malignant cases. These features include texture features, density features, morphological features extracted from the original tissue image or region of interest (ROI). For example, area, circularity, ratio of semi-axis are very typical and useful morphological features (Wei Qian et al. 2007); average intensity, mean gradient of region boundary, density uniformity are very common density features used for mass detection (Wei Qian, Li, and Clarke 1999); wavelet features, gray-level co-occurrence matrix (GLCM) features, run length features, local binary pattern (LBP) features, SIFT features are powerful texture features we used for breast cancer risk analysis (W. Sun, Tseng, Zheng, et al. 2015). Feature design is considered as an essential module for most existing CADx, but it is a time consuming and complicated task (D. Kumar, Wong, and Clausi 2015)(Roth et al. 2015)(Way et al. 2009)(Way et al. 2006). Moreover, the combination of well-

designed features may not necessarily produce expected performance without considering the correlation and interaction among different features. Feature selection algorithms like genetic algorithm (GA), sequential forward floating selection (SFFS) can help generate the optimum feature combinations (W. Sun, Tseng, Qian, et al. 2015)(W. Sun et al. 2014), but they become less efficient when the dimension of features is high. In addition, the performance and reproducibility of CADx systems are always controversial topics even for the commercially available systems (Leader et al. 2005)(Zheng et al. 2003). Because different image databases were used to develop different schemes and the results depend heavily on the difficulty of the selected cases (Nishikawa et al. 1994), and current CADx schemes are sensitive to small variations in the digital value matrices that result from operators and machines (Leijenaar et al. 2013)(van Tulder and Bruijne 2016). Some research group already made some progress in using deep learning algorithms on lung cancer diagnosis: Wei Shen et al (Shen et al. 2015a) proposed multiscale convolutional neural network structure on lung cancer diagnosis using 3D data, and they integrated information of ROIs acquired at different scales; Kumar Devinder et al (D. Kumar, Wong, and Clausi 2015) used autoencoder to analyze the diagnostic data from LIDC database.

3.2 MATERIALS

All the data we used in this study is from Lung Image Database Consortium and Image Database Resource Initiative (LIDC/IDRI) public database, which consists of diagnostic and lung cancer screening thoracic CT scans with marked-up annotated lesions (Armato et al. 2011)(Clark et al. 2013). At the time point of this study, there are 1018 cases collected from seven academic centers and eight medical imaging companies in this database with the CT scans slice thickness varies from 1.25 mm to 3 mm and reconstruction interval are between 0.625 mm and 3 mm. The clinical thoracic CT scan images of each case are associated to an XML file with the record of two-phase image annotation process evaluated by four experienced thoracic radiologists. Each radiologist reviewed each CT scan independently and labeled the lesions to

one of the three categories: nodule larger than 3 mm, nodule less than 3 mm, and non-nodule larger than 3 mm. The final opinions were made in the unblinded-read phase based on four anonymized marks. The ratings of 5 malignancy levels were given from each of the four radiologists to all the nodules larger than 3 mm, and level 1 and 2 represent benign nodules and level 4 and 5 denote malignant nodules. Figure 3.1 shows an example of a nodule in original CT scan images and its boundaries marked by four radiologists.

From these 1018 cases, we eliminated cases with no larger than 3 mm nodules or only non-nodule lesions, incomplete cases, and cases with missing truth files. To avoid the partial volume effects caused by different CT scanning protocols across different vendors, bi-cubic interpolation method was used to normalize CT scan volumes resulting in isotropic resolution at all directions. Since the size and shape of segmented nodules in the top layer and bottom layer might dramatically different from the rest of the slices, we removed these two slices if the segmented nodule volume contains three or more slices before interpolation. Then the segmented nodule area in each slice was arranged into a 52 by 52 pixel rectangular according to the following rules: if the segmented area can be fitted into a 52 by 52 pixel rectangular, it was placed to the center of this rectangular; otherwise it was downsampled to the size of 52 by 52 pixels. Every the rectangular was rotated to four different directions (0, 90, 180, 270), and converted to four single vectors with each representing one orientation. All the values in the vectors were down sampled to 8 bits. From these 1018 cases we generated 134668 vectors were obtained and each vector has 2704 elements. The distribution of malignancy levels of the data is shown in Table 3.1. All the intermediate samples (level 3) were eliminated, and 41372 benign samples (level 1 and 2) and 47576 malignant samples (level 4 and 5) were remained and used for this study.

Table 3.1: The distribution of malignancy levels of our dataset

Malignacy level	1	2	3	4	5	Total
Amount	15448	25924	45720	20520	27056	134668

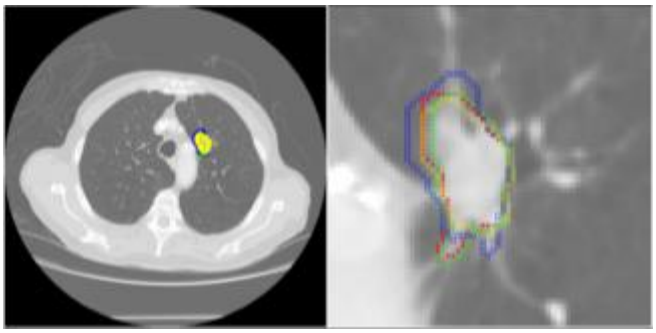

Figure 3.1: A nodule example in one slice of the original CT scan images with nodule's boundary marked by four radiologists (left) and the zoomed in image (right)

3.3 DEEP LEARNING METHODS

In this study, we designed and implemented three state-of-the-art deep structured schemes for nodule diagnosis: CNN, DBN, and SDAE. And for the comparison reason, we also extracted hand-crafted features from three different feature categories to classify the ROIs. In this section, we will introduce every scheme used in this study respectively.

3.3.1 Convolutional Neural Network

Convolutional neural network is the only deep learning algorithm we used without the need of unsupervised pre-training. Because the neural network with several full-connected layers is not practical for training when initialized randomly, the CNN structured we developed in this study is based on LeCun's model: a fully-connected layer followed by several convolutional layers and subsampling layers, and every layer has a topographic structure (Y. LeCun et al. 1989)(Yann LeCun et al. 1998). The algorithm begins by extracting random sub-patches from the ROIs mentioned above, with the size of the sub-patches referred to as the "receptive field size". In each layer the neuron is associated with a fixed two-dimensional position and its receptive field is corresponding to the input from previous layer (the first layer is corresponding to the original input image). Several neurons are connected to the same location at every layer, and each neuron is a linear combination of its corresponding neurons in previous layer with its set of input weights. The neurons at different locations are associated with the same set of weights, but different corresponding input patches (Bengio 2009).

The architecture of the proposed CNN is summarized in Illustration 3.1. It contains 8 layers: the first and last layers are input layer and output layer, and the 2, 4, 6 layers are convolutional layers, and the 3, 5, 7 layers are subsampling layers. In particular, the second layer has 12 feature maps connected to the input image through 12 5×5 kernels, followed by a 2×2 average pooling layer. The fourth layer has 8 feature maps, and they are all connected to previous layers through 96 5×5 kernels. After another average pooling layer, we obtained the sixth layer with 6 feature maps, and 48 5×5 kernels were used for this layer. In the eighth layer, the input shrunk to 3×3 matrices, and they were fully connected by using softmax non-linearity to the 2 output neurons associating with benign and malignant nodules. The output of each layer were contrast normalized and whitened before sending to the next layer (Hyvarinen and Oja 2000). The batch size was set to 100 and the learning rate was set to 0.1 for 100 epochs, the subsampling rate was constantly 2.

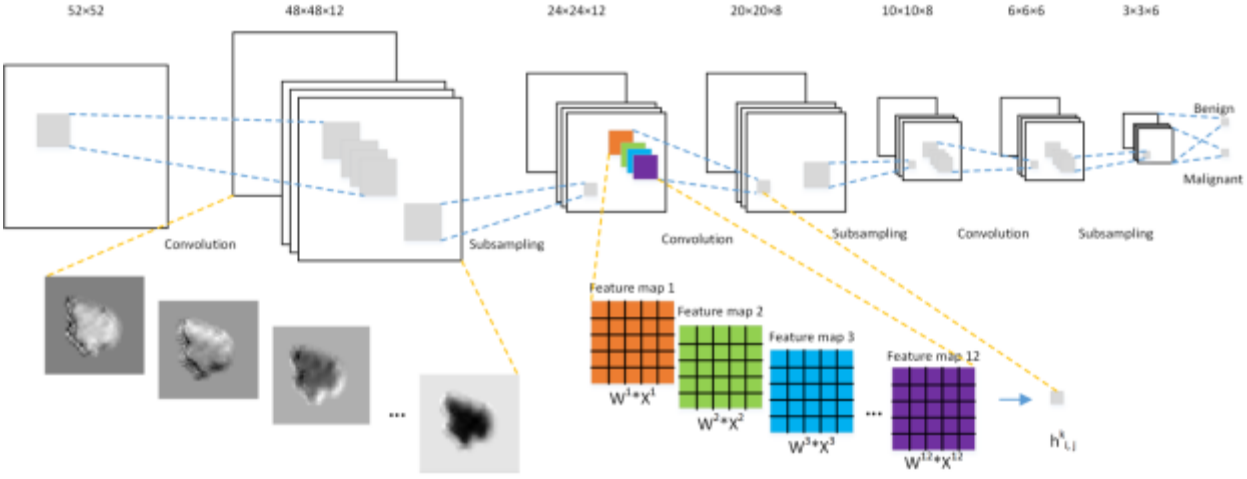

Illustration 3.1: The structure of CNN designed in this study. It demonstrates the original image, all the feature maps in convolutional layer, the process of subsampling layer. W^i is the weight matrix in each kernel, X^i is the pixels values of in a patch, and $h^k_{i,j}$ is one hidden unit in layer k at location (i, j).

3.3.2 Deep Belief Networks

Another deep learning algorithm we designed and tested in this study was DBN, it is a generative model that combined directed and undirected connections between variables. The model was obtained by training and stacking four layers of Restricted Boltzmann Machine (RBM) in a

greedy fashion. The RBM was used for the unsupervised learning as the start of the algorithm, and the meaningful computational features can be automatically extracted from the training process. The distribution of visible layer **x** can be calculated by computing the energy function of RBM:

$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{h}^{\mathrm{T}}\mathbf{W}\mathbf{x} - \mathbf{c}^{\mathrm{T}}\mathbf{x} - \mathbf{b}^{\mathrm{T}}\mathbf{h},$$

where **h** is the vector of hidden units, **b** and **c** are the bias vectors and **W** is the weight connecting two adjacent layers.

Each of the first three layers of our proposed scheme contains 400 units, and the top layer contains 1600 units. RBM doesn't allow the interactions either between the hidden units or between visible units with each other. The DBN we designed in this study followed the work of Hinton et al. (Hinton, Osindero, and Teh 2006), training greedily from lowest layer to the highest layer as an RBM, and the activations of previous layers were served as the input of next layer. The distribution of hidden unit h_j in the layer i follows the formula:

$$p\left(h_j^{(i)} = 1 \middle| \mathbf{h}^{(i+1)}\right) = sigm(\mathbf{b}^{(i)} + \mathbf{W}^{(i+1)^{\mathrm{T}}}\mathbf{h}^{(i+1)}),$$

the distribution visible unit x_j follows:

$$p\left(x_i = 1 \middle| \mathbf{h}^{(1)}\right) = sigm(\mathbf{b}^{(0)} + \mathbf{W}^{(1)^{\mathrm{T}}}\mathbf{h}^{(1)}).$$

When we reshape each weight vector into an image patch, each value was associated to the pixels at the same position of the input ROI. The positive and negative values of the pixel values at each position represent the increase or decrease the possibility of being 1 of the hidden units. This weight can be treated as the features extracted from the original ROI automatically by the computer. Once a layer is trained, all the parameters including the automatically learnt features were fixed until the whole training procedure for the multilayer DBN was finished. This greedy algorithm is shown to be optimizing the variational lower-bound on the data likelihood, if units in higher layers at least as have many units as each of the lower layers (Lee et al. 2011). The batch size was also set to 100, and the learning rate was set for 0.01 for 100 epochs. The brief idea of the DBN architecture used in this study is shown in Illustration 3.2.

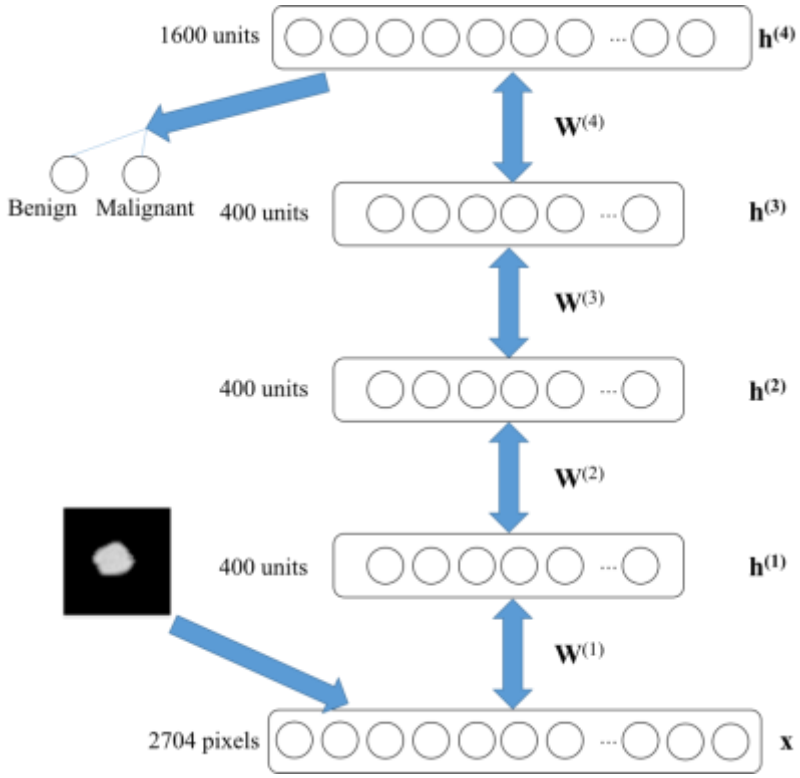

Illustration 3.2: The structure of DBN designed in this study. $\mathbf{h^{(i)}}$ is the vector of hidden units in hidden layer i, and $\mathbf{W^{(i)}}$ is the weight connecting two layers

3.3.3 Stacked Denoising Autoencoder

The last deep learning model we implemented and tested was SDAE (Bengio et al. 2007) and each autoencoder was stacked on the top of each other with the structure very similar to the DBN mentioned above. Autoencoder is another unsupervised algorithm that can automatically extract features from the data, and it is a type of feed forward neural network trained to reproduce the original input at the output layer instead of classifying them into different classes. The autoencoder consists encoder module and decoder module, and the encoder of layer *i* can be expressed as:

$$\mathbf{h}^{(i)} = \begin{cases} sigm(\mathbf{b}^{(i)} + \mathbf{W}^{(i)}\mathbf{h}^{(i-1)}), & i > 1 \\ sigm(\mathbf{b}^{(1)} + \mathbf{W}^{(1)}\tilde{\mathbf{x}}), & i = 1 \end{cases}$$

and the decoder of layer *i* can be expressed as:

$$\begin{cases} \mathbf{h}^{(i)} = sigm(\mathbf{c}^{(i)} + \mathbf{W}^{(n-i+1)^{\mathrm{T}}}\mathbf{h}^{(i-1)}), & i < n \\ \hat{\mathbf{x}} = sigm(\mathbf{c}^{(n)} + \mathbf{W}^{(1)^{\mathrm{T}}}\mathbf{h}^{(n-1)}), & i = n \end{cases}$$

where $\tilde{\mathbf{x}}$ is the original input vector with randomly added noise and $\hat{\mathbf{x}}$ is the output vector, **h** is the vector of hidden units, **W** is the weight between adjacent layers, **b** and **c** are the bias vectors (Vincent et al. 2008). All the parameters were optimized by minimizing the loss function below during the training process:

$$l = \frac{1}{2}\sum_{k=1}^{2704}(\hat{x}_k\text{-}x_k)^2$$

where $\hat{x}_k$ and x_k are the element in the noiseless input vector and output vector. For the discrimination purpose, the supervised classifier was added to the last layer of the encoder, and the whole model was trained as a feedforward-backpropagate neural network (Palm 2012).

There were 2000, 1000, and 400 hidden neurons in each autoencoder with corruption level of 0.5. Illustration 3.3 shows the structure of the proposed SDAE. The size of batches was set to 100, and the learning rate was set to 1 for all the 100 epochs. After the unsupervised autoencoder being well-trained, all the parameters were frozen and the weights were used to initiate the supervised neural network for classification.

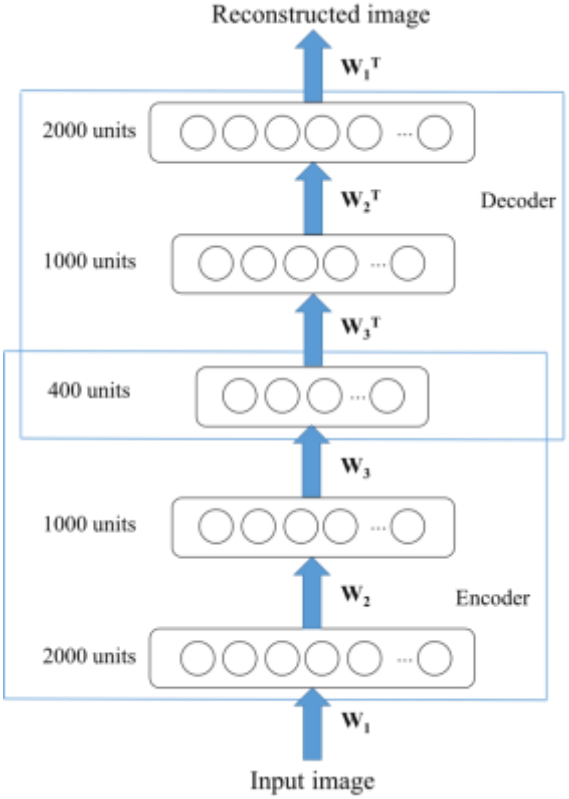

Illustration 3.3: The structure of designed SDAE. $\mathbf{W_i}$ is the weight matrix for layer i in encoder; $\mathbf{W_i^T}$ is the weight for decoder and it is the transpose of $\mathbf{W_i}$.

3.4 TRADITIONAL CAD SCHEME

For the comparison purpose, we extracted the traditional hand-crafted features from the same ROIs. From our previous experience of developing CADx systems, there are three major categories of computational features: morphological features, density features, and texture features. In this study, we implemented and tested several typical features in each category, and all these features were used in our previous researches and proved to be effective. 35 features were extracted from three categories, and Table 1 listed the descriptions of all these features. Then the kernel based support vector machine (SVM) was used to train the classifier using features from the same categories. We also combined all the features together, and trained the SVM classifier again using the features selected by SFFS.

Table 3.2: Descriptions of the traditional hand-crafted features used in this study

Category		Feature descriptions
Morphological		1) Area, 2) circularity, 3) ratio of semi-axis
Density		4) Average intensity, 5) standard deviation, 6) entropy
Texture	GLCM	7) Mean, 8) variance, 9) correlation, 10) uniformity, 11) inertia, 12) inverse difference, 13) contrast, 14) sum entropy, 15) homogeneity, 16) angular second moment
	Wavelet	17-20) Mean and 21-24) variance of HH, HL, LH, LL
	Run length	25-26) Long run emphasis, 27-28) short run emphasis, 29-30) low gray-level run emphasis, 31-32) high gray-level run emphasis, 33-34) run percentage calculated at 0 and 90 degree

To compare the performance of the deep learning features and traditional hand-crafted features, we tested all the algorithms based on the same ROIs extracted by the method mentioned above. There are some ROIs extracted from the same case, in order to completely separate the training data and testing data, we applied 10-fold cross-validation method for all the cases, and all the ROIs associated to the cases from the training folds were used for training, while the ROIs corresponding to the cases from the testing fold were used for testing. All the tests were two-sided and p value <0.05 were considered as statistically significant.

3.5 RESULTS

The three deep learning algorithms have many parameters including the filter size, learning rate, corruption level, etc. In order to find the best parameters for each algorithm, we tested all the combinations of the candidate parameters recommended by existing publications and our previous experience. The candidate parameters of each algorithm are listed in Table 3.3, and the parameter combinations used in the schemes achieved the top 3 performance for each algorithm are shown in Table 3.4.

Table 3.3: Tested combinations of parameters for CNN, DBN, and SDAE

CNN	Parameters	# of layers	# of Kernels in one layer	Alpha	Kernel size
	Canditate values	6, 8, 10, 12	48, 32, 24, 16, 12, 8, 6	0.01, 0.1, 1	3, 5
DBN	Parameters	# of layers	# of units in one layer	Learning rate	
	Canditate values	2, 3, 4, 5	100, 400, 700, 1000, 1300, 1600, 2000	0.01, 0.1, 1	
SDAE	Parameters	# of layers	# of units in one layer	Learning rate	Corruption level
	Canditate values	2, 3, 4, 5	36, 64, 256, 625, 1225, 2500, 3600, 4900	0.01, 0.1, 1	0.25, 0.5, 0.75

Table 3.4: The parameters used in the top 3 performance architectures of each deep learning algorithm and their performance based on 10-fold cross-validation

CNN				
# of layers	Architecture	Alpha	Kernel size	Accuracy
8	12, 8, 6	0.1	5, 5, 5	0.8175
8	12, 8, 6	0.1	5, 5, 3	0.8163
10	24, 16, 12, 8	0.1	5, 5, 5	0.8143
DBN				
# of layers	# of units in one layer	Learning rate		Accuracy
4	400, 400, 400, 1600	0.01		0.8220
3	400, 400, 400	0.01		0.8196
3	700, 700, 700	0.01		0.8188
SDAE				
# of layers	# of units in one layer	Learning rate	Corruption level	Accuracy
3	2500, 1125, 256	1	0.5	0.8001
3	3600, 1125, 256	1	0.5	0.7986
3	1125, 1125, 1125	0.1	0.25	0.7969

Figure 3.2 shows the kernels in the second layer and fourth layer of our CNN algorithm. The kernels in both layers cannot be directly analyzed with respect to what detectors they are as we expected (like curvy stroke detectors in hand written digit image recognition), and most of them are less well-defined detectors. However, there does seem to be some structure in the

kernels reflecting that the feature maps are still resembling the part of nodule corners and emphasizing structures of the nodules. Figure 3.3a shows the feature maps from the second layer: Figure 3.3a shows all the feature maps corresponding to one kernel in this layer, and Figure 3.3b shows the feature maps of one patch corresponding to all the 12 kernels in layer 2. We can easily see the contours of different nodules (Figure 3a), and each patch was decomposed into 12 patches with each one representing different texture of the nodule.

Figure 3.2: The visualization of the kernels of the second layer (a) and fourth layer (b) in the CNN.

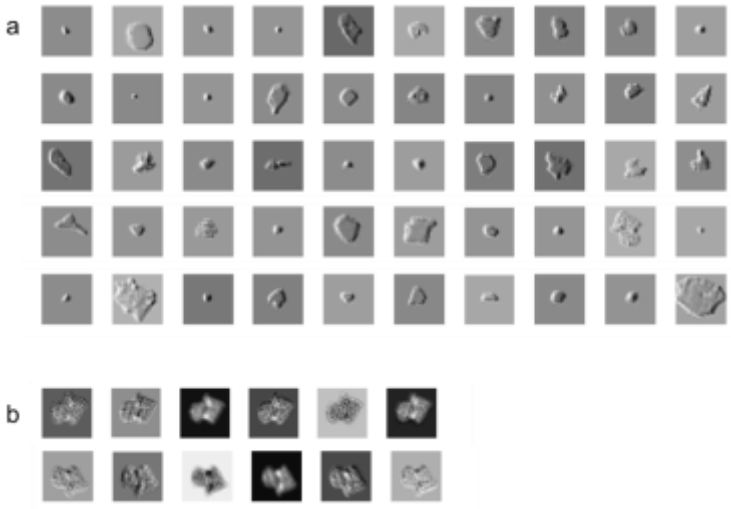

Figure 3.3: Some examples of the feature maps in the second layer of CNN. (a) Different patches in one feature map in layer 2. (b) One patch in 12 different feature maps in layer 2.

Figure 3.4 shows the visualization of the weights of the first layer in DBN. Each square represents the weight between one hidden unit and the pixel of the original image at the same position. The grey pixel denotes 0 in the weight matrix; white pixels represent positive values and increase the possibility of the hidden value being 1; black pixels represent negative values and decrease the possibility. These weights can be regarded as the automatically extracted features, and they have much more meaningful curvy stroke detectors compared to the kernels in CNN (Figure 3.2). Also, we can see the whiter pixels and darker pixels tend to be close together in spatially localized region. The extracted features from SDAE are shown in Figure 3.5, and you can see some meaningful detectors, but they are not as obvious as DBN.

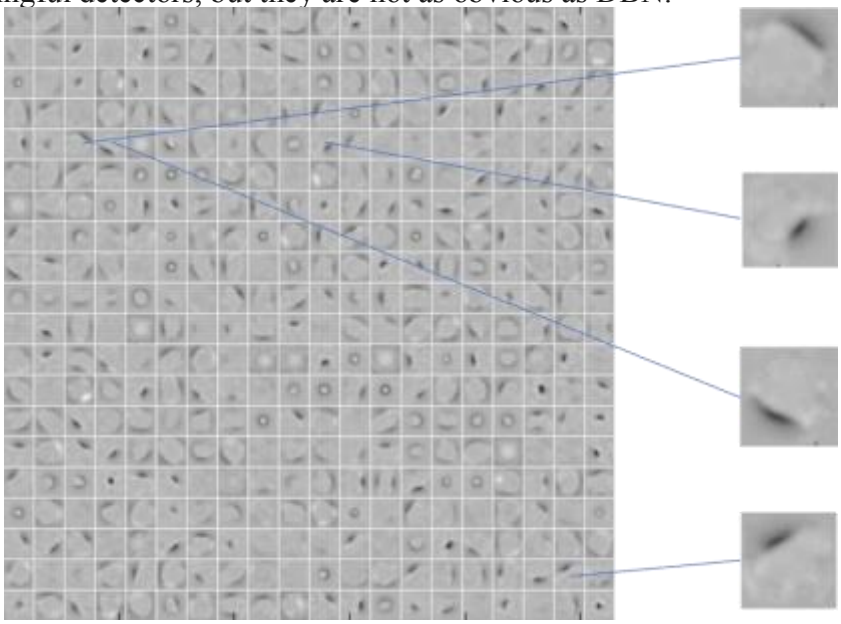

Figure 3.4: The visualization of 400 weights in the first layer of DBN. The amplified images on the right side are some example of weights representing curvy stroke detectors.

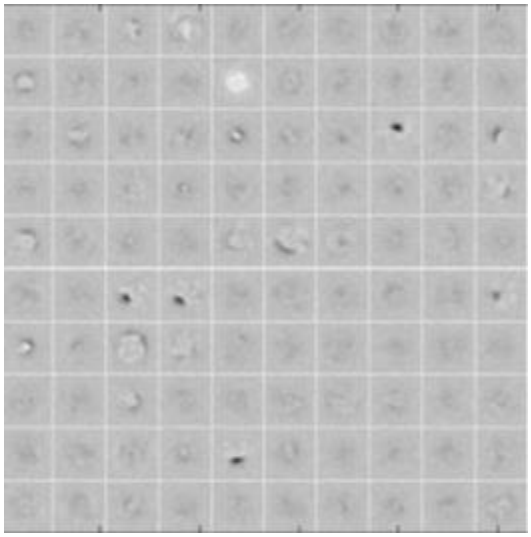

Figure 3.5: The visualization of 100 random weights in the first layer of SDAE.

To compare the performance of automatically generated features and traditional hand-crafted features we tested the seven different schemes using the same segmented ROIs: three deep learning algorithms, three categories of traditional hand-crafted features with SVM (i.e., density features, morphological features, and texture features), and the combination of all hand-crafted features with SVM. Table 3.5 shows the performances of these seven schemes. Since the merge of three different categories result in 34 features, SFFS was applied for feature selection and the final selected features are feature 1, 3, 8, 9, 18, 23, 26, 29. From the table, we can see DBN has the highest accuracy (0.8220) and AUC (0.8651±0.2026) among all these seven schemes, SDAE has the highest sensitivity (0.8354) and traditional CADx has the highest specificity (0.8743) which is slightly higher than DBN (0.8696).

Table 3.5: The comparison of accuracies and AUCs generated by the seven different schemes, and the highest value for each measurement is highlighted **bold**

	Sensitivity	Specificity	Accuracy	AUC	# of used features
CNN	0.7724	0.8376	0.8175	0.8483±0.1674	96
DBN	0.7186	0.8696	**0.8220**	**0.8651±0.2026**	400
SDAE	**0.8354**	0.7913	0.8001	0.8413±0.2941	**2000**
Density features with SVM	0.6042	0.6731	0.6676	0.7152±0.3672	3
Texture features with SVM	0.6411	0.7983	0.7409	0.7914±0.2419	3
Morphological features with SVM	0.6796	0.8463	0.7814	0.8242±0.1952	28
Combined features with SVM	0.6935	**0.8743**	0.8050	0.8443±0.2337	8

We also computed the p values of the classifiers for CNN, DBN, SDAE and traditional CADx system (the top four best schemes in Table 3.5), and they were all less than 0.0001 which were all statistically significant. Even though we used the term features for the comparison in Table 3.5, please be noted the meaning of the features in traditional CAD and deep learning schemes are different. Then we compared the predicted nodule malignancy scores of case generated from each one of these four schemes. After the normalization of these scores of each scheme and we performed F test on the scores from each pair of these schemes (Table 3.6). From the results, it can be noted that the predicted score is significantly different from all the other schemes, and the predicted scores from CNN and SDAE were not statistically different with each other at the level of critical value = 0.05. We also calculated the Akaike Information Criterion (AIC) (Akaike 1974) and Bayesian Information Criterion (BIC) (Schwarz 1978) for these four schemes (Table 3.7) and their ROC curves are shown in Figure 3.6. The results showed DBN has the best AIC and BIC.

Table 3.6: P values of F test on the predicted nodule malignancy scores using different schemes (Null hypothesis: true ratio of variances is equal to 1)

Scheme	Traditional CADx	CNN	DBN
CNN	0.02058		
DBN	2.2E-16	1.68E-12	
SDAE	0.04866	0.7306	1.36E-13

Table 3.7: AIC and BIC of four different schemes

	Traditional CADx	CNN	DBN	SDAE
AIC	1460.058	1475.598	1350.076	1486.608
BIC	1470.834	1486.373	1360.852	1497.383

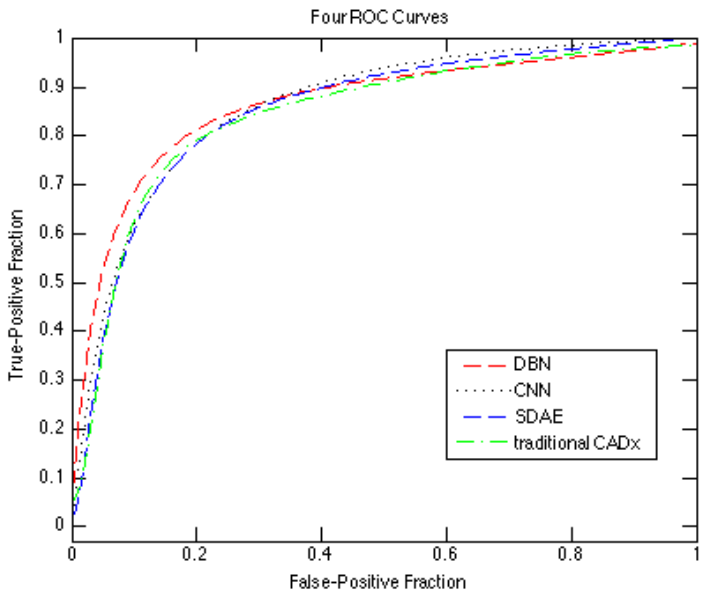

Figure 3.6: The ROC curves of CNN, DBN, SDAE and traditional CADx.

To compare the data differences mislabeled by DBN (the best performed deep learning algorithm in this study) and currently popular CADx, we calculated all the 34 hand-crafted features of all the nodules mislabeled by traditional CADx (group a) and DBN (group b), and compared their average values for each feature in both groups. No significant differences were found for all features extracted from the nodules in both groups (at critically value = 0.05). The largest difference is the standard deviation (feature 5) with average values from nodules in group a 2.35% larger than group b. Figure 10 shows some nodule examples mislabeled by DBN but correctly labeled by traditional CADx (Figure 3.7a) and nodules mislabeled by traditional CADx but correctly labeled by DBN (Figure 3.7b).

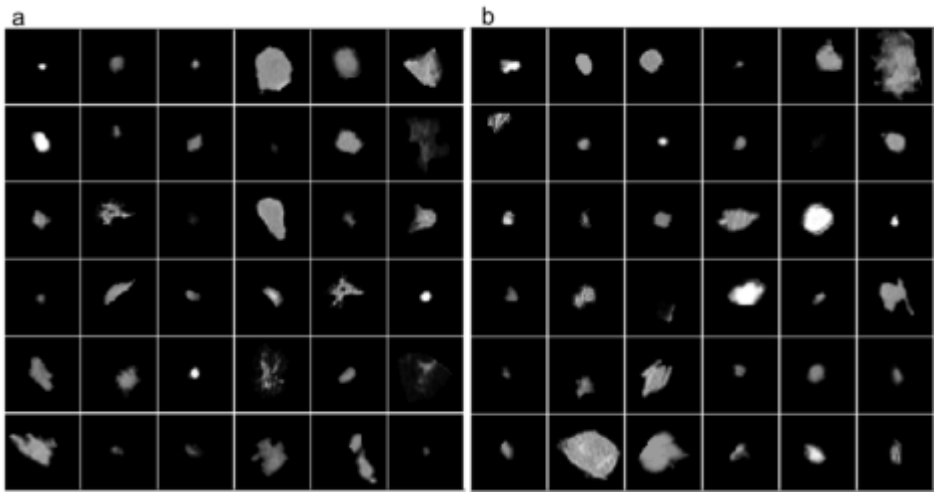

Figure 3.7: Some example nodules a) mislabeled by DBN but correctly labeled by traditional CADx; b) mislabeled by traditional CADx but correctly labeled by DBN.

Chapter 4: Conclusion

4.1 BREAST CANCER RISK ANALYSIS

Developing computerized scheme that can generate clinically acceptable results to stratify women into near-term high and low cancer risk group is of significant importance to improve efficacy of current population-based uniform mammogram screening paradigm. As a result, only a small fraction of women have to be screened more aggressively, while the majority can be recommended to get the screening in longer intervals until their cancer risk flag being changed in future assessment. Since in clinical practice, the radiologists qualitatively assess the CC view and MLO view mammograms of each breast to give more accurate evaluation, combining the information from two views in computerized scheme should increase the risk analysis performance. Based on this assumption, we developed and investigated a novel scheme to combine the information from MLO view and CC view mammograms for near-term cancer risk prediction. To the best of our knowledge, this is the first publication on breast cancer risk scheme using dual view ipsilateral mammograms. Compared to conventional risk models aiming to predict long-term or life time cancer risk, our model is focused on stratifying women into high risk and low risk groups of developing breast cancer in near-term (i.e., the next sequential mammography screening examination). The results showed our scheme can generate a higher AUC and lower FP rate compared to the single view scheme. The results also demonstrated that our scheme has a significantly better performance in predicting the development of breast cancer within 12 months after their "prior" negative screening than after 12 months.

To have a better understanding of utilizing the ipsilateral view mammograms, effectively combining the features from the different views is the most challenging step. In our study, we analyzed CC and MLO view cancer risk prediction individually, and we found that the prediction disagreements are more likely happened in less dense cases (BIRADS 2 and 3) and these case tend to have lower similarity scores. Due to this discovery, we designed the similarity score based dual view risk analysis scheme. Because the variations and consistency of CC view and

MLO view mammograms may vary dramatically from case to case, the correlations of the features from two views can also be very different. This brought a big challenge to the classifiers, for most of the supervised classification methods may not work well on highly correlated data. Until today, no ipsilateral view schemes have been reported for breast cancer risk analysis, only some literatures for dual view breast cancer detection CAD systems. In a recent paper, Maurice *et al*. (Samulski and Karssemeijer 2011) computed absolute differences of mass likelihood of each view, and used this as a similarity feature to detect the breast cancer. In the mass detection paper of Jun *et al*.(J. Wei et al. 2009), for each object, they used two scores from single view system (one score for each view) and one score from fused dual view system, and the final classifier was trained by these three scores. In another similar study by Xuejun *et al*.(X. Sun, Qian, and Song 2004), they combined the single view features and concurrent features together to train the classifier. In this study, we proposed a novel similarity based dual view system which considered the similarity of CC and MLO view mammograms. For the high similarity cases, the single view system will be used for risk prediction; for the low similarity cases, we incorporated two single view score, one fusion score and one similarity score for the final classification. From the results, we observed a significant AUC improvement and FP reduction using our similarity based dual view system compared to the single view system. The proposed system also outperformed the simple dual view system that did not consider the similarity information. The similarity measurement filtered the highly correlated dual view cases, so that the overall correlation of the two view features remained relatively low. For the cases with high similarity, two views will not give significant extra information compared to one view, but the cost of feature correlation increase cannot be overlooked.

In order to capture the characteristics of each pair of mammograms, identifying the effective computerized image features is important. Three different types of features were investigated in this study: single view features, similarity features and fused features. First, for single view features, we tested six different groups of features, including texture features and density features. All these features were extracted from both whole breast areas and dense

regions. Second, we proposed four similarity features to measure the degree of similarity each pair of mammograms. So the high similarity case will be sent to the single scheme and the low similarity case will be sent to high similarity scheme. Third, we developed an ipsilateral feature fusion formula to combine the information from two views. In addition to this formula, we also tested the other three fusion formulas: $\max(CC, MLO)$, $\min(CC, MLO)$, $\frac{1}{2}(CC + MLO)$. And we found the reported two equations are the best combination to integrate the two view features and then build the dual view classifier.

Compared to our previous studies, we combined dual view mammogram information to predict cancer risk for the first time. From our previous research results (W. Sun, Tseng, Qian, et al. 2015), we think certain amount of computational features is essential to make a better risk analysis. In this study, the dual view mammograms generated more features than single view. Several methods for combining the features from dual view mammograms were tested, and the results show combining dual view features with similarity information outperforms other strategies. Since the selected method generates a large number of features, choosing a suitable feature selector and classifier is important. In our scheme, we used EnSVM method which serves as a feature selector and a classifier at the same time. This method is good at dealing with large number of features and minimizes the influence of group effect. With this property, it is possible for us to incorporate and test features extracted from different views of mammograms and different sub-regions on each of them.

Despite the encouraging results, this study still has some limitations. First of all, our proposed scheme was tested in a limited dataset, so the efficiency and robustness of our algorithm on the general screening population need to be investigated in further studies. Second, more methods for efficiently combining the dual view features can be developed and compared. Third, we used some genomic biomarkers as the features in our system, the best combination of these features sand how to effectively combine them with image based features need to be investigated in future. The last but not the least, the situations where the number of features

exceeds the number of total sample size still remains a challenging and hot topic in machine learning and statistics areas. Many powerful classifiers are designed and published every year, and we will integrate the new classifiers to our system in future.

From the deep learning results we also found the potential of deep learning in breast cancer risk analysis. Our proposed scheme also achieved the desirable results. One of the biggest obstacle of using deep learning algorithm in breast cancer risk analysis is to identify the appropriate ROI input. Because the cases used in risk analysis haven't formed the obvious mass, so we cannot locate the suspicious areas like mass detection. Our proposed scheme provided a possible solution to generate the ROIs. Another alternative solutions will be studied in future.

4.2. LUNG CANCER DIAGNOSIS

This study supported deep learning algorithms with automatically generated features have comparable discriminative power to the currently popular CADx systems with traditional hand-crafted features in complicate medical image analysis like lung nodule CT image diagnosis, and the well-tuned deep learning algorithms have supreme performance than traditional CADx in terms of AUC and accuracy (Table 5). The results showed there is no observed bias in misclassified data from deep learning algorithms compared to traditional CADx, a deep learning based computerized lung nodule diagnosis scheme was able to differentiate the malignant and benign lung nodules with relatively higher AUC of 0.8651±0.2026.

To the best of our knowledge, this study is the first reported study on deep learning algorithms for lung cancer nodule image diagnosis, and systematically compared the performance of deep structured algorithms and currently popular CADx systems. A good feature should be invariant and selective at the same time. In the current popular CADx systems, all the computational features are manually designed. Designing and choosing features are difficult and time-consuming, the selected combination of certain features cannot guarantee satisfying results if not considering the correlations between each feature. Even the finalized features is able to produce good performance on the validation dataset, the the performance is still uncertain when

shifting to another dataset because the current CADx systems are very sensitive to small variations, which made the validity and reproducibility of the CADx systems a controversial topic (Leader et al. 2005)(Zheng et al. 2003). In the era of big data and radiomics, the large amount of accessible data can change the dataset dramatically. Deep learning algorithms with their ability of handling large scale of data and automatically generating computational features might have the potential in maintaining a sustainable and reliable performance on the ever-changing dataset. In this paper, we visualized the automatically generated features from CNN, DBN and SDAE, and compared their performance with traditional density features, morphological features and texture features. All these three deep learning algorithms generated meaningful features at different levels, but compared to the automatically generated features in simple image recognition tasks (Schmidhuber 2015), these features are less visually meaningful. One possible explanation is lung nodule diagnosis task is much more complicated than handwritten digit recognition task, even the radiologists in clinics cannot reach an agreement in a lot of times, so it will be harder to automatically extract visually meaningful detectors. The classification results showed all the three deep learning algorithms generated better performance than the scheme using each of the single categories of traditional features, and comparable results of the scheme using features combined from all the three categories. DBN achieved the highest AUC and lowest AIC, BIC among all the tested schemes (Table 5, 7), and the results are significantly different from CNN, SDAE and traditional CADx system. The AUCs, AICs, and BICs of the other three schemes are very close to each other. It was noted that DBN has the most meaningful automatically generated features compared to the CNN and SDAE, and it has also achieved the best performance.

Deep learning algorithms require input data has the same size, and this study provided a possible procedure of preprocessing the medical data so that the deep learning algorithms can be applied to these processed data. The original segmented nodules are of different sizes, and the length of a large nodule might span more than 200 pixels. If we use the rectangular contains the largest nodule as the input for the training and testing, it will end up more than 40000 pixels in

one sample, which makes the dimension of the data even larger than the number of entries. This situation should be generally avoided because the classifiers can hardly generate satisfying and reliable results because of the "curse of dimensionality" (Bellman 1957). On the other hand, if we downsample all the segmented nodules into a small matrix, much useful information will be lost. Considering the nodule size is an important feature (the single most important one (Wiemker et al. 2009)), we downsampled the large nodules and kept the original size of the smaller nodules so that every nodule image was fitted into a rectangular of same size. In this way, we preserved the size information for the deep learning algorithms, and the similar procedure can also be used for other deep learning based medical image analysis.

We also compared the mislabeled data from deep learning algorithm and traditional CADx. From the three tested deep learning algorithms, we chose the DBN, the one with the best performance, for the comparison. All the data mislabeled by DBN and traditional CADx systems were selected and we measured these data differences by computing all the 34 computational features discussed above. There is no statistically significant difference of every calculated feature between these two groups. This demonstrated DBN doesn't have any bias on the mislabeled data compared to the CADx system, which supported our hypothesis that automatically generated features by deep learning algorithms have great potential in medical image analysis.

Although the results are encouraging, we recognized this was a preliminary study. Firstly, we only tested limited number of layers in our implemented deep learning algorithms, but some deep learning algorithms applied in other areas have very complicated hierarchical structures and very deep layers (Szegedy et al. 2015). The depth of this structure is like the complicity of human's brain, so increase the number of layers might help improving the performance of the diagnosis. We cannot draw the conclusion that DBN is the better than CNN and SDAE for lung nodule diagnosis, because the deeper structured CNN or SDAE might have significantly improved results. Secondly, the optimum size of input for the deep algorithm has not been investigated. In this study, we chose 52 by 52 pixel input size because of the consideration of the

balance with the size of available dataset, computational efficiency, and the restrictions of CNN (the iterative of pooling and convolution in CNN will change the size of the image). Third, even though we used the largest available public lung cancer dataset, the size of the dataset is still a limitation for deep learning algorithms, because the numerous parameters inside the algorithms require a large dataset for good training results. With larger dataset, we can also test larger input ROIs with deeper layers structured algorithms.

In summary, we investigated and implemented three deep learning algorithms using LIDC lung cancer database, and the results showed the deep structured algorithms with automatically generated features can compete with traditional CADx systems using manually designed features. The potential of deep learning algorithms has been demonstrated in this study even using the structures with limited depth of layers. The performance of deep learning algorithms needs to be further tested using deeper layered algorithms and larger dataset in the future studies. If succeeds, the deep learning algorithms can significantly improve the accuracies of computerized lung cancer analysis systems, and we believe it holds promise as a scalable algorithm for learning hierarchical representations from other high-dimensional, complex medical image data.

References

Akaike, H. 1974. “A New Look at the Statistical Model Identification.” *IEEE Transactionson Automatic Control* 19 (6): 716–23. doi:10.1109/TAC.1974.1100705.

Amir, Eitan, Orit C. Freedman, Bostjan Seruga, and D. Gareth Evans. 2010. “Assessing Women at High Risk of Breast Cancer: A Review of Risk Assessment Models.” *Journal of the National Cancer Institute* 102 (10): 680–91. doi:10.1093/jnci/djq088.

Armato, Samuel G, Geoffrey McLennan, Drive Hawkins, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, et al. 2011. “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans.” *Medical Physics* 38 (2): 915–31. doi:10.1118/1.3528204.

Barley, Alexander, and Christopher Town. 2014. “Combinations of Feature Descriptors for Texture Image Classification.” *Journal of Data Analysis and Information Processing* 2 (3): 67–76.

Becker, Natalia, Wiebke Werft, Grischa Toedt, Peter Lichter, and Axel Benner. 2009. “PenalizedSVM: A R-Package for Feature Selection SVM Classification.” *Bioinformatics* 25 (13): 1711–12. doi:10.1093/bioinformatics/btp286.

Bellman, R. 1957. *Dynamic Programming*. *Princeton University Press Princeton New Jersey*. Vol. 70. doi:10.1108/eb059970.

Bengio, Yoshua. 2009. “Learning Deep Architectures for AI.” *Foundations and Trends® in Machine Learning* 2 (1): 1–127. doi:10.1561/2200000006.

Bengio, Yoshua, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. 2007. “Greedy Layer-Wise Training of Deep Networks.” *Advances in Neural Information Processing Systems* 19 (1): 153. doi:citeulike-article-id:4640046.

Berg, Wendie A., Cristina Campassi, Patricia Langenberg, and Mary J. Sexton. 2000. “Breast Imaging Reporting and Data System: Inter- and Intraobserver Variability in Feature Analysis and Final Assessment.” *American Journal of Roentgenology* 174 (6): 1769–77.

Bordes, Antoine, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2012. “Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing.” *International ...* 22: 127–35. http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS2012_BordesGWB12.pdf.

Boyd, Norman F., Lisa J. Martin, Michael Bronskill, Martin J. Yaffe, Neb Duric, and Salomon Minkin. 2010. “Breast Tissue Composition and Susceptibility to Breast Cancer.” *Journal of the National Cancer Institute* 102 (16): 1224–37. doi:10.1093/jnci/djq239.

Boyd, Norman F., Johanna M. Rommens, Kelly Vogt, Vivian Lee, John L. Hopper, Martin J. Yaffe, and Andrew D. Paterson. 2005. “Mammographic Breast Density as an Intermediate Phenotype for Breast Cancer.” *Lancet Oncology*. doi:10.1016/S1470-2045(05)70390-9.

Cireşan, Dan C., Alessandro Giusti, Luca M. Gambardella, and J??rgen Schmidhuber. 2013. “Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks.” In *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, 411–18. Springer

Berlin Heidelberg. doi:10.1007/978-3-642-40763-5_51.

Clark, Kenneth, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, et al. 2013. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository." *Journal of Digital Imaging* 26 (6): 1045–57. doi:10.1007/s10278-013-9622-7.

Claus, E B, N Risch, and W D Thompson. 1991. "Genetic Analysis of Breast Cancer in the Cancer and Steroid Hormone Study." *American Journal of Human Genetics* 48 (2): 232–42.

Cottle, Mike, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. Treister. 2013. "Transforming Health Care Through Big Data Strategies for Leveraging Big Data in the Health Care Industry." *Institute for Health Technology Transformation, Http://ihealthtran. Com/big-Data-in-Healthcare*.

Dean, Jeffrey, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc\textquotesingle aurelio Ranzato, et al. 2012. "Large Scale Distributed Deep Networks." In *Advances in Neural Information Processing Systems*, 1223–31. http://papers.nips.cc/paper/4687-large-scale-distributed-deep-networks.pdf.

Freeman, William T, and Edward H Adelson. 1991. "The Design and Use of Steerable Filters." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 13 (9): 891–906.

Gail, M H, L a Brinton, D P Byar, D K Corle, S B Green, C Schairer, and J J Mulvihill. 1989. "Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually." *Journal of the National Cancer Institute* 81 (24): 1879–86. doi:10.1093/jnci/81.24.1879.

Gail, Mitchell H., and Phuong L. Mai. 2010. "Comparing Breast Cancer Risk Assessment Models." *Journal of the National Cancer Institute* 102 (1): 665–68. doi:10.1093/jnci/djq141.

Giger, Maryellen L, Heang-Ping Chan, and John Boone. 2008. "Anniversary Paper: History and Status of CAD and Quantitative Image Analysis: The Role of Medical Physics and AAPM." *Medical Physics* 35 (12): 5799–5820. doi:10.1118/1.3013555.

Glide-Hurst, Carri K., Neb Duric, and Peter Littrup. 2007. "A New Method for Quantitative Analysis of Mammographic Density." *Medical Physics* 34 (11): 4491. doi:10.1118/1.2789407.

Greenspan, H., S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C.H. Anderson. 1994. "Overcomplete Steerable Pyramid Filters and Rotation Invariance." In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, 222–28. doi:10.1109/CVPR.1994.323833.

Haralick, Robert M., K. Shanmugam, and Its'Hak Dinstein. 1973. "Textural Features for Image Classification." *IEEE Transactions on Systems, Man, and Cybernetics* 3 (6): 610–21. doi:10.1109/TSMC.1973.4309314.

Hastie, Trevor, Robert Tibshirani, Jerome Friedman, and James Franklin. 2005. "The Elements of Statistical Learning: Data Mining, Inference and Prediction." *The Mathematical Intelligencer*. doi:10.1007/BF02985802.

Hendrick, R. Edward, and Mark A. Helvie. 2011. "United States Preventive Services Task Force

Screening Mammography Recommendations: Science Ignored." *American Journal of Roentgenology* 196 (2). doi:10.2214/AJR.10.5609.

Hinton, Geoffrey E, Simon Osindero, and Yee-Whye Teh. 2006. "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation* 18 (7): 1527–54. doi:10.1162/neco.2006.18.7.1527.

Hyvarinen, A., and E. Oja. 2000. "Independent Component Analysis: Algorithms and Applications." *Neural Networks* 13 (4–5): 411–30. doi:10.1016/S0893-6080(00)00026-5.

Irwig, L, N Houssami, and C van Vliet. 2004. "New Technologies in Screening for Breast Cancer: A Systematic Review of Their Accuracy." *British Journal of Cancer* 90 (11): 2118–22. doi:10.1038/sj.bjc.6601836.

Jonker, M. a., C. E. Jacobi, W. E. Hoogendoorn, N. J D Nagelkerke, Geertruida H. De Bock, and Johannes C. Van Houwelingen. 2003. "Modeling Familial Clustered Breast Cancer Using Published Data." *Cancer Epidemiology Biomarkers and Prevention* 12 (December): 1479–85.

Jørgensen, Karsten. 2012. "Is the Tide Turning against Breast Screening?" *Breast Cancer Research* 14: 107. doi:10.1186/bcr3212.

Keller, Brad M., Diane L. Nathan, Yan Wang, Yuanjie Zheng, James C. Gee, Emily F. Conant, and Despina Kontos. 2012. "Estimation of Breast Percent Density in Raw and Processed Full Field Digital Mammography Images via Adaptive Fuzzy c-Means Clustering and Support Vector Machine Segmentation." *Medical Physics* 39 (8): 4903–17. doi:10.1118/1.4736530.

Kopans, Daniel B. 2008. "Basic Physics and Doubts about Relationship between Mammographically Determined Tissue Density and Breast Cancer Risk." *Radiology* 246 (2): 348–53. doi:10.1148/radiol.2461070309.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances In Neural Information Processing Systems*, 1–9. doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007.

Kumar, Devinder, Alexander Wong, and David A. Clausi. 2015. "Lung Nodule Classification Using Deep Features in CT Images." In *12th Conference on Computer and Robot Vision*, 133–38. doi:10.1109/CRV.2015.25.

Kumar, Virendra, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven a. Eschrich, Matthew B. Schabath, Kenneth Forster, et al. 2012. "Radiomics: The Process and the Challenges." *Magnetic Resonance Imaging* 30 (9). Elsevier Inc.: 1234–48. doi:10.1016/j.mri.2012.06.010.

Leader, Joseph K, Thomas E Warfel, Carl R Fuhrman, Sara K Golla, Joel L Weissfeld, Ricardo S Avila, Wesly D Turner, and Bin Zheng. 2005. "Pulmonary Nodule Detection with Low-Dose CT of the Lung: Agreement among Radiologists." *Am J Roentgenol* 185 (4): 973–78. doi:10.2214/AJR.04.1225.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation*. doi:10.1162/neco.1989.1.4.541.

LeCun, Yann, L??on Bottou, Yoshua Bengio, and Patrick Haffner. 1998. "Gradient-Based

Learning Applied to Document Recognition." *Proceedings of the IEEE* 86 (11): 2278–2323. doi:10.1109/5.726791.

LeCun, Yann, Koray Kavukcuoglu, and Clément Farabet. 2010. "Convolutional Networks and Applications in Vision." *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*, 253–56. doi:10.1109/ISCAS.2010.5537907.

Lee, Honglak, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. 2011. "Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks." *Communications of the ACM* 54 (10): 95–103. doi:10.1145/2001269.

Leijenaar, Ralph T H, Sara Carvalho, Emmanuel Rios Velazquez, Wouter J C van Elmpt, Chintan Parmar, Otto S Hoekstra, Corneline J Hoekstra, et al. 2013. "Stability of FDG-PET Radiomics Features: An Integrated Analysis of Test-Retest and Inter-Observer Variability." *Acta Oncologica (Stockholm, Sweden)* 52 (May): 1391–97. doi:10.3109/0284186X.2013.812798.

Lohr, Steve. 2012. "The Age of Big Data." *The New York Times*, 1–5. doi:10.1126/science.1243089.

Ma, Shuangge, and Jian Huang. 2008. "Penalized Feature Selection and Classification in Bioinformatics." *Briefings in Bioinformatics* 9 (5): 392–403. doi:10.1093/bib/bbn027.

Madigan, M P, R G Ziegler, J Benichou, C Byrne, and R N Hoover. 1995. "Proportion of Breast Cancer Cases in the United States Explained by Well- Established Risk Factors." *Journal of the National Cancer Institute* 87 (22): 1681–85. http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L25336352%5Cnhttp://sfx.library.uu.nl/sfx?sid=EMBASE&issn=00278874&id=doi:&atitle=Proportion+of+breast+cancer+cases+in+the+United+States+explained+by+well-+established+risk+factors&sti.

Mikolov, Tomas, Anoop Deoras, Stefan Kombrink, Lukas Burget, and Jan Cernocký. 2011. "Empirical Evaluation and Combination of Advanced Language Modeling Techniques." In *INTERSPEECH*, 605–8.

Najafabadi, Maryam M, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. 2015. "Deep Learning Applications and Challenges in Big Data Analytics." *Journal of Big Data* 2 (1): 1. doi:10.1186/s40537-014-0007-7.

Nelson, Heidi D., Kari Tyne, Arpana Naik, Christina Bougatsos, Benjamin K. Chan, and Linda Humphrey. 2009. "Screening for Breast Cancer: An Update for the U.S. Preventive Services Task Force." *Annals of Internal Medicine*. doi:10.1059/0003-4819-151-10-200911170-00009.

Nicholson, Brandi T., Alexander P. LoRusso, Mark Smolkin, Viktor E. Bovbjerg, Gina R. Petroni, and Jennifer a. Harvey. 2006. "Accuracy of Assigned BI-RADS Breast Density Category Definitions." *Academic Radiology* 13 (2): 1143–49. doi:10.1016/j.acra.2006.06.005.

Nishikawa, R M, M L Giger, Kunio Doi, C E Metz, F F Yin, Carl J Vyborny, and R A Schmidt. 1994. "Effect of Case Selection on the Performance of Computer-Aided Detection Schemes." *Med Phys* 21 (2): 265–69.

Oeffinger, Kevin C., Elizabeth TH Fontham, Ruth Etzioni, Abbe Herzig, James S. Michaelson, Ya-Chen Tina Shih, and Louise C. Walter. 2015. "Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update from the American Cancer Society." *JAMA* 314 (15): 1599–1614.

Ojala, Timo, Matti Pietikäinen, and Topi Mäenpää. 2002. "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (7): 971–87. doi:10.1109/TPAMI.2002.1017623.

Pace, Lydia E, and Nancy L Keating. 2014. "A Systematic Assessment of Benefits and Risks to Guide Breast Cancer Screening Decisions." *Jama* 311 (13): 1327–35. doi:10.1001/jama.2014.1398.

Palm, Rasmus Berg. 2012. *Prediction as a Candidate for Learning Deep Hierarchical Models of Data*. Technical University of Denmark, Palm, 25. http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6284/pdf/imm6284.pdf.

Parmigiani, G, D Berry, and O Aguilar. 1998. "Determining Carrier Probabilities for Breast Cancer-Susceptibility Genes BRCA1 and BRCA2." *American Journal of Human Genetics* 62 (1): 145–58. doi:10.1086/301670.

Pinsky, Renee W., and Mark A. Helvie. 2015. "Mammographic Breast Density: Effect on Imaging and Breast Cancer Risk." *Journal of the National Comprehensive Cancer Network* 8 (10). Harborside Press: 1157–65. Accessed March 2. http://cat.inist.fr/?aModele=afficheN&cpsidt=23370070.

Qian, W, L H Li, and L P Clarke. 1999. "Image Feature Extraction for Mass Detection in Digital Mammography: Influence of Wavelet Analysis." *Medical Physics* 26 (3): 402–8.

Qian, W, L Li, S Sun, and R a Clark. 2000. "Wavelet-Based Image Processing for Digital Mammography." In *Proceedings of SPIE the International Society for Optical Engineering*. Vol. 4119. doi:10.1117/12.408648.

Qian, Wei, Laurence P Clarke, Maria Kallergi, and Robert a Clark. 1994. "Tree-Structured Nonlinear Filters in Digital Mammography." *IEEE Transactions on Medical Imaging* 13 (1): 25–36.

Qian, Wei, Lihua Li, and Laurence P. Clarke. 1999. "Image Feature Extraction for Mass Detection in Digital Mammography: Influence of Wavelet Analysis." *Medical Physics* 26: 402. doi:10.1118/1.598531.

Qian, Wei, Dansheng Song, Minshan Lei, Ravi Sankar, and Edward Eikman. 2007. "Computer-Aided Mass Detection Based on Ipsilateral Multiview Mammograms." *Academic Radiology* 14: 530–38. doi:10.1016/j.acra.2007.01.012.

Qian, Wei, Wenqing Sun, and Bin Zheng. 2015. "Improving the Efficacy of Mammography Screening: The Potential and Challenge of Developing New Computer-Aided Detection Approaches." *Expert Review of Medical Devices* 12 (5): 497–99. doi:10.1586/17434440.2015.1068115.

Raghupathi, Wullianallur, and Viju Raghupathi. 2014. "Big Data Analytics in Healthcare: Promise and Potential." *Health Information Science and Systems* 2 (1): 3.

doi:10.1186/2047-2501-2-3.

Roth, Holger, Le Lu, Jiamin Liu, Jianhua Yao, Ari Seff, Kevin Cherry, Lauren Kim, and Ronald Summers. 2015. "Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation." *IEEE Transactions on Medical Imaging* 35 (5): 1. doi:10.1109/TMI.2015.2482920.

Samulski, Maurice, and Nico Karssemeijer. 2011. "Optimizing Case-Based Detection Performance in a Multiview CAD System for Mammography." *IEEE Transactions on Medical Imaging* 30 (4): 1001–9. doi:10.1109/TMI.2011.2105886.

Schmidhuber, Jürgen. 2015. "Deep Learning in Neural Networks: An Overview." *Neural Networks* 61. Elsevier Ltd: 85–117. doi:10.1016/j.neunet.2014.09.003.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64. doi:10.1214/aos/1176344136.

Shen, Wei, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. 2015a. "Multi-Scale Convolutional Neural Networks for Lung Nodule Classification." In *International Conference on Information Processing in Medical Imaging*, 588–99.

———. 2015b. "Multi-Scale Convolutional Neural Networks for Lung Nodule Classification." In *Information Processing in Medical Imaging*, 588–99. doi:10.1097/00003072-199201000-00023.

Simard, P.Y., D. Steinkraus, and J.C. Platt. 2003. "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis." *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, 1–6. doi:10.1109/ICDAR.2003.1227801.

Smith, Robert A., Deana Manassaram-Baptiste, Durado Brooks, Mary Doroshenk, Stacey Fedewa, Debbie Saslow, Otis W. Brawley, and Richard Wender. 2015. "Cancer Screening in the United States, 2015: A Review of Current American Cancer Society Guidelines and Current Issues in Cancer Screening." *CA: A Cancer Journal for Clinicians* 65 (1): 30–54. doi:10.3322/caac.21261.

Smith, Robert A, Stephen Duffy, and Laszlo Tabar. 2012. "Breast Cancer Screening : The Evolving Evidence." *Oncology* 26 (5): 471–86.

Smith, Robert a, Deana Manassaram-Baptiste, Durado Brooks, Mary Doroshenk, Stacey Fedewa, Debbie Saslow, Otis W Brawley, and Richard Wender. 2015. "Cancer Screening in the United States, 2015: A Review of Current American Cancer Society Guidelines and Current Issues in Cancer Screening Robert." *Ca Cancer J Clin 2015;65:30–54* 65: 30–54. doi:10.3322/caac.21261.

Socher, Richard, Eh Huang, and Jeffrey Pennington. 2011. "Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection." *Advances in Neural Information Processing Systems*, 801–9. http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2011_0538.pdf%5Cnhttps://papers.nips.cc/paper/4204-dynamic-pooling-and-unfolding-recursive-autoencoders-for-paraphrase-detection.pdf.

Sun, Wenqing, Tzu-Liang (Bill) Tseng, Wei Qian, Jianying Zhang, Edward C. Saltzstein, Bin

Zheng, Fleming Lure, Hui Yu, and Shi Zhou. 2015. “Using Multiscale Texture and Density Features for near-Term Breast Cancer Risk Analysis.” *Medical Physics* 42 (6): 2853–62. doi:10.1118/1.4919772.

Sun, Wenqing, Tzu-Liang B. Tseng, Bin Zheng, Jianying Zhang, and Wei Qian. 2015. “A New Breast Cancer Risk Analysis Approach Using Features Extracted from Multiple Sub-Regions on Bilateral Mammograms.” In *SPIE Medical Imaging. International Society for Optics and Photonics,* 9414:941422. doi:10.1117/12.2076633.

Sun, Wenqing, Bin Zheng, Fleming Lure, Teresa Wu, Jianying Zhang, Benjamin Y. Wang, Edward C. Saltzstein, and Wei Qian. 2014. “Prediction of near-Term Risk of Developing Breast Cancer Using Computerized Features from Bilateral Mammograms.” *Computerized Medical Imaging and Graphics* 38 (5). Elsevier Ltd: 348–57. doi:10.1016/j.compmedimag.2014.03.001.

Sun, Xuejun, Wei Qian, and Dansheng Song. 2004. “Ipsilateral-Mammogram Computer-Aided Detection of Breast Cancer.” *Computerized Medical Imaging and Graphics* 28 (3): 151–58. doi:10.1016/j.compmedimag.2003.11.004.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. “Going Deeper with Convolutions.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9. doi:10.1109/ICCV.2011.6126456.

Tan, Maxine, Bin Zheng, Pandiyarajan Ramalingam, and David Gur. 2013. “Prediction of near-Term Breast Cancer Risk Based on Bilateral Mammographic Feature Asymmetry.” *Academic Radiology* 20 (12): 1542–50. doi:10.1016/j.acra.2013.08.020.

Tang, Xiaoou. 1998. “Texture Information in Run-Length Matrices.” *IEEE Transactions on Image Processing* 7 (11): 1602–9. doi:10.1109/83.725367.

van Tulder, Gijs, and Marleen de Bruijne. 2016. “Combining Generative and Discriminative Representation Learning for Lung CT Analysis with Convolutional Restricted Boltzmann Machines.” *IEEE Transactions on Medical Imaging* 35 (5): 1262–72.

Vincent, Pascal, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. “Extracting and Composing Robust Features with Denoising Autoencoders.” In *Proceedings of the 25th International Conference on Machine Learning*, 1096–1103. doi:10.1145/1390156.1390294.

Wang, Li, Ji Zhu, and Hui Zou. 2006. “The Doubly Regularized Support Vector Machine.” *Statistica Sinica* 16 (2): 589. http://www.stat.lsa.umich.edu/~jizhu/pubs/Wang-Sinica06.pdf.

Wang, Xingwei, Dror Lederman, Jun Tan, Xiao Hui Wang, and Bin Zheng. 2011. “Computerized Prediction of Risk for Developing Breast Cancer Based on Bilateral Mammographic Breast Tissue Asymmetry.” *Medical Engineering and Physics* 33 (8). Institute of Physics and Engineering in Medicine: 934–42. doi:10.1016/j.medengphy.2011.03.001.

Way, Ted W, Lubomir M Hadjiiski, Berkman Sahiner, Heang-Ping Chan, Philip N Cascade, Ella A Kazerooni, Naama Bogot, and Chuan Zhou. 2006. “Computer-Aided Diagnosis of Pulmonary Nodules on CT Scans: Segmentation and Classification Using 3D Active Contours.” *Medical Physics* 33 (7): 2323–37. doi:10.1118/1.2207129.

Way, Ted W, Berkman Sahiner, Heang-Ping Chan, Lubomir Hadjiiski, Philip N Cascade, Aamer Chughtai, Naama Bogot, and Ella Kazerooni. 2009. "Computer-Aided Diagnosis of Pulmonary Nodules on CT Scans: Improvement of Classification Performance with Nodule Surface Features." *Medical Physics* 36 (7): 3086–98. doi:10.1118/1.3140589.

Wei, Chia-hung, Yue Li, and Chang-tsun Li. 2007. "Effective Extraction of Gabor Features for Adaptive Mammogram Retrieval." In *Multimedia and Expo, 2007 IEEE International Conference on*, 1503–6.

Wei, Jun, Heang-Ping Chan, Berkman Sahiner, Chuan Zhou, Lubomir M Hadjiiski, Marilyn a Roubidoux, and Mark a Helvie. 2009. "Computer-Aided Detection of Breast Masses on Mammograms: Dual System Approach with Two-View Analysis." *Medical Physics* 36 (10): 4451–60. doi:10.1118/1.3220669.

Wei, Jun, Heang-Ping Chan, Yi-Ta Wu, Chuan Zhou, Mark A Helvie, Alexander Tsodikov, Lubomir M Hadjiiski, and Berkman Sahiner. 2011. "Association of Computerized Mammographic Parenchymal Pattern Measure with Breast Cancer Risk: A Pilot Case-Control Study." *Radiology* 260 (1): 42–49. doi:10.1148/radiol.11101266.

Wiemker, Rafael, Martin Bergtholdt, Ekta Dharaiya, Sven Kabus, and Michael C Lee. 2009. "Agreement of CAD Features with Expert Observer Ratings for Characterization of Pulmonary Nodules in CT Using the LIDC-IDRI Database." *Medical Imaging 2009: Computer-Aided Diagnosis* 7260 (1): 72600H. doi:10.1117/12.811569.

Wolfe, John N. 1976. "Risk for Breast Cancer Development Determined by Mammographic Parenchymal Pattern." *Cancer* 37 (5): 2486–92.

Ye, Gb, and Y Chen. 2011. "Efficient Variable Selection in Support Vector Machines via the Alternating Direction Method of Multipliers." *Of the International Conference on Artificial* 15: 832–40. http://www.stat.wisc.edu/~wahba/stat840/talks/sui/ye.chen.2011.pdf.

Zheng, Bin, Lara A Hardesty, William R Poller, Jules H Sumkin, and Sara Golla. 2003. "Mammography with Computer-Aided Detection: Reproducibility Assessment - Initial Experience." *Radiology* 228 (1): 58–62. http://radiology.rsnajnls.org/cgi/content/abstract/228/1/58.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 67 (2). Blackwell: 301–20. http://cat.inist.fr/?aModele=afficheN&cpsidt=16642045.

Vita

Wenqing Sun was born in Wuhan, China. He graduated with a bachelor's in Applied Physics from Huazhong University of Science and Technology, HUST. Later, he pursued Electrical and Computer Engineering M.S. from University of Texas at El Paso (UTEP) with a focus in medical imaging analysis. Then he went on to pursue his Ph.D. under the guidance of Dr. Qian from UTEP Electrical and Computer Engineering department.

Wenqing actively published seventeen quality publications during his graduate study. These include six peer reviewed journal papers, eleven peer reviewed conferences and two technical oral presentations. He also worked as a research intern and machine learning consultant at Hologic Inc.

After graduation, he plans on pursuing a career either in academia or industry in the fields of medical image analysis, computer vision and machine learning.

Contact Information: wsun2@miners.utep.edu

This thesis/dissertation was typed by Wenqing Sun.