UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

THYROID NODULE ULTRASOUND IMAGE ANALYSIS AND FEATURE EXTRACTION

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

XIAOMENG DONG
Norman, Oklahoma
2018

THYROID NODULE ULTRASOUND IMAGE ANALYSIS AND FEATURE EXTRACTION

A THESIS APPROVED FOR THE
GALLOGLY COLLEGE OF ENGINEERING

BY

_____
Dr. Theodore B. Trafalis, Chair

_____
Dr. Andrew H. Fagg

_____
Dr. Charles D. Nicholson

© Copyright by XIAOMENG DONG 2018
All Rights Reserved.

Table of Contents

List of Tables

List of Figures

Abstract

In this study, I introduce a novel workflow for extracting useful features in thyroid ultrasound images using deep learning and machine learning methods. The methodology combines Convolutional Auto-Encoder, Local Binary Patterns, Histogram of Oriented Gradients and professional image characterization together to extract useful information from medical images.

Multiple machine learning classifiers are used to build an effective thyroid tumor diagnosis model from extracted features. The experimental results show that Support Vector Machine with a specifically designed preprocessing scheme and a customized objective function outperforms human on the test set. The final model can effectively reduce the number of unnecessary biopsies and the number of missing malignancies.

Chapter 1 Introduction

Humans have been trying to teach computers how to recognize meaningful objects since the late 1960s. Back then, Artificial Intelligence (AI) pioneers thought Computer Vision could be accomplished in a summer project by attaching camera to computer and let the computer "recognize what they saw" [1]. However, as soon as they realized the complexity hidden behind the pixels, they concluded that Computer Vision was greatly more than a summer project with a question mark: why is it so hard to explicitly express the logic of visual perception, what we as humans take for granted?

In the next several decades, researchers began searching for answers by studying the visual systems of many biological entities. They developed quantitative analysis and rigorous mathematical models to simulate the visual systems of different species [2, 3]. One of the most influential mathematical model in computer vision that forms the architectural basis of most modern deep learning algorithms is the Neocognitron. (Figure 1)

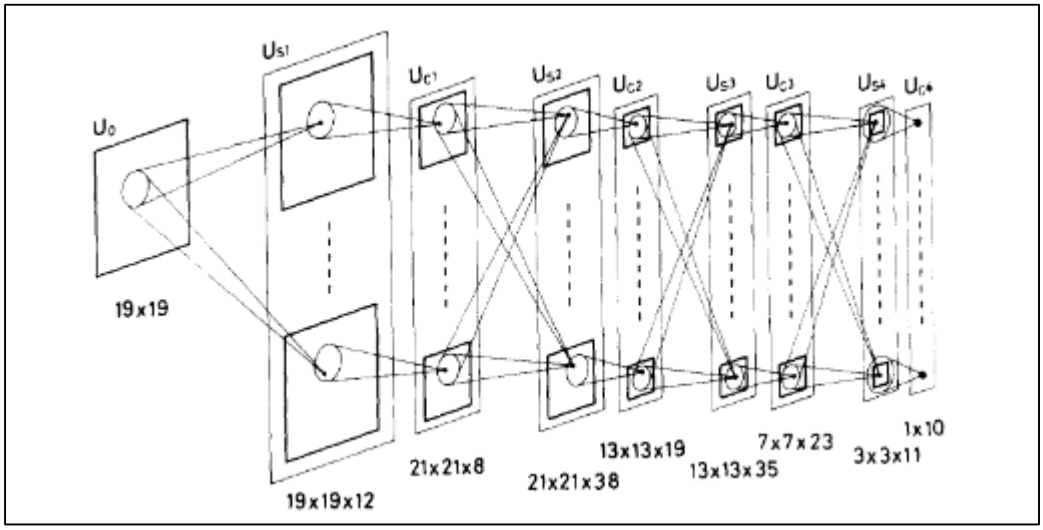

Figure 1 Structure of Neocognitron. (Figure from Fukushima, 1980)

Fukushima first proposed the concept of Neocognitron in the 1980s by mimicking mammal's visual cortex [4]. The main breakthrough of the Neocognitron is that it can be insensitive to position shifts and size changes. However, this discovery did not attract much attention at the time because there were no efficient techniques found to self-organize (train) such visual model. Designing and specifying all the parameters and details required to achieve certain recognition task for this model was extremely labor intensive, even impossible.

From the 80s to 90s, researchers developed efficient ways to automatically train network model by applying chain rules to recursively calculate the gradient of error with respect to each parameter, namely, the back-propagation algorithm [5]. In 1989 Yann LeCun implemented a network architecture similar to the Neocognitron with layers having local connections and shared weights. These architectures are later referred to as the Convolutional Neural Networks (CNN). LeCun trained the CNN using the back-propagation algorithm to recognize hand-written digits and achieved great performance [6]. His work is one of the earliest demonstrations that a CNN could be applied to real world visual applications and his system was later expanded for commercial use in banks and post offices.

The model complexity of CNN can be significantly less than the complexity of Neocognitron because CNN allows weight-sharing among filters and it also uses pooling to reduce network size. However, given the computing limit in the 90s, if stacking many layers together, the number of training parameters of CNN could easily become overwhelming. Furthermore, another obstacle which limits the depth of CNN is the vanishing gradient problem: the magnitude of gradient gradually decreases as error

propagates backwards. This phenomena is due to the fact that the gradient of error with respect to any parameter is calculated by chain rule. Longer the "chain" is, more likely the magnitude of gradient will be reduced by intermediate saturated nodes. Consequently, if the CNN model was too deep, the first several layers were very likely to be poorly trained back then. Therefore, CNN had very limited performance on more complex objects. The bad performance was once thought to originate from the convolutional model itself other than the lack of training. Consequently, CNN remained silent for another decade.

Although the potential of the CNN was masked during the late 90s, the rise of kernel machines [7] offered a valuable bridge for object recognition. However, the performance of kernel machines highly relies on the quality of features. Coming up with useful features from images has become a new challenge since then. Therefore, many computer vision researchers developed visual descriptors that can quantify useful visual information such as the shape, texture, motion. Some useful visual feature descriptors include Local Binary Patterns (LBP) [8], Histogram of Oriented Gradients (HOG) [9] and Grey Level Co-occurrence Matrices (GLCM) [10]. The common wisdom of image recognition at the time was to use visual descriptors to extract features from image and feed the features to a Support Vector Machine.

Visual descriptors can effectively extract information from images and they also serve as a way of reducing dimensionality which maps the pixel space of image to a lower dimensional vector (Figure 2). However, a great amount of information is lost and it does not perform well on an image recognition task like ILSVRC [11] (ImageNet

Large Scale Visual Recognition Challenge), which is also known as the Olympics of Computer Vision.

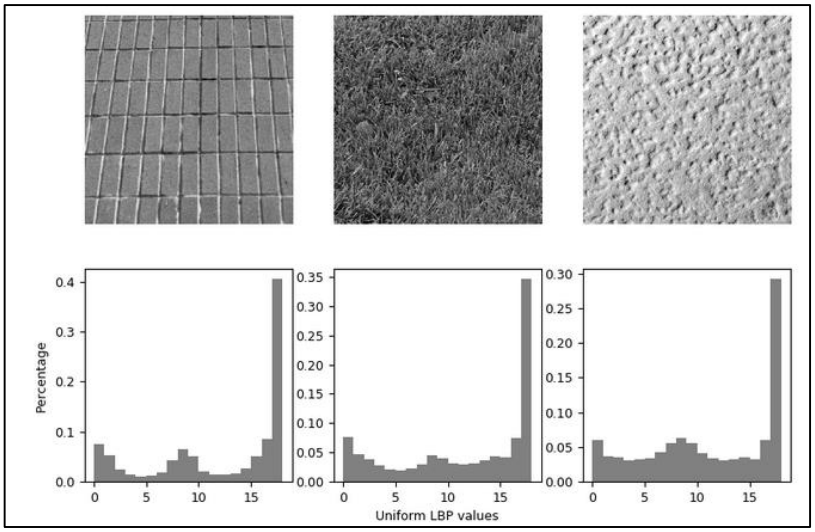

Figure 2 Illustration of how LBP map Images to a Vector

Things have greatly changed from 1998 to 2010, both data storage technology and computing power has skyrocketed. The use of GPUs alone increased the computational speed by a factor of 1000, which enabled us to build and train more complex models within feasible time limits. Deep learning, which saves the burden of feature engineering by building deep hierarchical representation of input, gradually became reality on many problem domains. In the field of Computer Vision, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton shocked the world by using a deep convolutional neural network to win the ILSVRC in 2012 [12]. Since then, deep learning began to attract more and more attentions among computer vision and AI researchers.

Over the last 6 years, deep learning has made remarkable achievements on image recognition, speech recognition, text processing and automatic game playing. For image recognition, it has boosted the performance of many computer vision tasks such

as face recognition, object recognition and image caption. It has been shown to be as good as humans at recognizing objects when the winner of ImageNet 2017 achieves 2.25% error rate [13] (lower than human error). More importantly, it is helping computer vision researchers achieve what was meant to be a summer project a long time ago.

In medicine and healthcare, the potential of applying deep learning techniques to solve medical vision problems has caught the attention of both industry and academia. For instance, IBM's Watson [14, 15] has been developing optimized cancer diagnosis and treatment plans based on patient's medical records. Google launched DeepMind Health [16, 17] to develop computer-aided diagnostic screening for disease and achieved great results in diagnosing diabetic retinopathy.

A lot of medical data is published and accessible for anyone who wants to make medical discoveries, such as TCIA (The Cancer Imaging Archive), IDA (LONI Image Data Archive) and many others. The number of publication on medical AI is also exploding, especially after Google published its deep learning framework TensorFlow [18] to the AI community.

In Biomedical Imaging, deep learning helped researchers to solve wide ranges of medical image recognition problems. Broad categories of those problems are: anomaly detection, diagnosis classification, segmentation, recognition and brain decoding, see Table 1.

Table 1 List of Deep Learning Applications in Biomedical Imaging

Anomaly classification[19-29]	Segmentation[30-38]	Recognition[39-44]	Brain decoding[45-46]
Gene expression pattern	Cell structure	Cell nuclei	Behavior
Cancer	Neuronal sturcture	Finger joint	
Alzheimer's disease	Vessel map	Anatomical structure	
Schizophrenia	Brain tumor		

One of the most important medical vision tasks is medical image diagnosis, for example, diagnosing whether a tumor is benign or malignant. A tumor (nodule) often results from cells that have their cell cycle broken by mutation. They can rapidly reproduce themselves and create their own base (tumor). Fortunately, most of those tumors are controlled by a membrane that prevents them from traveling elsewhere in our body (encapsulated), these tumors are described as benign tumor. However, some tumors may not be well-encapsulated and cancer cells can therefore travel through blood vessels to somewhere else (metastasis), in this case we have malignant tumor.

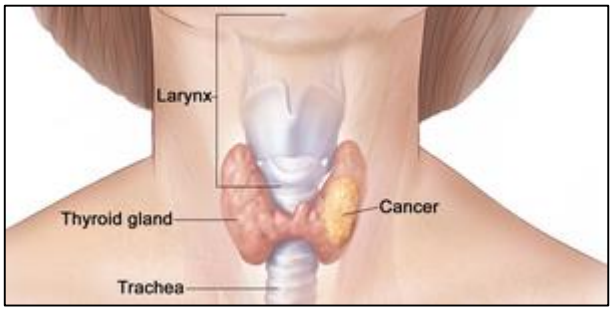

Figure 3 Illustration of Thyroid and Thyroid Cancer

A tumor can originate from any cells at any part of our body, thyroid is no exception (Figure 4). The development of thyroid nodules is a common pathology that happens most frequently in females around 20-55 age. The most common type of cancer associated with thyroid nodules is papillary cancer that represents about 75% to 85% of all thyroid cancer cases [47]. Fortunately, most of the thyroid nodules are benign [48]. Thyroid nodules are usually imaged by ultrasound, thyroid nodules can have many different appearance on ultrasound images (Figure 4). Radiologists will diagnose the tumor by assessing and characterizing the thyroid nodule ultrasound image.

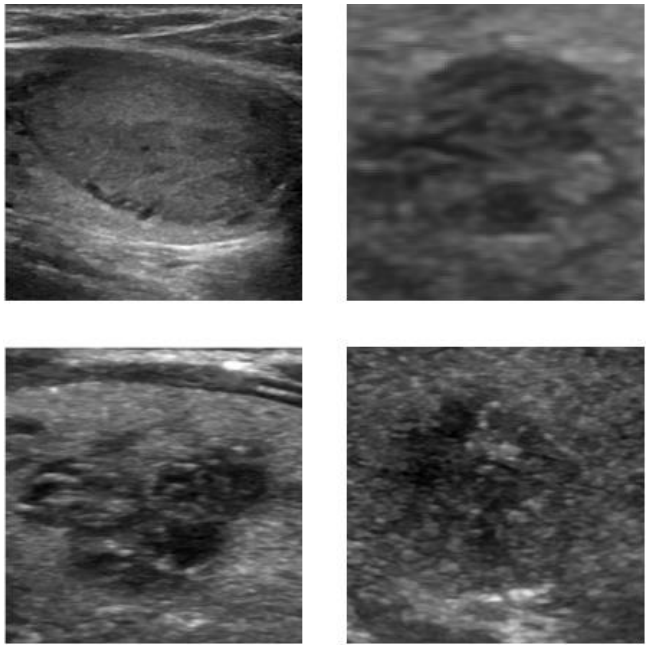

Figure 4 Examples of Thyroid Nodule Ultrasound Images

One of the most commonly used image characterization schemes for diagnosing thyroid tumors is the Thyroid Imaging Reporting and Data Systems (TIRADS) [49]. The TIRADS system includes several informative characterizations, such as shape, border, composition, calcification and many others. These characterizations are often referred as "TIRADS features" in many studies. In addition, TIRADS also provides a scoring system that helps radiologists to access the risk of malignancy. More details about TIRADS will be given in Chapter 4.

The rest of the thesis is organized as follows; Chapter 2 discusses the objective and scope of this study, Chapter 3 talks about the challenges of this thesis and more generally, the challenges of medical AI. Chapter 4 presents the dataset in the study. Chapter 5 illustrates the methodology used. Chapter 6 provides experimental results and analysis.

Chapter 2 Objective and Scope

The objective of this study is to analyze the medical records of patients with thyroid nodules through a deep learning and computer vision framework. The medical records used in this study include ultrasound biomedical images, biopsy results, radiologist's annotations and patient's demographic information. The ultimate goal of this study is to improve the current thyroid nodule diagnosis system and help radiologists better assess the risk of malignancy and make better decisions about a treatment plan.

The final decision will be weighed on both medical and economical perspective; namely, it aims to achieve a balance between risk of missing malignancy and the cost of unnecessary biopsy. The population target in this study mainly resides in the northeastern coast of United States (Figure 5), and most patients live near New York.

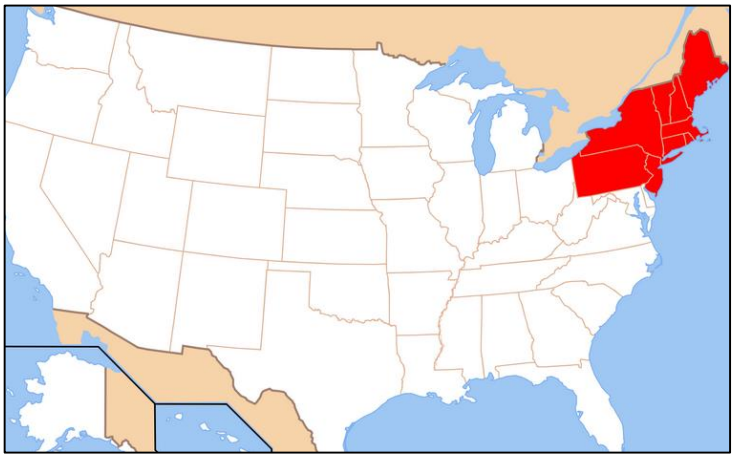

Figure 5 Population Target in This Study (Image from Wikipedia)

Chapter 3 Challenges in Medical AI

Before moving forward to the methodology, it is worth mentioning the challenges associated with this study and more generally, challenges that are prevalent in medical image recognition AI. This section will also briefly mention different approaches used in this thesis to tackle these challenges.

3.1 Hard to Make Decisions

It is very common that classifiers will produce a probabilistic output. The final decision will be the class with the highest probability. However, in medical practice, it is much more complicated. Imagine if a tumor is diagnosed benign with 80% chance, should people do a surgical biopsy or just ignore it?

The answer depends on two things: the cost of biopsy (both economical and mental) and spread rate of cancer cells. Unfortunately, these two factors may vary wildly from disease to disease, from individual to individual and from instance to instance.

Ideally we want to have an algorithm that can yield both low false negatives and low false discovery rates, but in reality, there is often a trade-off between them. From a machine learning perspective, in order to build a model that knows what we want, the objective function and model selection criteria have to account for this trade-off.

In the case of thyroid nodule diagnosis, the cost of biopsy is relatively low (~$200), but the cost of not detecting a malignant nodule is relatively expensive. Therefore, we want our false negatives to be minimal while false positives can be somewhat tolerable.

In the objective function, the weight of the positive class is modified to make sure that the positive data have enough significance (but not too much) in the optimization formulation. For the final model selection, radiologists provide us with a scoring function that is considered optimal by them on clinical and economical perspective. The scoring function is a weighted linear combination of the false negative rate, false discovery rate and accuracy. More details are provided later in Chapter 5.9.

3.2 Learning without Human's Guarantee of Success

If there is one thing that makes medical image recognition different among other computer vision problems, it would be learning something without a guarantee of success by human, which leads to the following two problems; (1) we do not know the performance limit and (2) human cannot help much in improving model performance.

In order to elaborate on these points further, I compare diagnosing malignancy with the ImageNet object recognition problem. ImageNet is a dataset that is used for identifying daily objects (like cars, chairs, cats etc. See Figure 6).

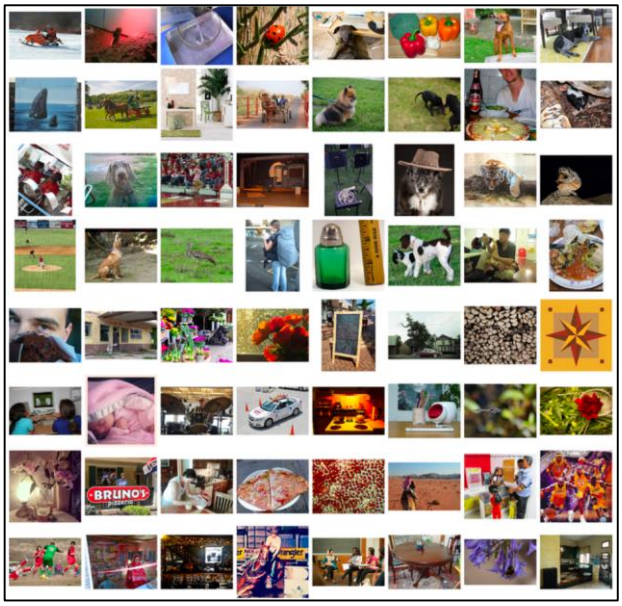

Figure 6 Sample Images from ImageNet Dataset

Given any image in the ImageNet dataset, we have a prior knowledge that the information in the image is enough to make a near-perfect decision (because human can do it, given enough experts). Therefore, a performance limit barely exists in the ImageNet problem. On the other hand, however, diagnosing the malignancy of a tumor from an image is equivalent to "judging a book by its cover", with a book representing the genotype of tumor, the cover being the phenotype of tumor. A Genotype can uniquely map to a phenotype but the opposite is not true. Therefore, we can think of a phenotype as a lower dimensional representation of the nature of tumor. Without further assumptions, we cannot recover the information of the higher dimensional space given its lower dimensional representation. Therefore, using "cover" to predict the quality of a "book" will never be perfect due to lack of information. In medical image diagnosis, neither can we perfectly diagnose malignancy of a tumor nor do we know where the performance limit is. It is tricky to answer the simplest question such as: can I achieve 90% accuracy on my dataset? Or what is the performance limit?

Furthermore, humans can achieve a high performance on the ImageNet recognition problem. Given any image, humans know where to look and what the decision rules are, even though humans may not be able to mathematically express the rules. If we encounter any performance problems, we can always tune parameters or adjust the model architecture by introspection of our own recognition behavior.

However, humans are not known to handle medical diagnosis well. Once AI has exceeded human performance, if the performance still does not meet our expectation, there is little we can do.

In this study, due to the lack of a known performance limit, I have chosen human performance to be the goal of performance. Once the performance exceeds human, I will settle with whatever results I end up with. Fortunately, the final algorithm greatly outperforms humans on the test set. More details are presented in Chapter 6.

3.3 More Complex Problem with Less Data

In ImageNet, we are required to identify whether the image is a cat or not (for simplicity). In medical imaging domain, instead of asking whether an image is a tumor or not, we are taking one step further by asking: what is the nature of that tumor? Intuitively speaking, the problem for medical diagnosis is much more complex than daily object recognition (Figure 7), thus medical diagnosis requires a more complex model to capture all the complexity.

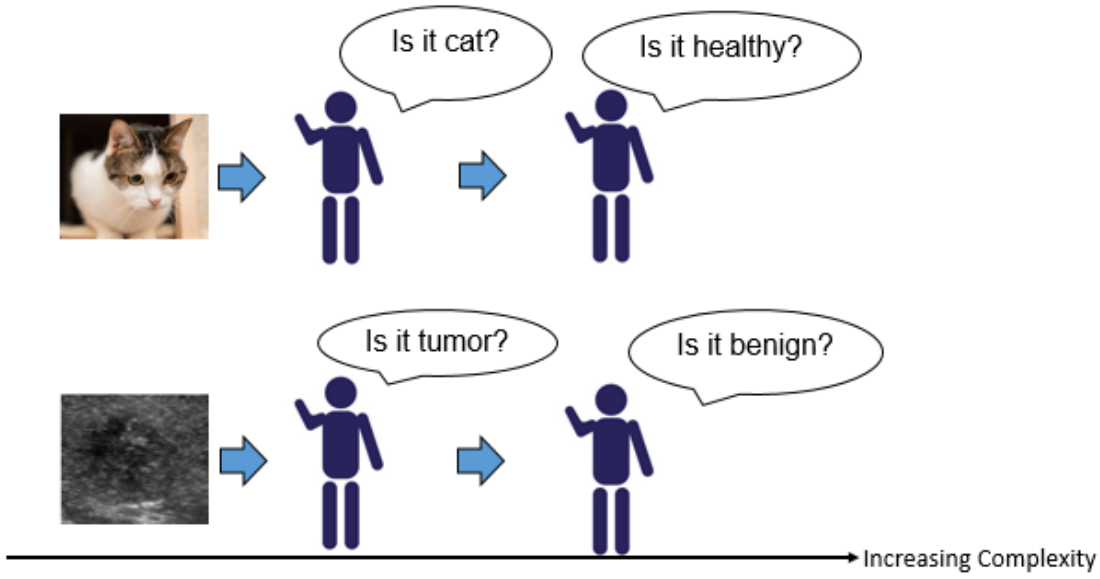

Figure 7 Object Recognition V.S. Medical Diagnosis

More model complexity means that more data is required to build a good model. However, in medical AI, instead of more data, the amount of available data is usually

orders of magnitude less than other recognition tasks. For a deep learning task, we are often used to image datasets with close to several million samples. However, for medical images, the sample size is often on the order of one hundred. The lack of medical data is mainly due to the expense of the medical data acquisition process. There are many legal issues regarding privacy of medical records as well.

Figure 8 demonstrates the dataset size in TCIA archive, one of the largest public cancer imaging database. The collection is sorted based on the number of subjects (sample size). As we can see, only one collection has more than 10,000 subjects and two collections with more than 1,000 sample size. Most of the others have sample sizes of only several hundred.

Collection	Cancer Type	Modalities	Subjects	Location	Metadata	Access	Status	Updated
National Lung Screening Trial	Lung Cancer	CT	26254	Chest	Yes	Limited	Complete	2013/03/01
CBIS-DDSM	Breast Cancer	MG	6671	Breast	Yes	Public	Complete	2017/09/27
LIDC-IDRI	Lung Cancer	CT, CR, DX	1010	Chest	Yes	Public	Complete	2012/03/21
CT Colonography	Colon Cancer	CT	825	Colon	Yes	Public	Complete	2011/10/31
NSCLC-Radiomics	Lung Cancer	CT, RTSTRUCT	422	Lung	Yes	Public	Ongoing	2016/05/20
PROSTATEx	Prostate Cancer	MR	346	Prostate	Yes	Public	Complete	2017/03/30
MyelomaTT3PET	Myeloma	PET	300	Whole Body	Yes	Public	Coming Soon	2016/04/16
Head-Neck-PET-CT	Head and Neck Cancer	PT, CT, RTSTRUCT, RTPLAN, RTDOSE	298	Head-Neck	Yes	Public	Complete	2017/11/27
TCGA-KIRC	Kidney Renal Clear Cell Carcinoma	CT, MR, CR	267	Renal	Yes	Public	Complete	2014/10/09
TCGA-GBM	Glioblastoma Multiforme	MR, CT, DX	262	Brain	Yes	Public	Complete	2014/05/08
RIDER Lung PET-CT	Lung Cancer	PT, CT	244	Lung	No	Public	Complete	2011/11/25
ISPY1	Breast Cancer	MR, SEG	222	Breast	Yes	Public	Complete	2016/08/31
HNSCC	Head and Neck Squamous Cell Carcinoma	CT, PT, MR, RTSTRUCT, RTPLAN, RTDOSE	215	Head-Neck	Yes	Public	Complete	2017/10/31
NSCLC-Radiogenomics	Non-small Cell Lung Cancer	PT, CT, SEG, SR	211	Chest	Yes	Public	Complete	2017/12/04
TCGA-LGG	Low Grade Glioma	MR, CT	199	Brain	Yes	Public	Complete	2014/09/04

Figure 8 ICIA Cancer Imaging Archive Dataset Size in Sorted Order

In this study, the sample size of ultrasound thyroid image is 3183. More details about dataset will be discussed in the Chapter 4. In order to deal with the scarcity of data, image augmentation is used to artificially generate medical images for training. Details of Image augmentation will be provided in Chapter 5.

3.4 Traditional CNN is Limited in Medical Image Recognition

Convolution is the basis of CNN and it works by having a kernel to capture specific local patterns and gradually assemble layers of local patterns together to form more general patterns. For example, given an image of a human face, a convolution may first extract edges in the first layer, then use those edges to construct simple shapes in the second layer and then use these shapes to determine higher-level features, such as facial shapes [50] (Figure 9).

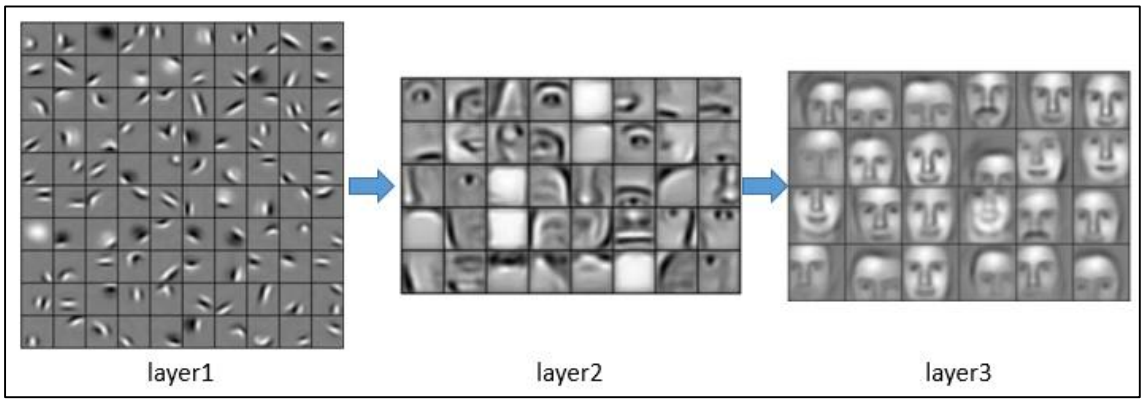

Figure 9 Hierarchy of Feature Representation by Convolution (Honglak Lee et al, 2009)

By using the Convolutional Neural Networks (CNN) architecture for generalization, we are essentially making an assumption: all specific local patterns in testing data are arranged by a similar rule as in training data. However, in medical images, this assumption does not hold.

The CNN architecture has been widely successful in recognizing images that has specific arrangements of local patterns. For example, all human faces follow certain rules: eyes are always above the nose, the mouth is always below the nose. Similarly, all cats image follow certain spatial arrangements: they all have four legs, their eyes are in round shape. Because these arrangements are consistent through training and testing, Google's Inception can effectively recognize cats and Facebook can automatically tag you in online images (Figure 10).

Figure 10 CNN is Successful at Recognizing Objects that Follow Specific Spatial Arrangement of Patterns, Such as Face Image

Medical images, on the other hand, do not have spatial feature arrangements for a CNN to capture. Unlike human faces, important information is not regulated in any spatial manner. For example, shape, size and border are all key characteristics describing a tumor which can appear in every imaginable manner (Figure 11). Thus, merely "remembering" what a benign nodule looks like will not guarantee us to successfully recognize a new benign tumor in the future.

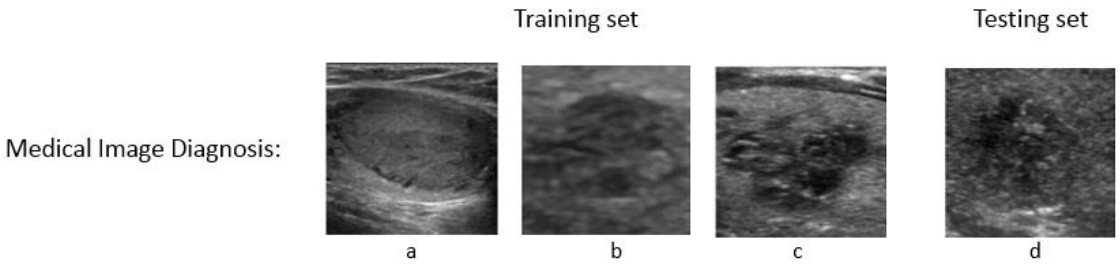

Figure 11 Different Appearance of Tumor

To make matters worse, similar appearing nodules can have wildly different and clinically important differences in pathologic categorization. In Figure 11, tumor d is very similar to tumor b. However, nodule b is benign by pathologic characterization, whereas nodule d is malignant.

Therefore, in this study, I do not use a deep convolutional network as an end-to-end predictor, instead, I use it as a feature extractor that "learns" what constitutes these ultrasound tumor images. Later, I use these learned features as attributes and feed them into other classifiers. More details can be found in Chapter 5.4.

3.5 Data is Usually Imbalanced

Medical data is usually imbalanced and current techniques dealing with imbalanced data usually fall into two categories: data manipulation and algorithm manipulation. Data manipulation usually involves subsampling data of the major class or replicating data of the minor class. Neither imbalanced data manipulation methods work in medical image problems. First, medical data is scarce and expensive, down sampling data is not an option. Second, creation of minor class medical image data is technically not possible.

Algorithm manipulation is easier to achieve and is therefore used in this study. As mentioned in section 3.1, data imbalance is dealt with by applying additional

weights in the positive class and creating a customized model selection scoring function, more details can be found in Chapter 5.9.

Chapter 4 Dataset

4.1 Ultrasound Images and Biopsy data

3183 ultrasound thyroid images that are collected by East River Medical Imaging center over 5-year period using General Electric Logiq E9 are used in this study (Figure 12). Each thyroid ultrasound image is evaluated by a trained radiologist and have the nodule (tumor) area cropped for further analysis.

Figure 12 Thyroid Nodule Image Data Acquisition

Each thyroid image has its own biopsy results represented by Bethesda Grade [51]. Bethesda grade divides the malignancy risk of a thyroid nodule into 6 categories, with label 1 being outlier, label 2 being benign and label 3 to 6 are associated with increasing risks as listed in Table 2. Label 6 has the highest probability of malignancy.

In this study, images with Bethesda grade 1 are deleted from the dataset. Next, all tumor images that have Bethesda grade from 3 to 6 are grouped to be suspicious (or positive), and the rest are tumor images that are diagnosed as benign (or negative). Under this grouping scheme, 2451 out of 3183 thyroid nodule images are associated with benignity, which is around 77% of the entire sample set.

Table 2 The Bethesda System for Reporting Thyroid Biopsy Results (Edmund and Syed, 2009)

I. Nondiagnostic or Unsatisfactory
Cyst fluid only
Virtually acellular specimen
Other (obscuring blood, clotting artifact, etc)
II. Benign
Consistent with a benign follicular nodule (includes adenomatoid nodule, colloid nodule, etc)
Consistent with lymphocytic (Hashimoto) thyroiditis in the proper clinical context
Consistent with granulomatous (subacute) thyroiditis
Other
III. Atypia of Undetermined Significance or Follicular Lesion of Undetermined Significance
IV. Follicular Neoplasm or Suspicious for a Follicular Neoplasm
Specify if Hürthle cell (oncocytic) type
V. Suspicious for Malignancy
Suspicious for papillary carcinoma
Suspicious for medullary carcinoma
Suspicious for metastatic carcinoma
Suspicious for lymphoma
Other
VI. Malignant
Papillary thyroid carcinoma
Poorly differentiated carcinoma
Medullary thyroid carcinoma
Undifferentiated (anaplastic) carcinoma
Squamous cell carcinoma
Carcinoma with mixed features (specify)
Metastatic carcinoma
Non-Hodgkin lymphoma
Other

4.2 Image Characterizations and TIRADS Features

In addition to the ultrasound image itself, image characteristics for 1434 images were evaluated by trained radiologists, according to the Thyroid Imaging Reporting and Data Systems (TIRADS). The TIRADS is a risk stratification system based on image characteristics proposed by the American College of Radiology (ACR) [49]. It is also a scoring system that sums the scores across multiple ultrasound features categories. The overall score indicates the probability of malignancy with a recommendation for either fine needle aspiration (FNA) or a follow-up checkup (see Figure 13).

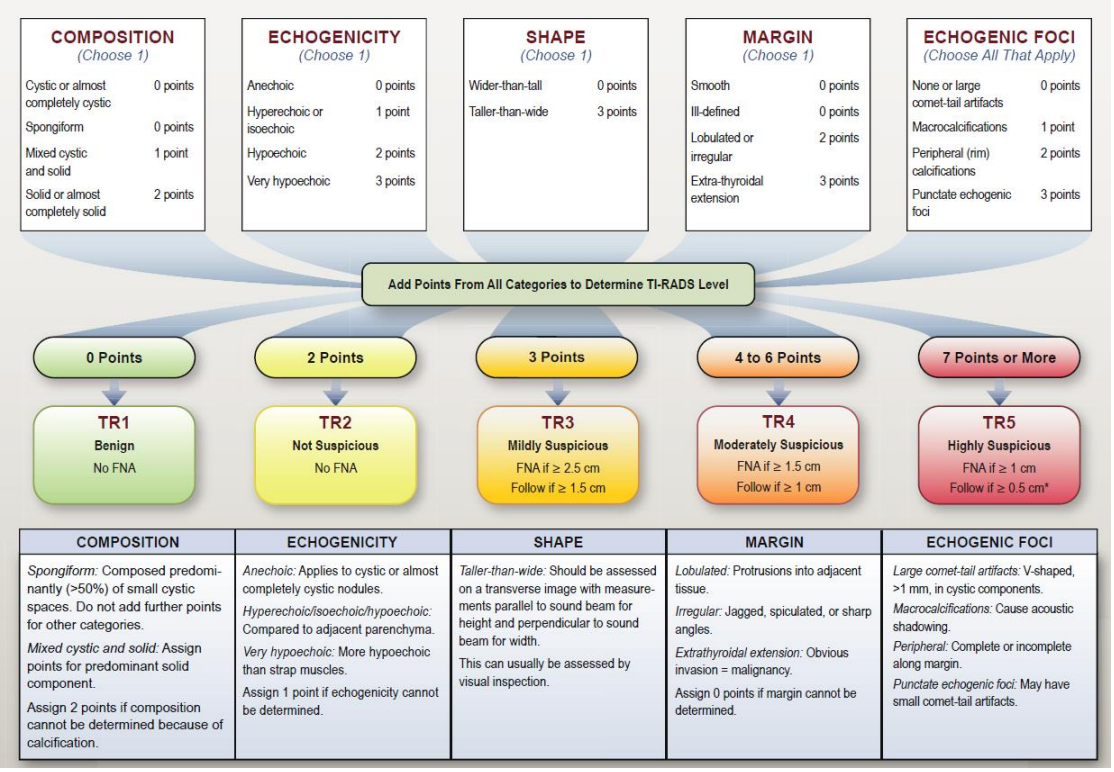

Figure 13 Thyroid Imaging Reporting and Data Systems (Franklin et al, 2017)

The ultrasound features considered in TIRADS include composition, echogenicity, shape, margin and echogenic foci. Each feature has multiple categorical values; for example, composition can be described in 4 categories: cystic, spongiform, mixed cystic and solid. Each category for a specific feature corresponds to a pre-assigned value and the overall score for a nodule is the summation of values among all features. Higher scores correspond to a higher probability of malignancy and an increasing recommendation for a biopsy. The TIRADS features are described by using the in-house developed medical image feature description software MINT v1.4 (Figure 14).

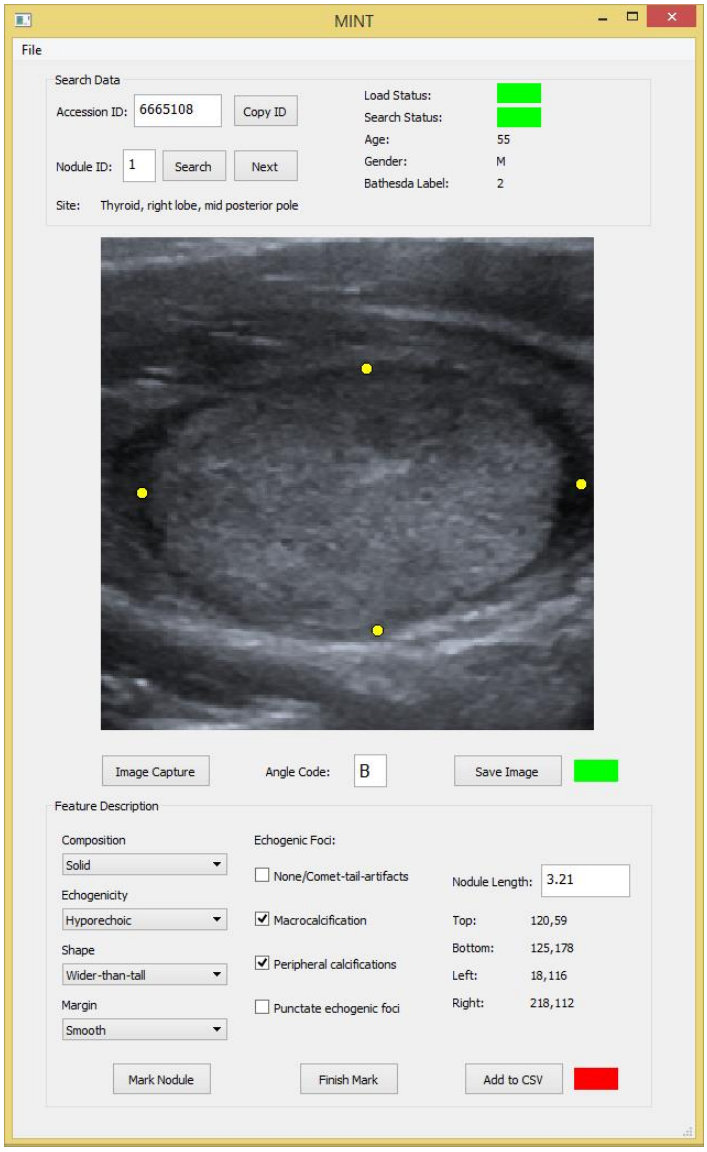

Figure 14 Medical Image Characterization Using MINT (Dong, 2017)

In addition to TIRADS features, the coordinates of top/bottom/left/right boundary of nodule is also recorded for later preprocessing and alignment. Moreover, some demographic information is also recorded, such as age and gender. The detailed statistics of TIRADS characterization and demographic information in this study are displayed in Table 3.

Table 3 Statistics of TIRADS Characterization and Demographic Information in This Study

Retrospective Observations		n=1434
Age		56.3 ± 14.6
Gender	female	1113 (77.6%)
	male	321 (22.7%)
Composition	Cystic	42 (2.9%)
	Spongiform	233 (16.3%)
	Mixed Cystic and solid	362 (25.2%)
	Solid	797 (55.6%)
Echogenicity	Anechoic	18 (1.3%)
	Hyperechoic	338 (23.6%)
	Hypoechoic	1018 (80%)
	very hypoechoic	60 (4.2%)
Shape	Wider-than-tall	1368 (95.4%)
	Taller-than-wide	66 (4.6%)
Margin	Smooth	1201 (83.8%)
	Ill-defined	140 (9.8%)
	Lobulated/irregular	91 (6.4%)
	Extra-thyroidal extension	2 (0.1%)
Comet-tail artifacts	Yes	267 (18.6%)
	No	1167 (81.4%)
Macro calcifications	Yes	79 (5.5%)
	No	1365 (94.5%)
Peripheral calcifications	Yes	30 (2.1%)
	No	1404 (97.9%)
Punctate echogenic foci	Yes	442 (30.8%)
	No	992 (69.2%)
Nodule Length (cm)		1.8 ± 1.0 (0.3-8.4)
Nodule label	Benign	1121 (78.17%)
	Suspicious	313 (21.83%)

There are, in total, 1434 images that have a complete set of characterizations evaluated by radiologists. As shown in Table 3, most characterized nodules (78.17%) are benign, patients are mostly around middle to elder age and are mostly female (77.6%). At last, all data available for the study is summarized in Figure 15 by a Venn diagram.

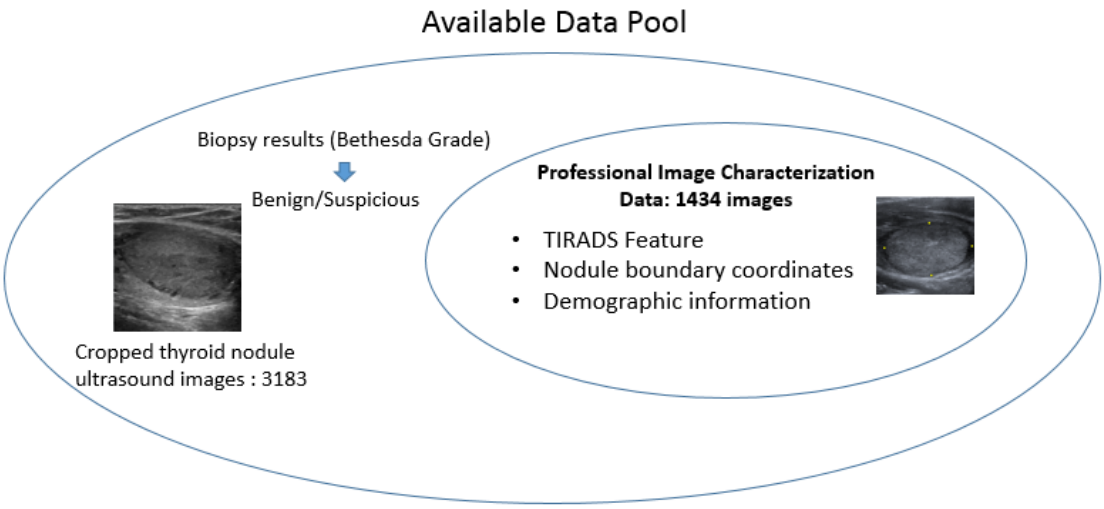

Figure 15 Summary of Available Data for the Study

Chapter 5 Methodology

5.1 Methodology Workflows

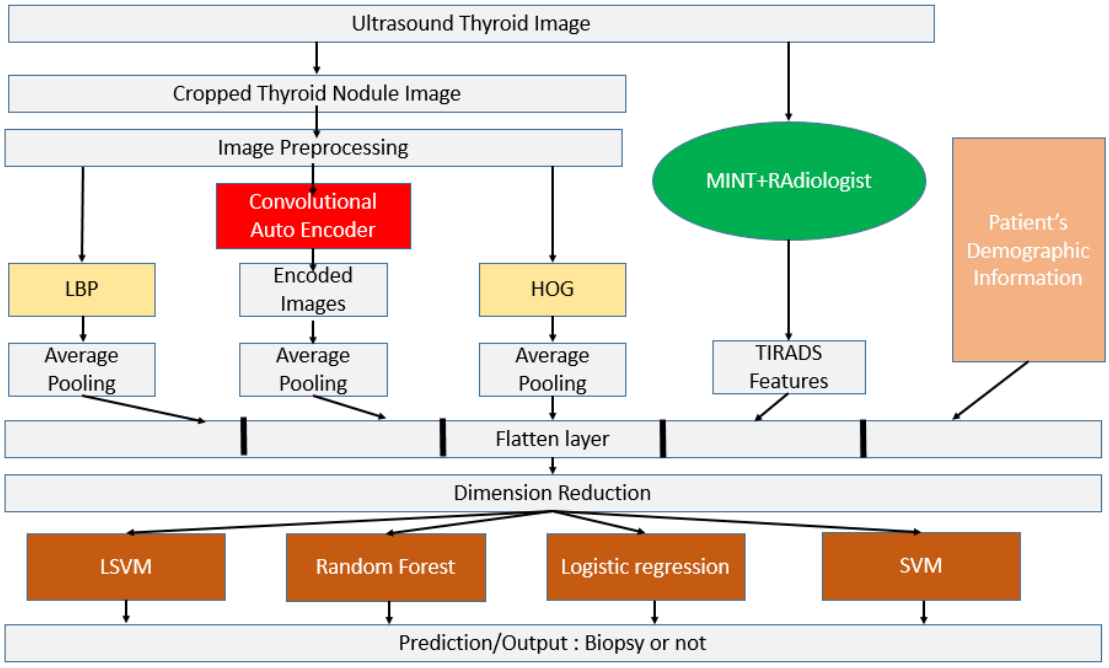

Figure 16 Methodology Workflow in This Study

The overall workflow in this study is described in Figure 16; each component will be discussed in detail later in this chapter. The general idea of this workflow is: gather useful features from multiple sources and build classifier on top of them.

Given the ultrasound thyroid image, radiologists will crop a sub-image that contains a nodule. The cropped nodule image will go through image preprocessing to align all nodule together and resize based on the boundary coordinates provided by MINT.

After preprocessing, I applied two computer vision descriptors: Local Binary Pattern (LBP) and Histogram of Oriented Gradients (HOG) on the nodule image to extract specific information. LBP can provide a texture information of the image and HOG can provide shape and directional patterns in the image. Next, I trained a

Convolutional Auto-Encoder (CAE) to extract useful patterns from the nodule image. The useful pattern is stored at the encoded layer of CAE.

After extracting image features using LBP, HOG and CAE, an additional pooling layer is applied to further reduce dimensionality and preserve the spatial pattern. Next, I flatten the pooled images as vectors and concatenate them together with TIRADS features and patient's demographic features.

Now, the problem becomes a machine learning problem. During the machine learning experiment design, nested cross validation is used to determine the optimal preprocessing methods and the optimal hyper-parameter for four classifiers.

Finally, models with the best validation score for each algorithm are evaluated on a separate test set to calculate the final statistics and scores.

5.2 Data workflow

In order to better understand the methodology, it is helpful to know how the data flow is handled in the study. Please note that the data workflow is not "data flow" in TensorFlow context; instead, it is a graph that specifies how the available data is being used for training, validation and testing. The data workflow contains a combination of unsupervised feature learning and supervised classification, as shown in Figure 17 below.

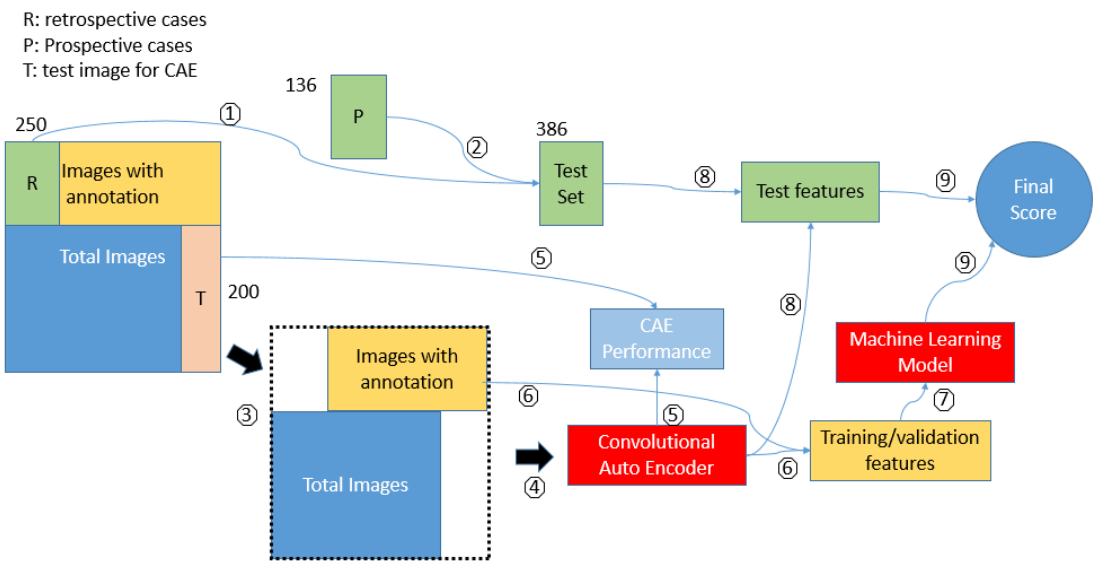

Figure 17 Data Workflow in the Study

There are in total of 9 steps in the data workflow:

(1) 250 cropped nodule images with professional characterizations are held out as retrospective test data.

(2) As the study proceeds, there are 136 new test cases that also joining the test sets, the size of testing data is further expanded to 386.

(3) Hold out 200 images from uncharacterized nodule images as test set for CAE, it is used to test the performance of CAE models.

(4) Use the rest of the images to train CAE with image augmentation.

(5) Test the reconstruction of CAE on test nodule images (T).

(6) Use CAE to extract features for the rest of the characterized nodule images.

(7) Use nested cross validation to build several good performing machine learning models based on the training features.

(8) Once a machine learning model finishes training and best model of each algorithm is selected from validation data, feed all characterized test sets into the CAE to extract image features.

(9) Use those test features to make predictions using the machine learning model trained before, and see the test performance.

5.3 Image Preprocessing

Given the cropped nodule image, a preprocessing is required to align the nodule center of all images and then standardize image sizes.First, the center of each nodule is calculated by the boundary coordinates (Figure 18). Next, the center of all nodule images are aligned together at center (Figure 19).

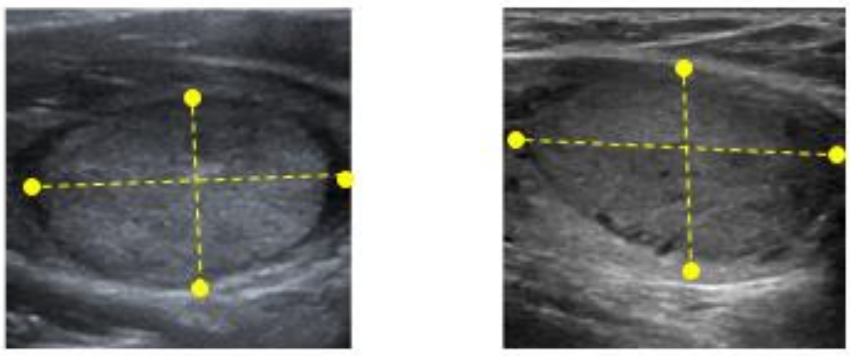

Figure 18 Center is the Intersection of two Boundary Lines

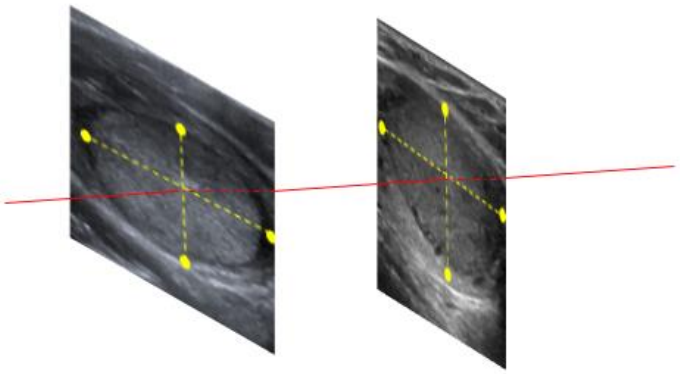

Figure 19 Align Nodule Center for all Images

After the alignment, the size of all images is standardized to 224x224 by stretching with respect to each center. During the stretch, the horizontal and vertical stretch ratio are recorded for image reconstruction later. Finally, the grey value pixel is rescaled between 0 and 1 for the convenience of CAE training.

5.4 Convolutional Auto-Encoder

The concept of Convolutional Auto-encoder (CAE) originates from combining the concept of an Auto-encoder (AE) and Convolutional Neural Networks (CNN) together. Similar to an AE, the essential idea of a CAE is to gradually "compress" the two dimensional images to a lower dimensional representation. This is called the encoding process. Next, CAE will extract the encoded image back to its original image, which is called the decoding process. The objective of this process is to make sure that the reconstructed image is as similar to the original image as possible.

There are two goals that can be achieved by CAE's encoding and decoding process. First it serves as a dimensionality reduction method that can effectively get rid of high frequency noises in the input. Second, and more importantly, by learning an efficient way of compression and extraction, CAE can extract very useful patterns from

the image. Therefore, CAE is not only a dimension reduction method, it is also an effective image feature extractor.

The general model architecture of CAE is described in Figure 20. First, the input images are filtered by a 2-D convolutional layer. Next, the size is reduced by pooling layers. Multiple convolution-polling layers are stacked together until it reaches our expected pattern size. Sometimes multiple convolution layers can be stack together to allow for more non-linearity.

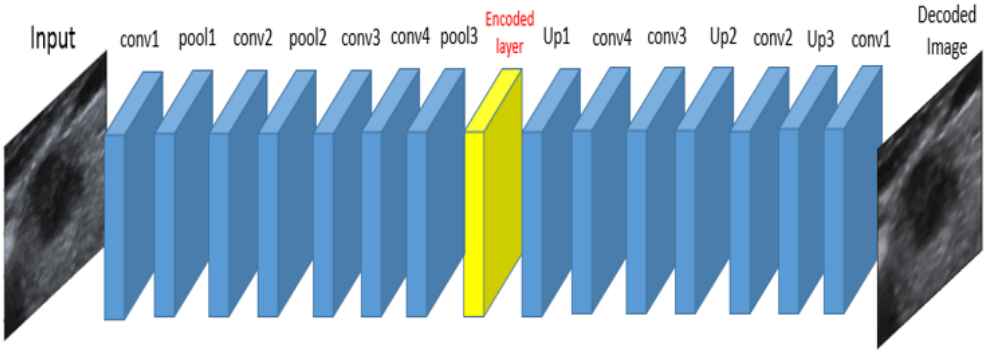

Figure 20 The General Architecture of Convolutional Neural Network

The middle layer, with the lowest number of dimensions is called the encoded layer. Encoded layer is the place that stores the most compact patterns of input images. The rest of the network (decoder) is usually symmetric to the encoding process with a pooling layer now becoming an up-sampling layer. The weights in the decoder can either be the same as encoder (tied-weights) or different from encoder (untied-weights).

The encoded layer is a lower dimensional representation of the input image that not only has the high frequency noise removed, but also contains important patterns about the input image. Subsequent supervised classifiers can use the compact patterns in encoded layer as features, in which case, CAE can be used as a feature extractor.

As it is mentioned in the Chapter 3.4, convolution is not suited for end-to-end task for medical image recognition. Therefore, in this study, CAE is used as feature extractor instead. CAE can be easily constructed using the TensorFlow library.

5.5 Image Augmentation

It is stated in Chapter 3.3 that the medical data is usually not enough to train most deep learning models. In order to mitigate the negative effect caused by lack of data, image augmentation methods can be used. Image augmentation is a way of artificially creating new training images to compensate the lack of data. Some simple image augmentation are achieved by rotation, shearing, shifting or cropping of initial images. Image Augmentation provides an efficient way to expand the number of training data. Moreover, since the convolution process is rotation sensitive, image augmentation offers a way of helping the CNN recognize the rotation of images.

Before Image augmentation, it is important to make sure that the augmentation operation would not destroy any useful patterns in the original image. For example, in nodule image diagnosis, the width and height provide very valuable information on diagnosis, if rotating the image, the width and height information is no longer valid.

In this study, the image augmentation specifications are shown in table 4. All grey-scale images are rescaled between [0, 1]. Then width and height are randomly shifted up to 20% of entire size, if images are shifted outside of size limit, then fill in constant 0 in those pixel. During the shifting, there is a random zooming effect up to 25% applied to the tumor image as well. At last, the horizontal axis of image may randomly flip.

Table 4 Image Augmentation Specification Used in This Study

Image Augmentation Parameter	Value
rescale	1/255
width shift range	20%
height shift range	20%
zoom range	25%
horizontal flip	TRUE
fill mode	constant,0

5.6 Average Pooling

Pooling layer is extensively used in almost all convolution neural networks. It is often placed after a convolution layer to reduce the dimensionality. Moreover, a pooling layer provides robustness to the model and is the primary reason why Convolutional Neural Network (CNN) can be insensitive to position shift [52, 53].

The most common pooling methods are max pooling and average pooling. Pooling partitions the input image into a set of sub-regions such that for each sub-region, it outputs the max or average value of the sub region (Figure 21). The pooling layer serves to progressively reduce the size of image, as well as reduce the number of training parameters.

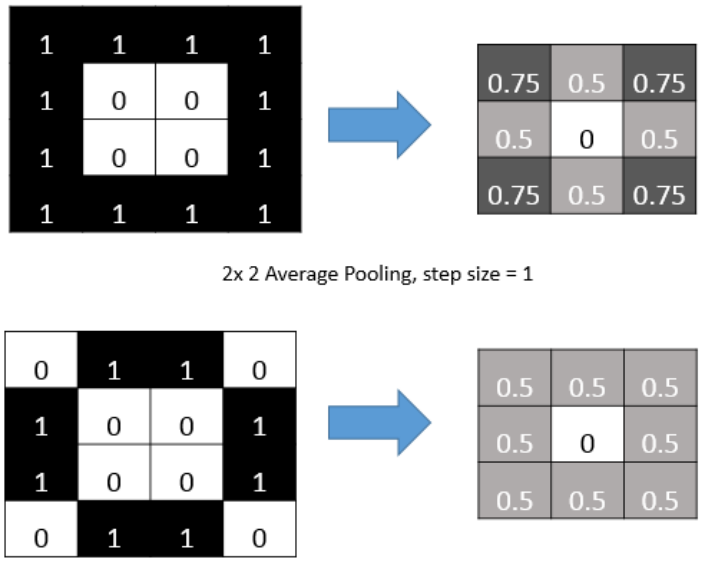

Figure 21 Illustration of Average Pooling

The pooling layer provides invariance to modern CNN, it allows CNN to correctly recognize similar objects with differences. For instance, in Figure 21, although the original top and bottom images are not the same, they all have circular pattern. Applying pooling is able to extract their circular pattern and as a result, their pattern after pooling is almost the same. This indicates that pooling can effectively increase the robustness in recognizing similar patterns.

5.7 Local Binary Patterns

Local Binary Patterns (LBP) [8] are a simple, popular, yet efficient visual descriptor in computer vision. It is mainly used to describe the texture of images. Figure 22 shows how LBP is calculated in an arbitrary pixel, with a neighborhood of 8 points on a circle of radius 1.

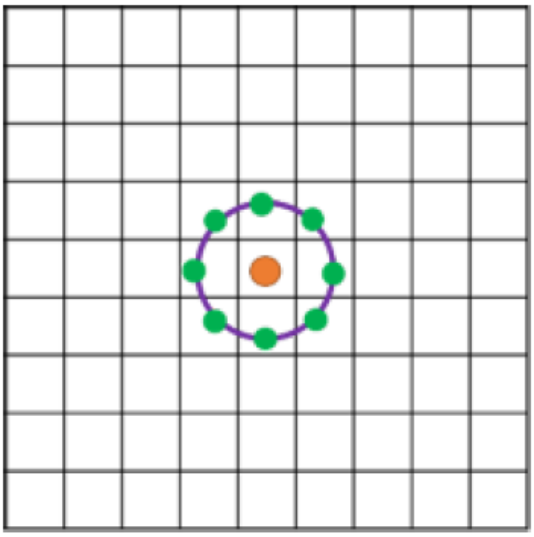

Figure 22 Demonstration of Local Binary Pattern Calculation of Arbitrary Pixel

The LBP code of a pixel value g_c located at (x_c,y_c) is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \quad (1)$$

Where P is the total number of points in the neighborhood of radius R and s is a unit step function defined by:

$$s(x) = 1 \; if \; x \geq 0$$
$$s(x) = 0 \; if \; x < 0 \quad (2)$$

The input and output of an LBP descriptor usually has the same dimension. The output of each pixel represents the LBP code of the original image. Figure 23 displays an example of an LBP descriptor using a thyroid nodule image. A histogram can also be obtained from the LBP code to statistically represent texture information of image and LBP features are described as both LBP images and histogram counts. In this study, the LBP is created through the skimage package in python.

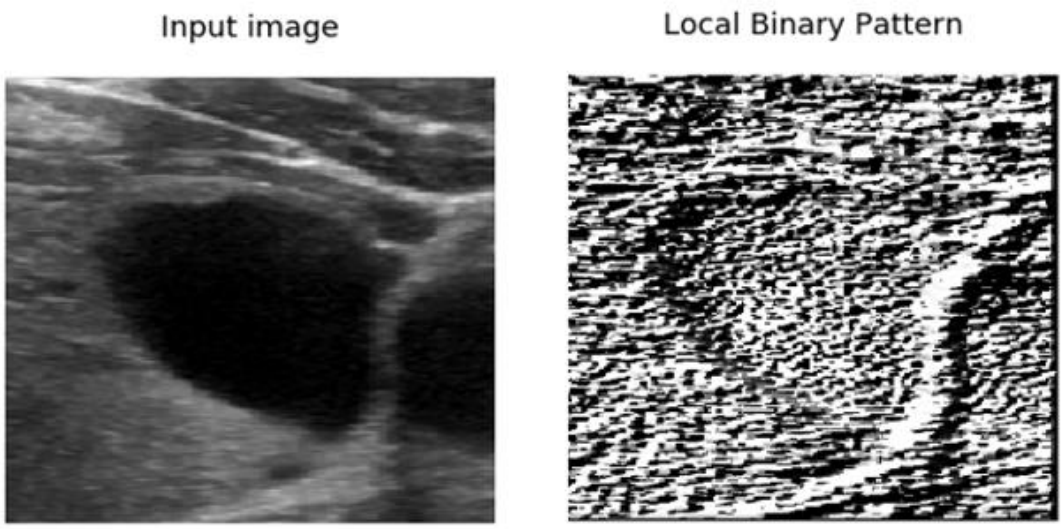

Figure 23 Local Binary Pattern of one Nodule Image

5.8 Histogram of Oriented Gradients

Histogram of Oriented Gradients (HOG) is designed to capture the appearance and shape of a local object. [7] The directional pattern in the image can be depicted by the histogram of local intensity gradients. Given the input image, a centered filtered mask computes the magnitude and orientation of the gradients. The magnitude and gradients can capture contour and some texture information. Then, the image window is divided into small spatial cells, which quantize the gradients into a local 1-D histogram of gradients over all the pixels in the cell (Figure 24). The histogram divides the gradient angle range into a fixed number of bins and the magnitude of gradient is used to vote into the orientation histogram.

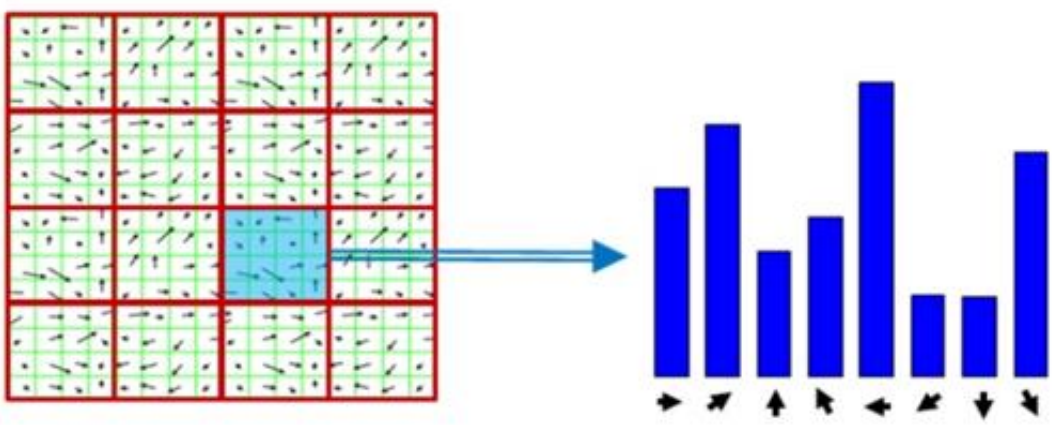

Figure 24 Creation of Cell Gradient Histogram

The HOG image of a nodule image is shown below in Figure 25. It is capturing the shape of the nodule, as well as some directional patterns outside of the boarders. The histogram count of each direction within a pooling layer can be used as a feature for later machine learning. The HOG image is generated by the skimage package.

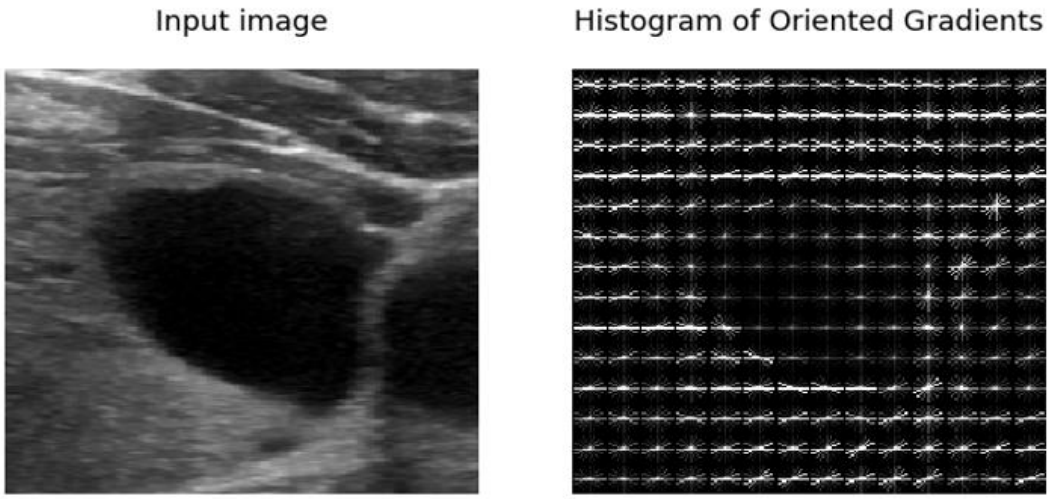

Figure 25 Histogram of Oriented Gradients of Nodule Image

5.9 Model Selection Score Function

As it was mentioned in the section 3.1, the score function needs to be carefully determined so that it can correctly reflect the best trade-off between false positive and

false negative. The false negative and false positives are components of the confusion matrix (Figure 26).

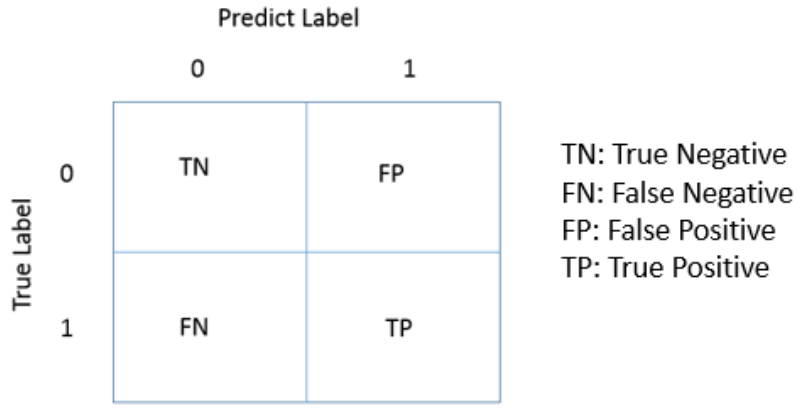

Figure 26 Confusion Matrix

Several important metrics in the study are accuracy, false discovery rate (FDR), false negative rate (FNR), sensitivity and precision. Each one of them is calculated from the confusion matrix with the following expression:

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FP+FN} \tag{3}$$

$$\text{FDR} = \frac{FP}{FP+TP} \tag{4}$$

$$\text{FNR} = \frac{FN}{FN+TP} \tag{5}$$

$$\text{Sensitivity} = \frac{TP}{FN+TP} = 1 - \text{FNR} \tag{6}$$

$$\text{Precision} = \frac{TP}{FP+TP} = 1 - \text{FDR} \tag{7}$$

In the context of this study, we want the FDR to be low because the final algorithm should reduce unnecessary biopsy cases, and at the same time, we want the FNR to be low because it is also important to reduce the number of missing malignancy

cases. At last, accuracy is also an important metric to give us an overall performance measure.

The convention of score function is that a higher score indicates a better performance. Therefore, minimizing FDR and FNR means maximizing sensitivity and precision (from equation 6 and 7). A final score function is determined by radiologists teams by weighing both economical cost and clinical risks caused by FDR and FNR.

$$\text{Final Score} = \text{Accracy} + 1.85\ \text{Precision} + 4.73\ \text{Sensitivity} \quad (8)$$

It is known that some terms in equation (8) are repetitive to some degree. For example, both precision and sensitivity are using the same TP in the calculation and the use of accuracy involves all terms used to calculate precision and sensitivity. However, despite repetitiveness of information, the equation provides a nice and easy framework for radiologists to weigh each metric and is therefore used as score function in the study. As we can notice in equation (8), reducing false negative is more than twice important than reducing false positive.

Chapter 6 Experimental Results and Analysis

This chapter will discuss the implementation, training, validation and testing details of Convolutional Auto-Encoder and multiple supervised classification models such as random forest, logistic regression, linear support vector machines and non-linear support vector machines.

6.1 Convolutional Auto-Encoder Experiments

After image preprocessing, the CAE model with the architecture shown in Figure 27 is constructed to extract patterns.

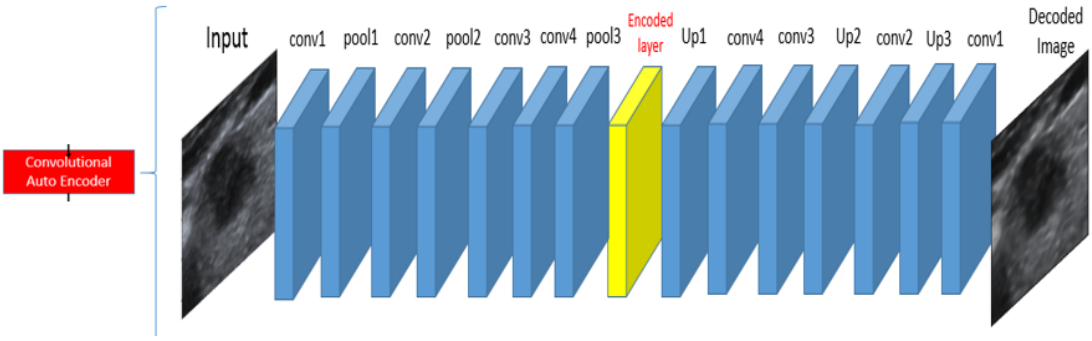

Figure 27 Convolutional Auto-Encoder Model used in This Study

The encoding and decoding structure are entirely symmetric with respect to the encoded layer at the center. There are three convolution-pooling layer combinations on each side. Experiments showed that the use of two consecutive convolutional layers near the middle yields a better reconstruction of the original input. The specification of the CAE model for encoding layer is specified in table 5. Since the model is symmetric, the decoding specification is the same as encoding with each layer reversed.

Table 5 CAE Specification in the Study

Layer	Width after layer	Height after layer	#filter after layer	kernel size/pool size	strides
input	224	224	1	NA	1
conv1	224	224	64	(3,3)	1
pool1	112	112	64	(2,2)	1
conv2	112	112	32	(3,3)	1
pool2	56	56	32	(2,2)	1
conv3	56	56	16	(3,3)	1
conv4	56	56	8	(3,3)	1
pool3	28	28	8	(2,2)	1
Encoded	28	28	8	NA	1

During the training, I used image augmentation methods mentioned in Chapter 5.5 to create different replication of training data for each epoch. The objective function is binary cross entropy, which has proved to be a very effective objective function for image data. The optimizer used to train the CAE is RMSprop optimizer, which is an unpublished, adaptive learning rate method proposed by Hinton in a course [54]. The training configuration of CAE in this study is listed in table 6.

Table 6 Training Specification of CAE

Optimizer	rmsprop
Momentum	0.9
Initial Learning Rate	0.01
End learning rate	0.0001
number of epoch per decay	20
learning rate decacy factor	0.94
learning rate decay type	exponential
Batch Size	32
Maximum Epoch	2000
weight decay	0.00005

The training data is specified in chapter 5.2. After training, I used the test set to visualize the reconstruction. Figure 28 shows the comparison between the original image and reconstructed image for two nodules.

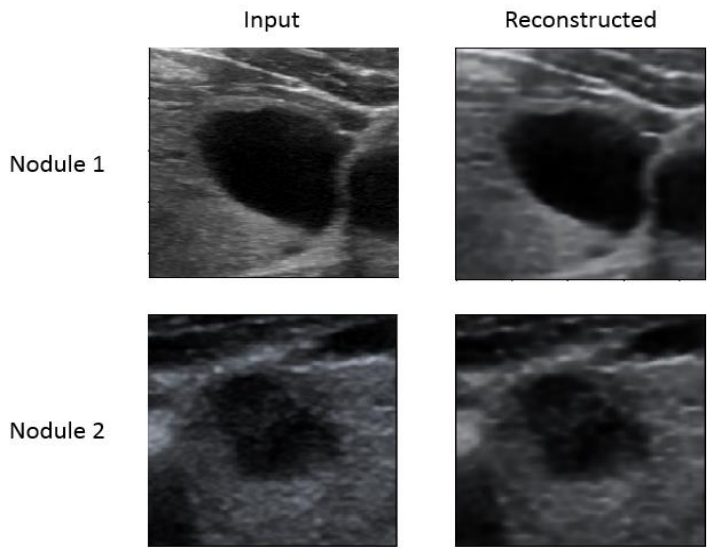

Figure 28 Input Image V.S. Reconstruction Test Image by CAE

There is a high degree of similarity between the input image and the reconstructed image, which ensures that our encoded layer, as a low dimensional representation that contains almost the same degree of information as the original image. Now the next question is: can encoding and decoding process extract useful patterns?

Fortunately, the encoded layers can be visualized for examination. The encoded feature for nodule 1 is shown in Figure 29 below.

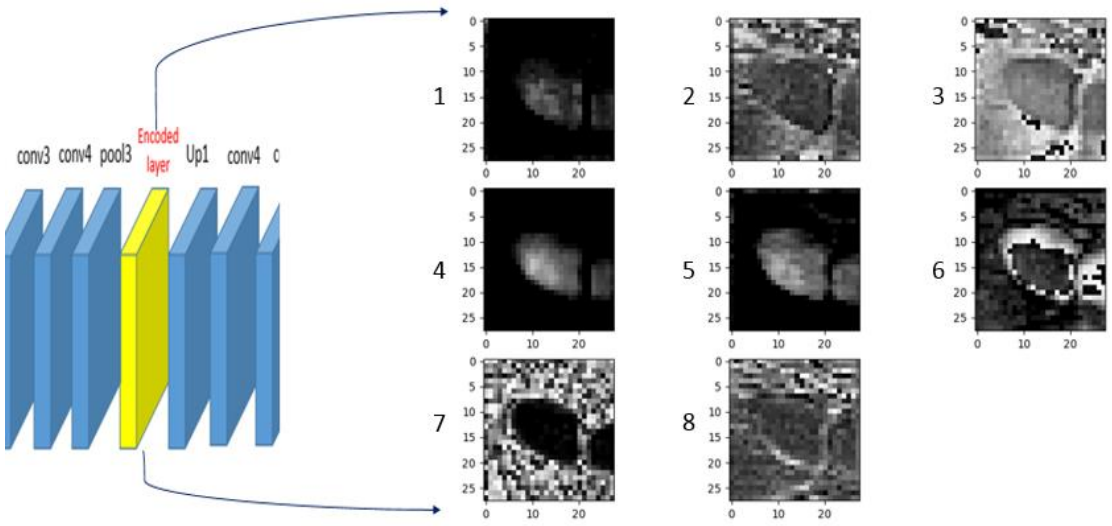

Figure 29 The Encoded layer of CAE for Test Nodule#1

In the encoded layer, all pixel values are between zero and one, with zero being entirely dark (no signal) and one being entirely bright (full signal). The brightness of pixel represents the corresponding significance in each encoded grey scale image. As we can see from Figure 28, there are in total of 8 parallel images at encoded layer, where each encoded image is focusing on different pattern. For example, images 6 focuses on the boarder of nodule. Image 1, 4 and 5 pay attention to the inner area. Image 2, 3, 7 and 8 look at surrounding tissues. The patterns in the encoded layer provide evidence to statement made earlier: CAE is not only a dimension reduction method, it is also an image feature extractor.

An average pooling layer with size 4x4 and stride 4 is applied after the CAE to extract pattern information while preserving the spatial pattern. After the average pooling, the feature is flattened and concatenated with features from LBP, HOG, TIRADS and demographic information. The next step is to use these features for supervised classification.

6.2 Nodule Diagnosis Classification

6.2.1 Design of Experiments

Now the problem becomes a binary classification: deciding whether a nodule is benign or suspicious. Note that the target is not benign/malignant because our target is derived from Bethesda Grade, which comes from biopsy tests. Although biopsy test can provide a very accurate answer, it still cannot confidently determine the malignancy of a nodule. Therefore, all nodules other than benign are lumped together as suspicious nodules.

The feature now has 1381 dimensions and the detail of these features is listed in Table 7. There are 8 parallel images at CAE's encoded layer, each encoded image has dimension 28x28. After a 4x4 non-overlapping average pooling, the dimension of each encoded image now becomes 7x7.

The total number of features from CAE is 7x7x8=392. The LBP and HOG features are calculated in similar manner. LBP Histogram pooling is calculated from a 10-bin histogram for each pooling region of LBP data. HOG histogram is calculated in a similar manner using 9 bins for angle.

Table 7 Size of Features for Each Component

Feature	Calculation	# Dimension
CAE Encoded Image Pooling	7x7x8	392
LBP Image Pooling	7x7	49
LBP Historgram Pooling	7x7x10 (LBP value)	490
HOG Histogram Pooling	7x7x9(angle)	441
TIRADS Features		7
Demographic Information		2
Total		1381

The training sample size is 1184 and it is the number of characterized images with 250 retrospective test cases being held out. Since the 1381 features are gathered from multiple data sources, there are plenty of features with redundant information and many features may have limited information in diagnosing a nodule. Therefore, in order to proceed, it is necessary to apply feature selection and feature reduction methods.

In this study, the essential idea of building a supervised learning method is to create an overall pipeline that includes preprocessing, feature selection, dimension reduction and model selection together (Figure 30). Then I use two nested10-fold cross validation (CV) to do parameter optimization. The outer CV uses a random search to find the best preprocessing workflow and the inner CV uses a grid search for model selection.

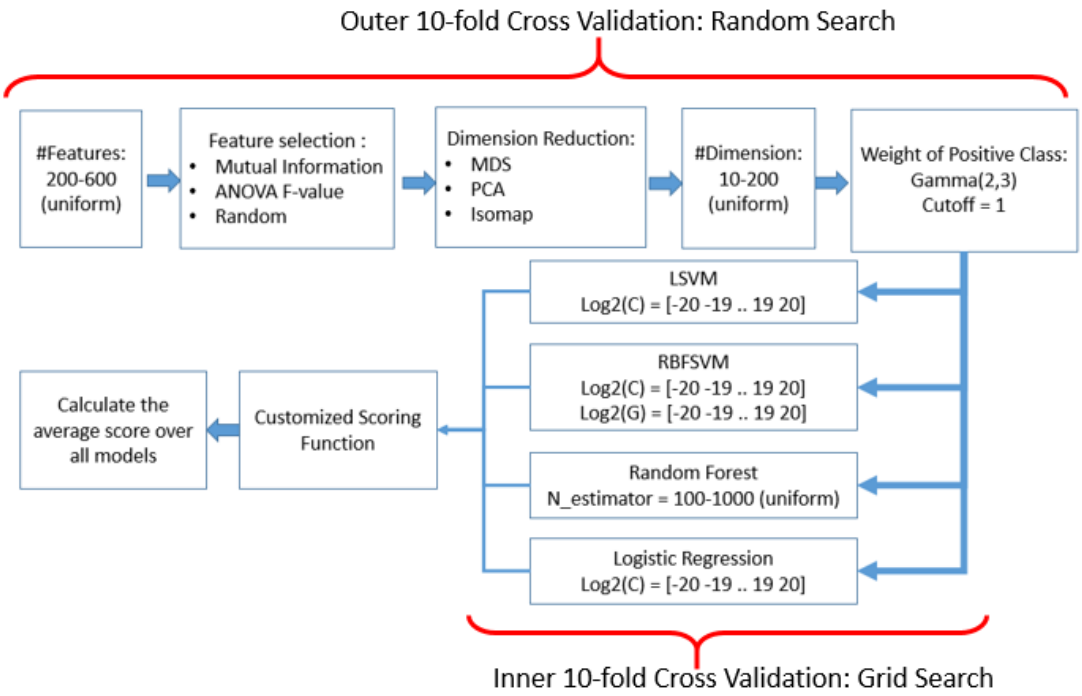

Figure 30 Nested Cross Validation Modeling Scheme

The cross validation on the outer layer is used to decide the best suite of preprocessing parameters that gives us the highest model selection scores. The outer CV contains the following component:

(1) The number of features is drawn from a uniform distribution from [200,600].

(2) Select a feature selection scheme, from Mutual Information (MI), F-Value and Random Selection with each method having equal probability. If using random selection, the feature index is recorded.

(3) Select dimension reduction method from MDS, PCA and Isomap (each method has equal probability).

(4) The number of reduced dimensions is further reduced to a smaller number, which is drawn from a uniform distribution within [10,200].

(5) Randomly generate the weight for the positive class from gamma (2, 3) distribution, with probability distribution function shown in blue in Figure 31. If the weight happens to be less than 1, then manually assign 1 as a weight, because positive class should not be less important than negative class.

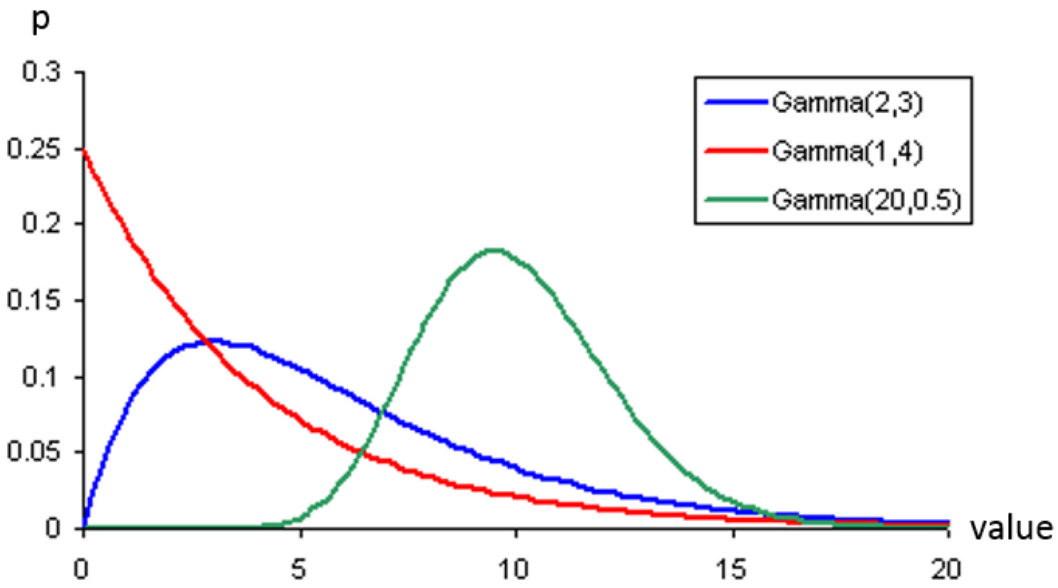

Figure 31 Gamma Probability Distribution in the Study

There are infinite number of preprocessing configurations because many preprocessing parameters are drawn from a random distribution. In order to fully explore the parameter space with finite computational resource, random search optimization is used in outer CV to find out the best preprocessing workflow.

The random search in this study is inspired by Bergstra and Bengio's paper [55]. In their study, they demonstrated that when the number of hyper-parameters is large, mostly only a small fraction of parameters are indeed important to final performance. Using a grid search is computationally expensive and may not be able to fully explore the space of important parameters. In Figure 32, vertical axis means the space of unimportant parameter and horizontal axis represents the space of important parameter.

If using 9-point grid search, only 3 values of important parameter are explored and it is less likely to find out the sweet spot of important parameters. Random search, on the other hand, can more effectively explore the space of important parameters and can therefore more likely to find the performance sweet spot.

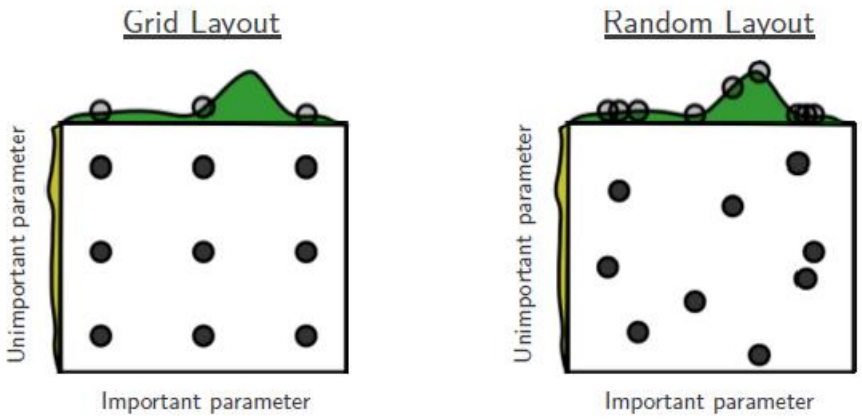

Figure 32 Grid Search V.S. Random Search (Bergstra and Bengio, 2012)

Figure 33 shows the nested cross validation scheme. There are mainly four steps involved in nested cross validation:

1. Randomly generate 1000 data preprocessing schemes.
2. Divide the training the set to 10 outer folds for each preprocessing scheme.
3. Use a grid search to determine a model hyper-parameter configuration for each inner fold (for example, c and g in RBF SVM).
4. Further divide the rest of training data into inner 10 folds. The score for each fold is calculated by equation (8). If a null model is encountered (such as one that predicts all positive), then the score is assigned as 0.

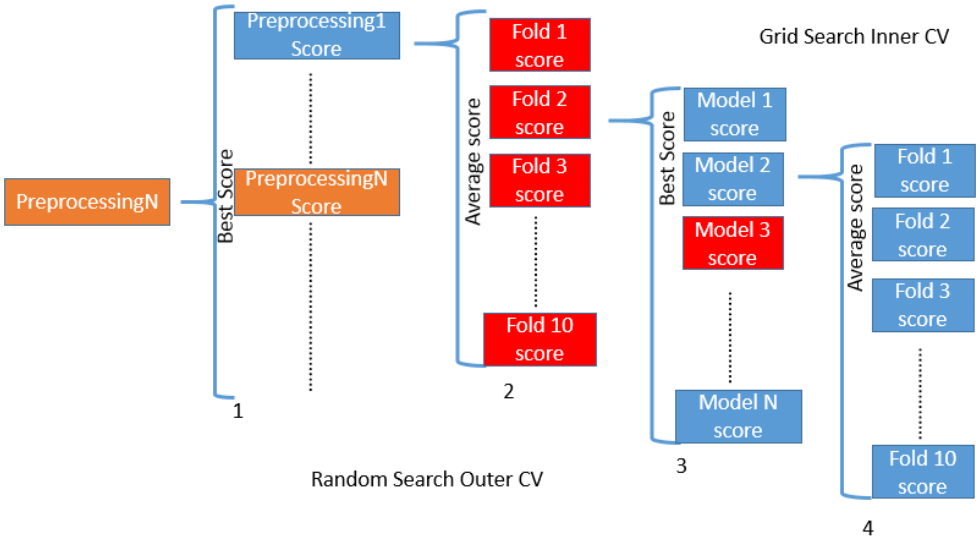

Figure 33 Nested Cross Validation Scheme in the Study

The score in step 1 and 3 is calculated by averaging scores from multiple folds in step 2 and 4. The optimal preprocessing scheme and the optimal hyper-parameter of each classifier is chosen by picking the best average score. Once the best preprocessing scheme for each classifier is determined, I retrained the inner grid search using the best preprocessing configuration for each classifier on entire training data.

For each algorithm, the whole nested cross validation takes about 2 days using a 20-core node on a supercomputer. After finding the best preprocessing scheme and best model for each supervised classifier, I applied them on the test set and compare the test performance.

6.2.2 Experimental Results

The best preprocessing method and best parameter for each algorithm is listed on Table 8. Their corresponding performance on the test set is listed on Table 9. The test results for each classifier can provide us with information about the problem itself.

Table 8 Best Model for Each Algorithm

Classifier	Outer CV					Inner CV
	Feature Selection	# Features	Dimension Reduction	# Dimension	weight	best parameter
Random Forest	ANOVA F-Value	576	MDS	195	2.21	n = 700
Logistic Regression	MI	244	MDS	37	3.73	log2(c) =2.18
Linear SVM	MI	396	MDS	104	4.33	log2(c) =4.46
RBF SVM	MI	435	PCA	198	8.92	log2(c) =13.84, log2(g) = 4.83

Table 9 Test Performance for Each Algorithm

Classifier	Performance Metrics			Final Score
	Accuracy	Precision	Sensitivity	
Random Forest	0.666	0.323	0.187	2.15
Logistic Regression	0.575	0.316	0.458	3.33
Linear SVM	0.544	0.305	0.505	3.50
RBF SVM	0.417	0.308	0.888	**5.19**
Human Performance	0.262	0.254	0.860	4.80

The Random forest produces the highest accuracy. However, since the dataset itself is imbalanced, accuracy can be misleading because it tends to predict the dominant class more likely than the minor class, which leads to a very low sensitivity and therefore a very high false negative rate. Therefore, the random forest yields the lowest overall score based on our scoring function.

The performance of Logistic Regression and Linear SVM is very similar, even though Logistic Regression has slightly better accuracy and Linear SVM has better recall. These two are both linear models and we can infer from their performance that the real decision boundary is highly non-linear under their best preprocessing scheme.

Support Vector Machines (SVM) with radial basis function kernel (RBF) surprisingly produces the highest score (even higher than human performance).

Although the accuracy is not the best, it is able to minimize the false negative the most, which produces the highest sensitivity and therefore the highest score.

Human performance has lowest accuracy and highest sensitivity, which is not a surprise. In order to reduce false negatives, radiologists tend to recommend biopsy for most thyroid nodules. Comparing the SVM with human performance, we can see that the SVM model with the RBF kernel has higher sensitivity and higher precision, which means that the SVM model can both reduce unnecessary biopsy cases and reduce the number of missing malignancy cases.

6.2.3 Hypothesis Test

In order to ensure the consistency of experimental results, a hypothesis test is conducted on the final SVM model. When the optimal training configuration is fixed, I divide the entire training data into 10 folds and train 10 different SVM models with each model using 9 folds. At last, I evaluate each model against the test dataset. The test score for each model is shown in Table 10. I assume the human score is constantly 4.80 on the test set.

Table 10 Test Score of Different Final Models

Model	Test Score
SVM1	4.7965
SVM2	4.8523
SVM3	5.0063
SVM4	4.9761
SVM5	5.0675
SVM6	5.1948
SVM7	4.9693
SVM8	5.1358
SVM9	5.0219
SVM10	5.0314

The null hypothesis H_0 is: The performance of final SVM models has an average score that is no greater than human performance score.

Since the final SVM model results from multiple comparisons, thus the hypothesis test significance level α needs to be corrected to account for multiple comparison. In this study, I correct α using Bonferroni correction [56] by equation (9).

$$\alpha' = \frac{\alpha}{N} \tag{9}$$

$$\alpha = 0.05 \tag{10}$$

$$N = 3 \tag{11}$$

Where α' is the corrected significance level and α is 4the original significance level, which is assumed to be 0.05 (equation 10). N is the total number of comparisons, there are 3 comparisons in finding the classification algorithm with maximal score. The corrected hypothesis significance level α' then becomes 0.0167.

Given the performance scores in Table 10, unpaired bootstrap sampling with randomization is used conduct the one-sided hypothesis test (right-side test). The number of bootstrap resampling is set at 10 million.

The final p value for the null hypothesis H_0 is 0.00016, which is less than the corrected significance level 0.0167. Therefore, I can reject the null hypothesis and claim that the final SVM models has an average score greater than human performance score. The higher score means that the SVM model can provide better tumor diagnosis considering both economical cost and clinical risk. Therefore, the final model can be clinically helpful to radiologists in determining whether a nodule should undergo biopsy or not.

Conclusion

This study introduced how machine learning can be used in biomedical imaging to help radiologists improve cancer diagnosis. It also proposed a new framework for extracting useful information for thyroid ultrasound images. There are several key points in this thesis:

(1) End-to-end Convolutional Neural Networks have limited applicability in thyroid nodule diagnosis due to lack of data and lack of spatial arrangements in ultrasound nodule images.

(2) Medical Image Augmentation can be a primary solution to the lack of data problems in medical AI.

(3) Convolutional Auto-Encoder is not only a dimension reduction technique, but also an effective feature extractor for medical image.

(4) The feature extraction workflow produces features for multiple classifiers to diagnose thyroid tumor. The best performing model is RBF SVM, which is trained by nested cross validation and it outperforms human experts on the test set.

Future Work

The rise of deep learning and AI has the potential to improve and reform future medical diagnosis system. This thesis has only explored a tiny corner of a much bigger picture. Even in the field of thyroid tumor diagnosis, there are many promising future directions that are worth exploring.

First, future studies can try to predict some expert-determined features. If a machine can characterize a thyroid tumor image on a professional level, it can greatly lower the cost of medical data acquisition process. Moreover, having such system can help automate the thyroid nodule diagnosis workflow, which can significantly save labor costs and make high quality medical diagnosis more affordable.

Second, one can also study the importance of different features and transform them into medical knowledge and improve current thyroid diagnosis system. Studying the importance of different features can also help us eliminate irrelevant information in the input and build better performing models.

Last but not least, there are many new deep learning model architectures that are worth exploring. Future researchers are encouraged to design/try new model architectures for medical images to overcome limitations of current models. For example, CNNs cannot handle the recognition of the same objects from different viewpoints, which makes CNN bad at recognizing ultrasound images because most ultrasound images are captured from different angles. Capsule Networks, on the other hand, can be a promising model for ultrasound image because study has shown that the capsule architecture is capable of recognizing equivalence of images under different viewpoints [57].

References

[1] Richard S (30 September 2010). Computer Vision: Algorithms and Applications. Springer Science & Business Media. pp. 10–16. ISBN 978-1-84882-935-0.

[2] Papert, S (1966-07-01). "The Summer Vision Project". MIT AI Memos (1959 - 2004). Retrieved 2 August 2016.

[3] Margaret AB (2006). Mind as Machine: A History of Cognitive Science. Clarendon Press. p. 781. ISBN 978-0-19-954316-8.

[4] Fukushima,K. (1980), Neocognitron: A self-organizing neural network model for a mechanism pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4),193-202

[5]Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. Nature, 323, 533--536.

[6] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2323. DOI: 10.1109/5.726791

[7] Cortes, C., Vapnik, V. (1995). "Support-vector networks". Machine Learning. 20 (3): 273–297.

[8] T. Ojala, M. Pietikainen, and T. Maenpa (2002). "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." IEEE Transactions on pattern analysis and machine intelligence, vol 24, pp: 971-987.

[9] N. Dalal, and B. Triggs (2005). "Histograms of oriented gradients for human detection." Computer Vision and Pattern Recognition. CVPR 2005. IEEE Computer Society Conference on. vol. 1. IEEE, 2005.

[10] Robert M Haralick; K Shanmugam; Its'hak Dinstein (1973). "Textural Features for Image Classification". IEEE Transactions on Systems, Man, and Cybernetics. SMC-3 (6): 610–621.

[11] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115, 211-252.

[12] Alex K & Ilya S & E. Hinton G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. . 10.1145/3065386.

[13] Hu, J & Shen L & Sun G. (2017). Squeeze-and-Excitation Networks. arXiv preprint arXiv:1709.01507

[14] Ferrucci D, Brown E, Chu-Carroll J et al (2010). Building Watson: An overview of the DeepQA project. AI magazine;31(3):59-79.

[15] IBM Watson for Oncology. IBM. http://www.ibm.com/smarterplanet/us/en/ibmwatson/watson-oncology.html, 2016. Last Accessed on May 2018

[16] DeepMind Health. Google DeepMind. https://www.deepmind.com/health, 2016. Last Accessed on May 2018

[17] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA.;316(22):2402–2410. doi:10.1001/jama.2016.17216

[18] Tensor flow ;Abadi, M et al(2015). "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." CoRR abs/1603.04467

[19] Plis SM, Hjelm DR, Salakhutdinov R et al. Deep learning for neuroimaging: a validation study. Frontiers in neuroscience 2014;8.

[20] Hua K-L, Hsu C-H, Hidayati SC et al. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets and therapy 2015;8.

[21] Suk H-I, Shen D. Deep learning-based feature representation for AD/MCI classification. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, 2013, 583-90.

[22] Roth HR, Lu L, Liu J et al. Improving Computer-aided Detection using Convolutional Neural Networks and Random View Aggregation. arXiv preprint arXiv:1505.03046 2015.

[23] Roth HR, Yao J, Lu L et al. Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications. Recent Advances in Computational Methods and Clinical Applications for Spine Imaging. Springer, 2015, 3-12.

[24] Li Q, Cai W, Wang X et al. Medical image classification with convolutional neural network. In: Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on. 2014. p. 844-8. IEEE.

[25] Cireşan DC, Giusti A, Gambardella LM et al. Mitosis detection in breast cancer histology images with deep neural networks. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, 2013, 411-8.

[26] Ypsilantis P-P, Siddique M, Sohn H-M et al. Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. PloS one 2015;10(9):e0137036.

[27] Zeng T, Li R, Mukkamala R et al. Deep convolutional neural networks for annotating gene expression patterns in the mouse brain. BMC bioinformatics 2015;16(1):1-10.

[28] Cruz-Roa AA, Ovalle JEA, Madabhushi A et al. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, 2013, 403-10.

[29] Bar Y, Diamant I, Wolf L et al. Deep learning with non-medical training used for chest pathology identification. In: SPIE Medical Imaging. 2015. p. 94140V-V-7. International Society for Optics and Photonics.

[30] Li Q, Feng B, Xie L et al. A Cross-modality Learning Approach for Vessel Segmentation in Retinal Images. IEEE Transactions on Medical Imaging 2015;35(1):109 - 8.

[31] Ning F, Delhomme D, LeCun Y et al. Toward automatic phenotyping of developing embryos from videos. Image Processing, IEEE Transactions on 2005;14(9):1360-71.

[32] Turaga SC, Murray JF, Jain V et al. Convolutional networks can learn to generate affinity graphs for image segmentation. Neural Computation 2010;22(2):511-38.

[33] Helmstaedter M, Briggman KL, Turaga SC et al. Connectomic reconstruction of the inner plexiform layer in the mouse retina. Nature 2013;500(7461):168-74.

[34] Ciresan D, Giusti A, Gambardella LM et al. Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in neural information processing systems. 2012. p. 2843-51.

[35] Prasoon A, Petersen K, Igel C et al. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, 2013, 246-53.

[36] Havaei M, Davy A, Warde-Farley D et al. Brain Tumor Segmentation with Deep Neural Networks. arXiv preprint arXiv:1505.03540 2015.

[37] Roth HR, Lu L, Farag A et al. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer, 2015, 556-64.

[38] Stollenga MF, Byeon W, Liwicki M et al. Parallel Multi-Dimensional LSTM, With Application to Fast Biomedical Volumetric Image Segmentation. arXiv preprint arXiv:1506.07452 2015.

[39] Xu J, Xiang L, Liu Q et al. Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology images. IEEE Transactions on Medical Imaging 2015;35(1):119 - 30.
[40] Chen CL, Mahjoubfar A, Tai L-C et al. Deep Learning in Label-free Cell Classification. Scientific reports 2016;6.

[41] Cho J, Lee K, Shin E et al. Medical Image Deep Learning with Hospital PACS Dataset. arXiv preprint arXiv:1511.06348 2015.

[42] Lee S, Choi M, Choi H-s et al. FingerNet: Deep learning-based robust finger joint detection from radiographs. In: Biomedical Circuits and Systems Conference (BioCAS), 2015 IEEE. 2015. p. 1-4. IEEE.

[43] Roth HR, Lee CT, Shin H-C et al. Anatomy-specific classification of medical images using deep convolutional nets. arXiv preprint arXiv:1504.04003 2015.

[44] Roth HR, Lu L, Seff A et al. A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, 2014, 520-7.

[45] Gerven MAV, De Lange FP, Heskes T. Neural decoding with hierarchical generative models. Neural Computation 2010;22(12):3127-42.

[46] Koyamada S, Shikauchi Y, Nakae K et al (2015) Deep learning of fMRI big data: a novel approach to subject-transfer decoding. arXiv preprint arXiv:1502.00093.

[47] Mitchell RS; Kumar V; Abbas AK; Fausto, N. Robbins Basic Pathology. Philadelphia: Saunders. ISBN 1-4160-2973-7. 8th edition.

[48] S. Guth, et al (2009) "Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination." European Journal of Clinical Investigation, vol. 39, pp: 699-706.

[49] E. Grant, et al (2015) "Thyroid ultrasound reporting lexicon: white paper of the ACR thyroid imaging, reporting and data system (TIRADS) committee." Journal of the American College of Radiology, vol. 12, pp: 1272-1279.

[50] Lee, H et al. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proceedings of the 26th International Conference On Machine Learning, ICML 2009. 77. 10.1145/1553374.1553453.

[51] Edmund SC, Syed ZA (2009). The Bethesda System for Reporting Thyroid Cytopathology, American Journal of Clinical Pathology, Volume 132, Issue 5, 1, Pages 658–665, https://doi.org/10.1309/AJCPPHLWMI3JV4LA

[52] Y. Boureau, J. Ponce, and Y. LeCun (2010). A theoretical analysis of feature pooling in visual recognition, In Proceedings of the 27th International Conference on Machine Learning, ICML-10, 111-118.

[53] Goodfellow I, Bengio Y, and Courville A(2016). Deep Learning. The MIT Press.

[54] Tieleman, T and Hinton G (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning

[55] Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-parameter Optimization. J. Machine Learning Res., 13, 281–305.

[56] Bonferroni, C. E. (1936) Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

[57] Sara S, Nicholas F, and Hinton G.(2017) Dynamic routing between capsules. In Advances in Neural Information Processing Systems, pp. 3859–3869.