Deep learning for magnetic resonance imaging-genomic mapping of invasive breast carcinoma

by

Qian Liu

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Biochemistry and Medical Genetics
University of Manitoba
Winnipeg

Table of Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

AUC	the area under the receiver operating characteristic (ROC) curve
CCLE	Cancer Cell Line Encyclopedia
CHi-C	map physical contacts between chromatin regions in cell nuclei using high-throughput sequencing
CPM	count per million
DA	denoising autoencoder
DL	deep learning
ER	estrogen receptor
FDA	the US Food and Drug Administration
FDR	false discovery rate
FN	False negative
FP	False positive
FPKM-UQ	Upper Quartile Fragments per Kilobase of transcript per Million mapped reads
Gd-DTPA	gadolinium diethylenetriamine penta-acetic acid
GSEA	gene set enrichment analysis
GWAS	genome-wide association studies
HER2	human epidermal growth factor receptor 2
ICGC	the International Cancer Genome Consortium
kde	kernel density estimation
KEGG	Kyoto Encyclopedia of Genes and Genomes
LASSO	least absolute shrinkage and selection operator
LME	Linear Mixed Effect
MR	magnetic resonance
MRI	magnetic resonance imaging
mRNA	messenger RNA
MSE	mean square error
N	lymph node metastasis
PAM	Prediction Analysis for Microarrays
PCA	principal component analysis
Perp	perplexity
PR	progesterone receptor
ReLU	Rectified Linear Unit
ROC	receiver operating characteristic
ssGSEA	single-sample gene set enrichment analysis
T	tumor size
t-SNE	t-Distributed Stochastic Neighbor Embedding
T1WI	T1-weighted image

T2WI	T2-weighted image
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive
TCIA-BRCA	The Cancer Genome Atlas Breast Cancer collection
TN	True negative
TNM	breast cancer TNM staging system
TP	True positive

Abstract

Objective

It has been believed that traditional handcrafted radiomic features extracted from magnetic resonance imaging (MRI) of tumors are normally shallow and low-ordered. Recent advancement in deep learning technology shows that the high-order deep radiomic features extracted automatically from tumor images can capture tumor heterogeneity in a more efficient way. We hypothesize that MRI-based deep radiomic phenotypes have significant associations with molecular profiles of breast cancer tumors. We aim to identify MRI-based signatures that can explain the potential underlying genetic mechanisms and predict the molecular classification of invasive breast cancers.

Methods

We developed a new deep learning model to retrospectively extract 4,096 MRI-based radiomic phenotypes from primary breast cancer tumor collected by The Cancer Imaging Archive (TCIA). These phenotypes of the tumors are then associated with genomic features (commercialized gene signatures, expression of risk genes, and pathways activities) of the corresponding molecular profiles (e.g. gene expression) and other clinical features collected from The Cancer Genome Atlas (TCGA). We developed novel association and classification methods to select the most-predictive radiogenomic features for the clinical phenotypes, including tumor size (T), lymph node metastasis(N) from breast cancer TNM staging system which is widely used in clinic, and status of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2).

Results

We find that transcriptional activities of various genetic pathways and gene signatures are positively associated with more than 1000 of the 4,096 MRI-based radiomic phenotypes. These radiomic phenotypes are also associated with the mRNA expression of the risk genes identified from two other genome-wide association studies. Higher performances are obtained in the prediction of HER2 status, ER status and tumor size (T) than PR status and lymph node metastasis (N). These identified MRI-based radiomic phenotypes also show significant power to stratify the breast cancer tumors, which may have a significant clinical impact.

Conclusion

Our radiogenomic approach for identifying MRI-based imaging signatures may pave potential pathways for the discovery of genetic mechanisms regulating specific tumor phenotypes and may enable a more rapid innovation of novel imaging modalities, hence accelerating their translation to personalized medicine.

Acknowledgement

I would first like to express my sincere gratitude to my supervisor Dr. Pingzhao Hu for his persistent support, encouragement, and guidance in overcoming numerous obstacles I have been facing through my thesis research during the past 2 years I spent at the University of Manitoba. His insightful instruction in the computational biology course also opens my mind to bioinformatics world. Furthermore, his passion and enthusiasm for academic research has encouraged me profoundly, leading me to continue my PhD work with him.

I would also like to extend my gratitude to my supervisory committee members Dr. Leigh Murphy and Dr. Yang Wang for their feedbacks, suggestions and encouragement which have kept me on track during this thesis research and have been particularly instrumental during key academic milestones including all the review meetings at both department-level and university-level.

I'm also grateful to all the faculty and staff members in the Department of Biochemistry and Medical Genetics who have supported me in various ways. Special thanks to my colleagues at the Hu-Lab and my department for all the communication, cooperation, and of course friendship.

Last but not least, I would like to dedicate this thesis to my family for supporting me spiritually throughout my master's program and my life in general.

1 Chapter 1: Background and Introduction

1.1 Dynamic Contrast-enhanced T1-weighted Magnetic Resonance Imaging

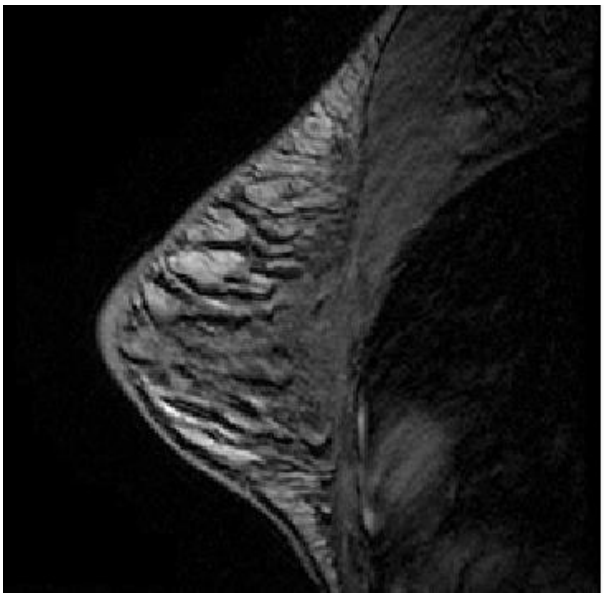
Medical imaging is widely involved in different phases of disease management due to its interpretability, real-time monitoring, noninvasiveness, etc [1]. Originating from simple X-ray imaging, medical imaging has developed a variety of branches according to different imaging theories. To form an image, there must be contrasts between contents. In the human body, these contrasts can come from natural properties of the tissue and/or introducing contrast media purposely.

The contrasts of normal X-ray imaging come from the differences between radiation absorption by tissues [2]. Bones have a much higher density than surrounding soft tissues in terms of the elemental composition, thus having a higher X-ray absorption, which provides a perfect natural contrast. This mechanism supports X-ray imaging to become the pervasive and key examination method for fracture. However, the limitation of the X-ray is its weakness in distinguishing amongst soft tissues, due to the lack of natural radiation absorption differences among these soft tissues including the gastrointestinal tract. Introducing artificial high X-ray absorption liquid into these soft tissues through body vessels can help to address this problem, which is known as contrast-enhanced X-ray imaging. However, limited by the dependence on vessels in this procedure, it still cannot be expected that every soft tissue can be well detected by enhancement X-ray imaging.

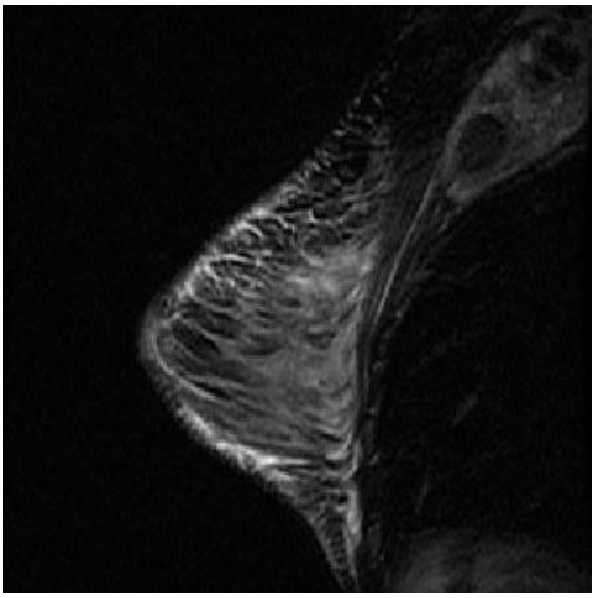
Different from X-ray imaging, the contrasts of normal magnetic resonance imaging (MRI) are determined by the interaction of a proton spin with external magnetic fields[3] and this procedure is controlled by multi-parameters[2]. These parameters mainly include the density of hydrogen atom proton in the tissues and magnetic relaxivity of the tissues. The spin of positively

charged hydrogen atom proton could generate a tiny magnetic momentum. This tiny magnetic momentum, along with a stronger external magnetic field, would force the proton to undergo a spinning-top-like precessional motion. If the external magnetic field is consistent, the precession of proton would also follow a consistent rate called Larmor frequency[4]. The stable spin system is able to absorb energy when an electromagnetic wave of the same Larmor frequency was introduced [3], which is the so-called magnetic resonance. It's not hard to understand that more hydrogen atom protons in the tissues can potentially contribute more tiny magnetic momentum, thus can be used to generate more detectable signals later, while less content of hydrogen in the tissues will likely end up with fewer signals. The difference would partially form the contrasts of the image constructed using these signals. However, the signal is not determined only by the density of spin proton, magnetic relaxivity also plays a role[2]. After the applied electromagnetic wave is canceled, the excited spin would start to relax to the stable status. This process is called relaxation, and there are two ways of relaxation, namely T1 relaxation and T2 relaxation. T1 refers to the way the spin releases the absorbed energy to the around substance (spin-lattice interaction), while T2 refers to the spins canceling the energy out within themselves(spin-spin interaction)[3]. T1 and T2 are intrinsic characteristics of organics and differ a lot among tissues. T1 relaxation is longer than T2 relaxation, and usually, they are independent but would happen simultaneously[4]. So, how would the T1 and T2 influence the MR signal? T1 time can influence the spin intensity indirectly if we add another electromagnetic wave before previous T1 relaxation has completed. Since not all of the spins can be excited in this case, those tissues with fast T1 relaxation would be more likely to be excited, therefore would contribute stronger MR signal, which turns out to be bright in the MR image. Tissues with longer T1 relaxation would appear dark in the image due to lack of excited spins. As for the T2 relaxation,

we can take advantage of the T2 time differences among tissues directly by setting the MR signal acquiring time. If we wait a while instead of acquiring the signal right after the excitation, then short T2 tissues would become dark because the relaxation has completed, and no signal would be detected. At the same time the long T2 tissues have not completed the relaxation, so the signal is still detectable to be bright in the image[4]. Therefore, by changing the combination of acquiring parameters, different intrinsic properties could be emphasized, forming the different types of MR images such as the T1-weighted image and the T2-weighted image. Because they reflect different information about the tissues, the grey levels or colors of the same tissue in different types of images differ. For example, fat tissue in T1-weighted MR images is bright, while it is dark in T2-weighted images (**Figure 1-1**).



T1-weighted image



T2-weighted image

Figure 1-1: Comparing the T1-weighted image with the T2-weighted image for the same body slide. Fat tissue in the T1WI is bright, while in the T2WI it is dark.

To be brief, the contrast of MRI is complex, mainly depending on proton density, T1 time, and T2 time. By changing excitation and acquisition strategies, we can generate a variety type of MR images that emphasize different intrinsic properties of the tissues, such as the T1-weighted images, the T2-weighted images, etc. The flexibility of emphasizing different contrast sources (i.e. the intrinsic properties of tissues) builds MRI a good reputation for soft tissue and tumor identification even without introducing external contrast agents.

By analogy to enhanced X-ray imaging, MRI could also employ external contrast agents to improve its sensitivity and specificity[5][6]. The most commonly used MRI enhancement agent is Gd-DTPA (gadolinium diethylenetriamine penta-acetic acid) which can reduce tissue's T1 time, therefore, forming a brighter contrast in the T1-weighted MR images. Another advantage of enhanced imaging is its ability to capture the contrast agents wash-in and wash-out time, which have important clinical diagnostic meaning. For example, some tumors show a contrast wash-in and wash-out behavior compared with their benign lesion, which can be used for differential diagnosis[7].

1.2 Breast Cancer and Dynamic Contrast-enhanced T1-weighted Magnetic Resonance Imaging in Breast Cancer

According to the latest Global Cancer Statistics published in 2018, breast cancer still remains the most commonly diagnosed cancer and the leading cause of death for women[8]. Breast cancer is known as a complex polygenetic disease, which means the risk of developing breast cancer is influenced by a large number of genes, which potentially results in the diversity of breast cancer patient' clinical outcomes among individuals and populations. A lot of efforts have been invested in identifying breast cancer risk genes, classifying different types of breast cancer, and staging the severity of the breast cancer.

Biomarkers like estrogen receptor (ER) status, progestogen receptor (PR) status, and human epidermal receptor 2 status are widely used in the clinic to gain a rough prospect of breast cancer patients' prognosis and response to endocrine therapy. Estrogen and progesterone are sex steroid hormones which affect growth, differentiation, and function of the mammary gland by binding to their receptors. While human epidermal growth factor is a common simulating factor that would increase the proliferation of cells, and it also acts through binding to its receptor in the cells. If cancer cells have estrogen receptors, the cancer is called ER-positive and will be influenced by estrogen. Usually, ER-positive patients have better survival than ER-negative patients and more response to endocrine therapy[9]. Likewise, if cancer cells have progesterone receptors, the cancer is called PR-positive and it will be able to receive signals from progesterone. If the cancer cells have mutation in the HER2 gene, too many human epidermal growth factor receptor twos would be made, and this type of breast cancer is called HER2-positive. HER2-

positive cancer usually has a rapid growth rate, but the patient is also more likely to benefit from the corresponding targeted treatment. ER, PR, and HER2 status are important clinic characteristics of breast cancer because such information can tell whether a patient would benefit from endocrine therapy or not. These three metrics can be tested through the immunohistochemistry examination.

Traditional TNM staging system is still the most commonly used method in the clinic to stage breast cancer, but a final decision of the stage might not be possible to be made before the surgery is performed [10]. In the system, T(Tumor) is used to simply describe the size of the tumor, N(Node) is used to describe nearby (regional) lymph nodes that are involved and M (Metastasis) is used to describe whether the cancer has spread from one part of the body to another[10]. The detail of the staging levels of the metrics T and N is shown in **Table 1-1**.

Table 1-1: Tumor(T) and Node(N) metrics in TNM staging system.

Metric	Level		Meaning
Tumor (T)	T1	T1mi	0.1cm across or less
		T1a	more than 0.1 cm but not more than 0.5 cm
		T1b	more than 0.5 cm but not more than 1 cm
		T1c	more than 1 cm but not more than 2 cm
	T2		more than 2 centimetres but no more than 5 centimetres
	T3		bigger than 5 centimetres
	T4	T4a	the tumour has spread into the chest wall
		T4b	the tumour has spread into the skin and the breast might be swollen
		T4c	the tumour has spread to both the skin and the chest wall
		T4d	inflammatory carcinoma – this is a cancer in which the overlying skin is red, swollen and painful
	TX		tumour size can't be assessed.
	Tis		ductal carcinoma in situ (DCIS)
Node (N)	N1		there are cancer cells are in the lymph nodes in the armpit, but the nodes are not stuck to surrounding tissues.
	N2	N2a	there are cancer cells in the lymph nodes in the armpit, which are stuck to each other and to other structures.

	N2b	there are cancer cells in the lymph nodes behind the breastbone (the internal mammary nodes), which have been seen on a scan or felt by the doctor. There is no evidence of cancer in lymph nodes in the armpit.
N3	N3a	there are cancer cells in lymph nodes below the collarbone
	N3b	there are cancer cells in lymph nodes in the armpit and behind the breastbone
	N3c	there are cancer cells in lymph nodes above the collarbone
NX		the lymph nodes can't be assessed (for example, if they were previously removed)
N0		there are no cancer cells in any nearby nodes

To take more genomic information into account, PAM50 (Prediction Analysis for Microarrays) intrinsic subtyping was developed and has been widely accepted nowadays due to its excellence of capturing biological features and predicting clinical outcomes of breast cancer. PAM50 uses expression information of 50 genes to classify patients into 4 subtypes: Luminal A, Luminal B, Her2-enriched and Basal-like [11][12]. Previous studies have confirmed the different expression patterns of these 4 subgroups using the large public databases such as The Cancer Genome Atlas (TCGA) and Cancer Cell Line Encyclopedia (CCLE)[12].

Besides the subtyping, identifying genes that associated with increased risk of breast cancer is also important and it is actually the foundation of gene-level subtyping. As biology entered the genomic era, it has been widely recognized that a large number of genes contribute to human cancer onset and development. Many genome-wide association studies (GWAS) in the cancer genetics field have been working on identifying genes or loci that are associated with the risk of cancer and they have already achieved fruitful results, which significantly impact today's medicine. For example, recently, Baxter et al. used Capture Hi-C, a technology to map physical contacts between chromatin regions in cell nuclei using high-throughput sequencing, to annotate 63 of the established breast cancer risk loci and then identified CHi-C interaction peaks

involving 110 putative target genes mapping to 33 of the loci[13]. Wu et al. performed a transcriptome-wide association study to evaluate associations of genetically predicted gene expression with breast cancer risk in 122,977 cases and 105,974 controls of European ancestry and identified 179 genes whose predicted expression was associated with breast cancer risk at false discovery rate (FDR) < 1.05×10^{-3} [14]. These genes are considered as breast cancer susceptibility genes or risk genes, which could be used as a short list of candidates for further studies.

Previously, different gene signatures, such as Prosigna[15], MammaPrint[16], Oncotype DX[17], were also generated as biomarkers for different clinical purposes. Each of these gene signatures include a list of genes curated based on tumor subtypes and other clinical characteristics and genomic profiles. Each of these gene signatures can be summarized into a patient-specific risk score using its gene expression profiles following interpretable algorithms. The scores can be used to predict patients' prognosis, recurrence, and therapeutic benefits. Many of these signatures have been well validated[18]. However, the cost of testing the expression levels of the genes in the lists are still expensive, usually thousands of USA dollars, which limited their clinical practice.

Similar to gene signatures, the patient-specific risk score can be inferred at pathway level. The definition of a biological pathway is a series of actions and reactions among molecules within a cell that leads to a certain product or a change in a cell[19]. The biological pathways could be subtyped into three main categories according to their involvement in different biological functions. These three types are namely metabolism pathways, gene-regulation pathways, and signaling pathways. Kyoto Encyclopedia of Genes and Genomes (KEGG) database collected 182 pathway maps representing the existing knowledge on each of the three

categories of the biological pathways[20]. Traditional case and control –based genomic experiments would identify some significant genes which are expressed differentially between the case group and the control group. Using the defined gene list, associated pathways could be identified as well. This is the so-called gene set enrichment analysis (GSEA) and it would put the relative pathways' activity scores, which could provide statistical evidence for the repression or activation of these pathways. However, GSEA has to be limited by the hypothesis of the designed case and control experiment, due to its requirement of a phenotypical label to define the differentially expressed gene sets. To make it more flexible, single-sample gene set enrichment analysis (ssGSEA) was proposed to calculate a single sample's pathway activity scores through its gene expression profile to get higher-ordered and biologically more interpretable features of the individual sample[21].

From the point of view of phenotype, medical imaging is applied to all phases of the management of spatially and temporally heterogeneous solid cancers like breast cancer in current clinic practice. MRI is well-known to be superior to X-ray imaging or ultrasound, but this doesn't mean MRI can replace X-ray and ultrasound. Actually, although some explorations try to extend MRI as a new screening method for breast cancer other than mammography [22], the clinical reality is that the benefit does not compensate for the relatively higher costs for MRI, in most cases [23]. The strategy of current breast cancer screening, which combines mammography, clinical examination and ultrasound, has been validated to have good performance in detecting early-stage and low-risk breast cancer[24]. But MRI still plays an important role in detecting high-risk multifocal breast cancer and provides better staging information[25] with a reported sensitivity of over 90% for detecting malignant invasive breast cancer[8][22]. Most importantly, MRI, especially dynamic enhanced MRI, appears to have higher sensitivity than mammography

or ultrasound in detecting tumors in those females who have inherited susceptibility of breast cancer[26][27], which indicates MRI's potential position in the genetic-based personal medicine.

Generally speaking, after the medical images are acquired, radiologists would make the diagnostic decisions based on their training and experience. As for the dynamic enhanced MR images, radiologists could observe the information contained in the images in two aspects, which are the morphologic and kinetic information during contrast medium wash-in/wash-out. Because there are a lot of parameters and variances in both acquisition procedure and individual intrinsic properties as we mentioned previously, which also increase the complexity and subjectivity of MRI related healthcare studies, diagnosis especially differential diagnosis should be done carefully. Therefore, since the application of MRI technology has been introduced in the clinical practice, people have begun to explore quantitative and objective observation standards. Current quantitative measurements generated by computer-aided diagnosis tools range from the simple features as the size and volume of the tumor, to the complex features as kinetic curves (wash-in/wash-out times and patterns). They have already been well applied in the clinical practice and performed very well in supporting the diagnosis of breast cancer[28].

1.3 Radiomics in Breast Cancer

As an expansion of computer-aided diagnosis, radiomics is defined as using high-throughput and large number (usually larger than 200) of quantitative features extracted from medical images by advanced mathematic techniques to describe disease phenotypes thereby predicting the clinical characteristics and outcomes[29]. Actually, radiomics intends to make the best use of the medical images by thoroughly mining the information embedded in the image and exploring any possible representativeness of it.

A typical workflow of radiomic-based study is as follows. Raw images are acquired in the first place following a standard and uniform procedure to make sure the origin of the data is comparable. Then the radiologists or oncologists need to be involved to mark the tumor regions in each of the images based on their experience so that subsequent feature extraction methods could be applied. There have been a large number of feature extraction algorithms published in the recent decade, and their representativeness and predictive ability are identified and evaluated later on[29].

We have learned from the principle of dynamic enhanced MRI mentioned above that MRI is a complex process and it has the ability to capture different intrinsic properties of the tissues. It is reasonable for us to expect that dynamic enhanced MR image could potentially contain the surrogate information of the disease about clinical characteristics/outcomes, pathology, genomics, etc. MRI has all the advantages in terms of non-invasive, real-time, and rich enough for us to mine. If significant associations can be identified between the radiomic features and other known important features of the disease, then the MRI could serve as a "one-stop for everything" tool to better diagnose, monitor, and treat the disease.

Zhu et al. reported 38 quantitative radiomic features (**Table 1-2**) extracted from the dynamic enhanced MR images of 91 breast cancer patients, which characterize the size, shape, morphology, enhancement texture, kinetics, and variance kinetics of the breast tumor[30].

Table 1-2: The 38 semi-auto dynamic enhanced MRI-based radiomic features of breast cancer reported by a previous study.

Feature category*	Name
Size features	Lesion volume
	Lesion diameter
	Lesion surface area
	Maximum length
Shape features	Sphericity
	Irregularity
	Surface-to-volume ratio
Morphological features	Margin sharpness
	Variance of margin sharpness
	Variance of radial gradient histogram
Enhancement texture features	Contrast
	Correlation
	Difference entropy
	Difference variance
	Angular second moment (energy)
	Entropy
	Inverse difference moment

	Information measure of correlation
	Information measure of correlation
	Maximum correlation coefficient
	Sum average
	Sum entropy
	Sum variance
	Sum of squares (variance)
Kinetic curve features	Maximum enhancement
	Time to peak
	Uptake rate
	Washout rate
	Curve shape index
	Enhancement at first postcontrast time point
	Signal enhancement ratio
	Volume of most enhancing voxels
	Total rate variation
	Normalized total rate variation
Enhancement-variance kinetics features	Maximum variance of enhancement
	Time to peak at maximum variance
	Enhancement variance increasing rate
	Enhancement variance decreasing rate

However, their radiomic features are semi-automatically extracted. Actually, although there are some imaging biomarkers identified by previous radiomics studies following such a

workflow [31], they are still subjective due to the requesting of radiologists to segment the regions of interest[32]. What's more, it has been suggested that these handcrafted biomarkers, such as the tumor size and shape, are shallow and may not fully represent the heterogeneity of images[33]. These limitations must be overcome so that radiomics can be better explored on the right track.

1.4 Radiogenomics in Breast Cancer

From the genomic side, the flourishing of genomic biomarkers has successfully revealed some biological mechanisms of diseases and performed very well in supporting clinical decisions[34]. However, although the development of biological technologies has highly reduced the cost of genome sequencing, it is still not that practical to test these genomic biomarkers in vivo and to use them in the real-time clinical routine[35]. Therefore, the attempt of integrating radiomics and genomics together has driven a brand-new discipline which is called radiogenomics. Radiogenomics aims to solve clinical problems by taking account of both radiomics and genomics.

For example, Zhu et al. investigated their 38 semiauto- dynamic enhanced MR image features and reported significant associations between the image features and genomic signatures[35]. However, since their image features were obtained under the radiologists' prior knowledge, the objectivity was still doubted.

Besides the subjectivity, another critical issue in the radiogenomics analysis is the increasing computing complexity in both the feature extraction and later on association and prediction analyses, which is due to the nature of the high-dimensional and large-scale multi-omics data.

1.5 Deep Learning in Radiomics and Radiogenomics

Deep learning is a technique under the broad term of artificial intelligence which is not new but comes afresh in recent years due to the development of advanced hardware techniques in the computer science field. Artificial intelligence is broadly defined as computer programs trained to learn, reason, perceive, infer, communicate and make decisions as humans do. Under this umbrella, machine learning is the approaches of letting computers not just be programmed to be smart at something, but actually learn from its experience and improve its performance over time. Deep learning is a kind of machine learning which uses a specific model called deep artificial neural network to learn knowledge. The naming of artificial neural network was inspired by the networks of neurons in the human brain. Deep artificial neural network is composed of multiple non-linear data processing layers in its mathematic architecture to extract features with multiple levels of abstraction (**Figure 1-2**) [36].

Deep learning is not a new technique but was remained under explored for many years because of the computing complexity. However, recently it has been warmly re-embraced in a lot of domains due to the rapid development of computer hardware and its power for handling big data. Like other learning methods, in order to solve different problems, deep learning could be implemented in a supervised or unsupervised way, which depends on whether the label information is given to the model or not. Supervised learning algorithms would use a guiding label to train a model and the model is often used to predict this label in an unlabeled dataset later on[37], while unsupervised learning algorithms do not force the model to learn a certain label, but it is more data-driven instead of hypothesis-driven[38][39].

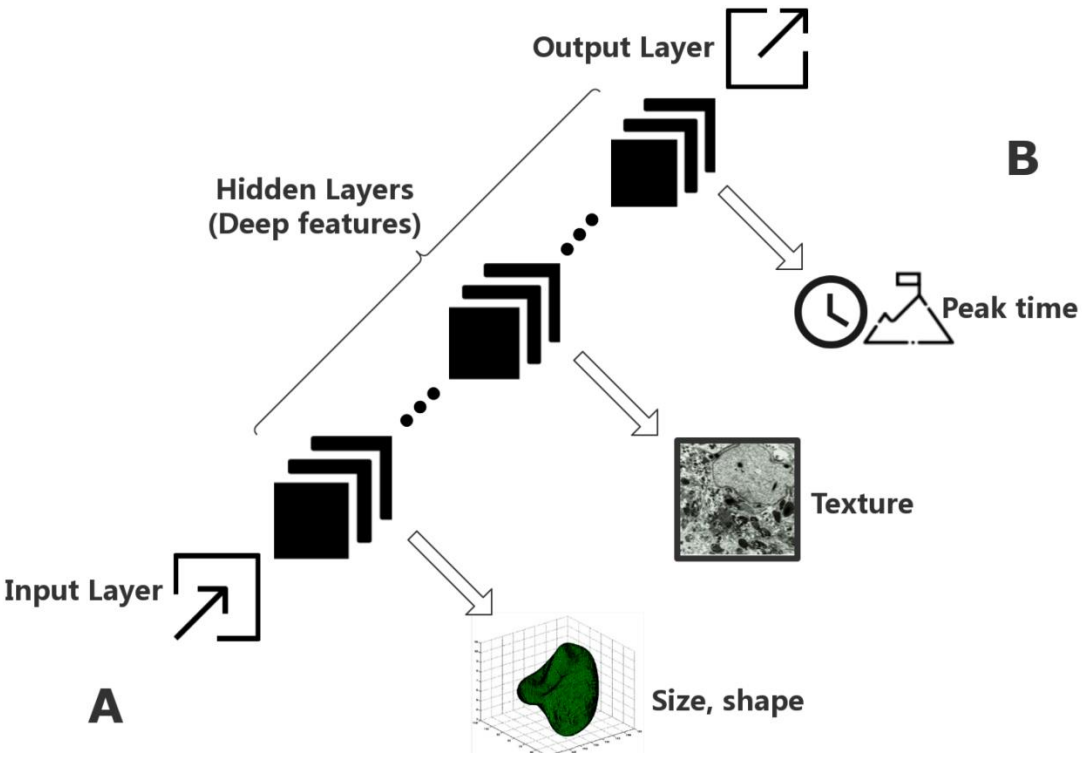


Figure 1-2: Deep features with multi-level abstraction. Deep artificial neural network is composed of multiple non-linear data processing layers in its mathematic architecture and each hidden layer represents features with multiple levels of abstraction.

Recently, deep learning has been applied to perform radiomic-based analyses. Li et al. developed a supervised deep learning model to automatically extract features from glioma patients' MR images and these features were estimated to have tumor grading significance[40]. However, they didn't perform further exploration of the genomic surrogate of these deep image features, which made their study limited to radiomics instead of radiogenomics.

Among the unsupervised learning algorithms, autoencoder is a new technology that uses the data itself as the learning objective or label. Therefore, it is also known as self-labeled or self-supervised deep learning. Traditional autoencoders may face an invalid learning problem

when the number of hidden nodes is larger than the input size, which means the autoencoder would figure out a tricky way to finish the training task by just outputting the result that is exactly the same as the input data without learning anything. To avoid this potential risk, denoising autoencoders come up with adding some noise to the input data on purpose. Vincent et.al brought the concept of denoising autoencoders into deep learning and built a specialized feature extraction architecture [41]. The key idea of denoising autoencoders as mentioned above is to add random noise in the raw data before it is input into the network. After the encode and decode processes, raw data would be reconstructed from the noisy data, while the compact and efficient representations of the raw data could be learned as well [41].

Features extracted by a data-driven method under the unsupervised way are believed to have higher flexibility of representing the intrinsic pattern of the data being analyzed than supervised hypothesis-driven methods. Since the radiomics has a core goal of identifying as many as possible high-quality surrogate biomarkers, there is a slightly larger tendency in this field to use unsupervised learning like denoising autoencoders to extract features from medical images.

Deep learning has been believed to have the power to extract robust features from images, the increasing number of attempts to introduce deep learning to radiomics have proved this. However, deep learning-based auto- image features have not been well studied in radiogenomics yet. Most of the completed radiogenomic studies were still done in a semiauto- way. Radiogenomics is eager for automatic and powerful algorithms to push its limits, while deep learning could provide the solutions.

2 Chapter 2: Motivation, Hypothesis and Research Objectives

2.1 Motivation

MRI especially dynamic enhanced MRI has the potential to encompass clinically relevant deep information due to MRI theory at the proton level, which is worthy to be dug up. Due to the severity and complexity of breast cancer, a large number of genetic and clinical research has been performed, which provides us with validated and valuable biomarkers to focus on. Meanwhile, the advancements in related biological techniques and computational techniques equip us to dive into radiogenomics deeply without considering too much about the cost and time.

The limitation of being subjective or lack of genomic information makes the image biomarkers identified by previous radiogenomic studies hard to be translated into clinical reality. The need of objectively mining deeper information from the complex medical images and exploring their genomic potential motivates us to identify radiomic biomarkers which can be acquired automatically by an advanced and effective deep learning model like denoising autoencoder, and also have strong genomic and clinical significance. An ideal thought is that the associations between auto- radiomic features and genomic signatures as well as the clinical outcomes could be learned thoroughly. The identified significantly associated radiomic biomarkers thereby can be used to predict those important genomic biomarkers and clinical outcomes, supporting personal medicine decision-making in an indirect way.

It is expected that if the same radiomic features and genomic features can be significantly associated with the same clinical outcomes, then the radiomic features can be potentially used as a cost-saving non-invasive way to replace genomic signatures for cancer diagnosis, prognosis or

therapeutic use. Furthermore, they can also help deepen the understanding of cancer's initiation and progression (**Figure 2-1**).

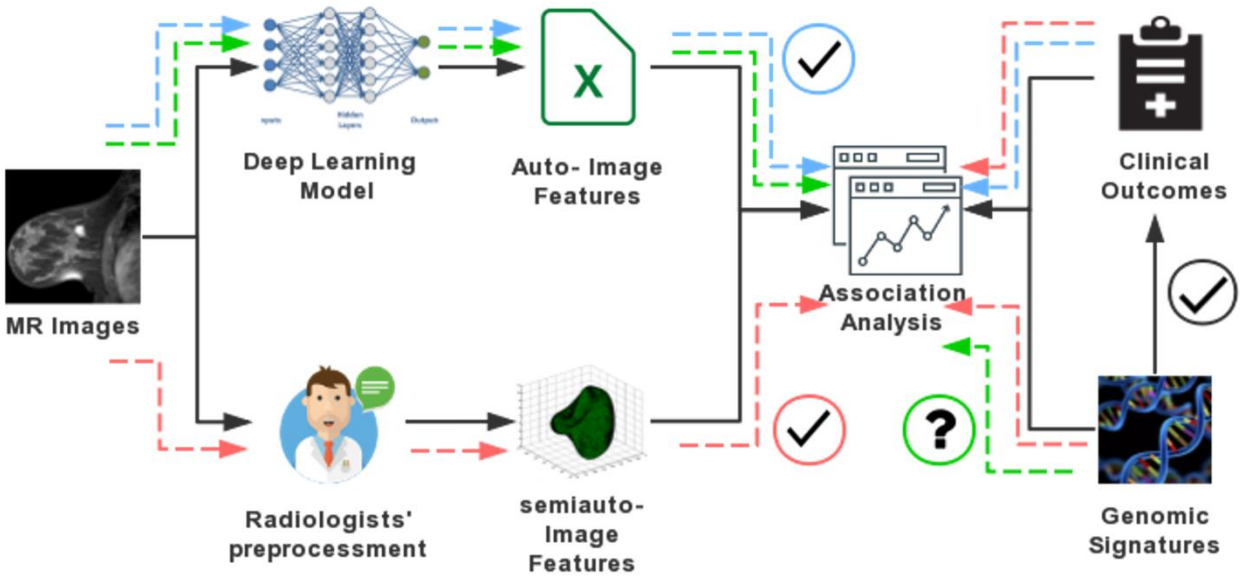


Figure 2-1: A flowchart illustrating the rationale and significance of this project. The red dash line represents the workflow of traditional semiauto-image biomarker identification. It has been well studied and supported by genomic associations. The blue dash line represents the workflow of auto-image biomarker identification. It has been well studied the association of the features with clinical but not genomic features. The green dash line represents the workflow of this study. We intend to identify auto-image biomarkers and study their association with both clinical and genomic features.

2.2 Hypothesis

In this study, we hypothesize that dynamic enhanced MRI-based auto- image features extracted by denoising autoencoder models have significant association with breast cancer patients' genomic profiles. We also hypothesize that these image features could also be used to predict patients' clinical outcomes. Therefore, the radiogenomic findings could potentially be used for supporting personal medicine decision.

2.3 Research Aims

We have three aims:

1. Develop a robust deep learning model in a denoising autoencoder way to automatically extract auto-image features from breast cancer dynamic enhanced MR images;

2. Identify the significant associations between these auto-image features with genomic features;

3. Predict clinical outcomes using these auto-image features.

3 Chapter 3: Materials and Methods

3.1 Data Sources

Three publicly available datasets were involved in this retrospective study, namely MR image data, genomic data, and clinical data. These datasets were downloaded from The Cancer Image Archive (TCIA) and The Cancer Genome Atlas Breast Cancer collection (TCGA-BRCA)[44]. TCGA-BRCA is the counterpart of TCIA-BRCA project. TCIA provides images and TCGA provides matched clinical, genetic, and pathological data for the same patients. Several preprocessing procedures were done before the feature extraction and further statistical analyses were performed. The downloading and preprocessing procedures of the three data sources are shown in **Figure 3-1**. The details of these procedures are discussed in the following subsections.

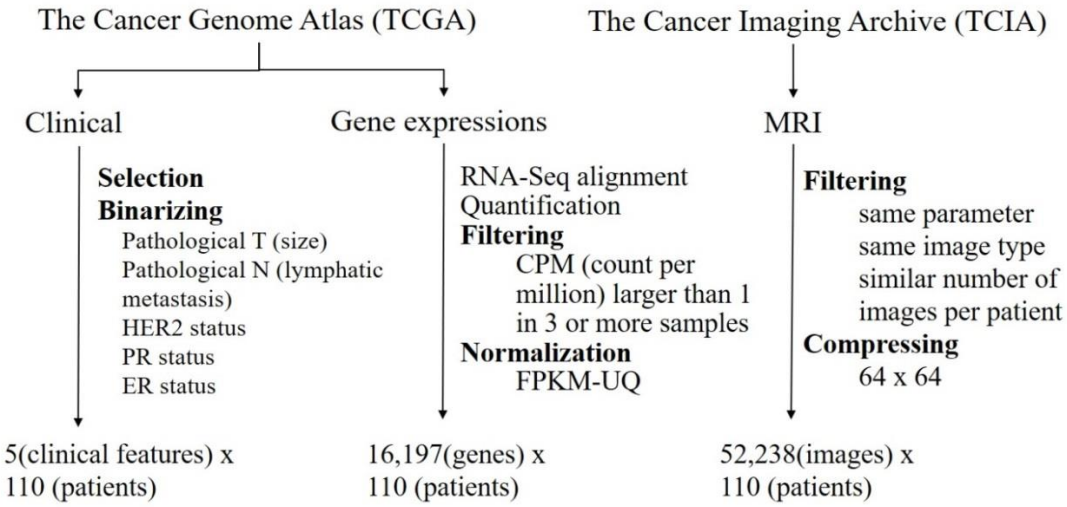


Figure 3-1: A flowchart showing the downloading and preprocessing procedures of the three data sources used in this study. Bold items were done by us, while RNA-Seq alignment and quantification were done by the TCGA database.

3.1.1 MRI data

We downloaded the raw MR images of 139 breast cancer patients stored in TCIA-BRCA[42]. These MR images were acquired under different devices and machine parameters, which resulted in different types such as the T1-weighted images, the T2-weighted images, and so on[43]. As discussed in the Chapter 1, different tissue intrinsic properties are emphasized in these types of images; the same tissue shows different grey levels (colors) in different types of MR images. For example, fat tissue in the T1-weighted MR images is bright, while in the T2-weighted images it is dark (**Figure 1-1**). Therefore, to keep our study comparable between patients and reduce the error caused by differences of image series, we only focused on the T1-weighted dynamic enhancement series from each study-level image sets as shown in **Figure 3-2**. It should be noted that one MRI series is like a 3-dimensional movie with a series of frames. Images within one series look very similar in contrast and resolution because they were acquired under the same parameters. The only difference is that they show different body slices. The organization structure of the MRI raw data is shown in **Figure 3-2**.

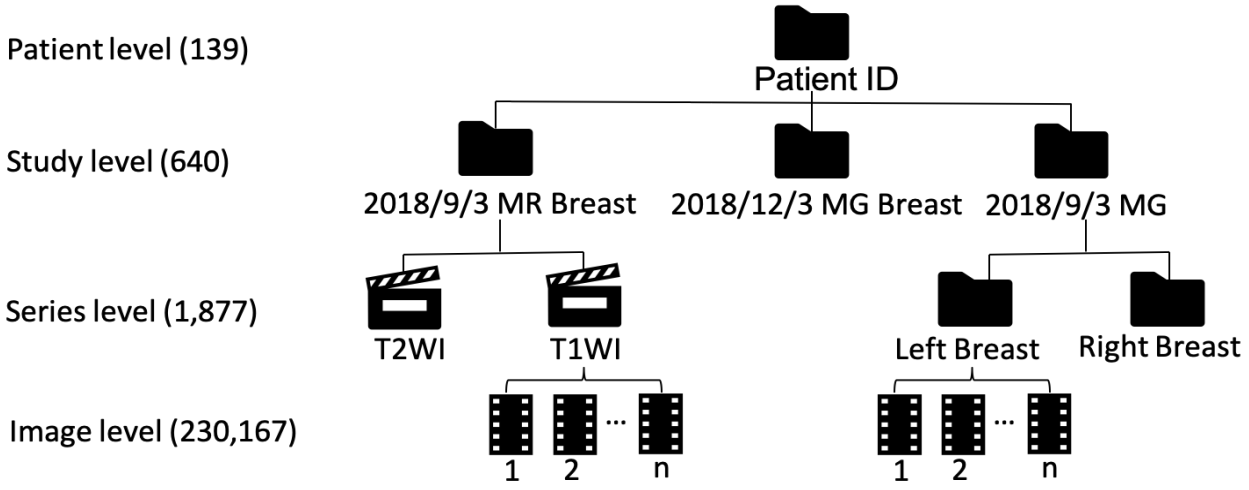


Figure 3-2: The organization structure of TCGA breast cancer images. The total number of patients in TCIA-BRCA collection is 139. The top-level of TCIA image archive structure is the individual patients. Then according to the different time points and exam modalities, each patient is classified into different "study", where MR refers to magnetic resonance images while MG refers to mammography images. We focus on only MR in this study. Under each study-level, "series" level image set is a combination of similar images that were acquired under the same device parameters but different body slides. While within a series (same patient, same date, same parameter setting), different numbers of images were acquired at different body slices of the same patient. These images look very similar in contrast and resolution to one other but different only by the body slice.

It should be also noted that each patient has a different number of images as shown in **Figure 3-3**. To avoid potential bias, we filtered out three patients who have more than 2,000 images. After the filtering, we were left with 52,238 images from the remaining 110 breast cancer patients.

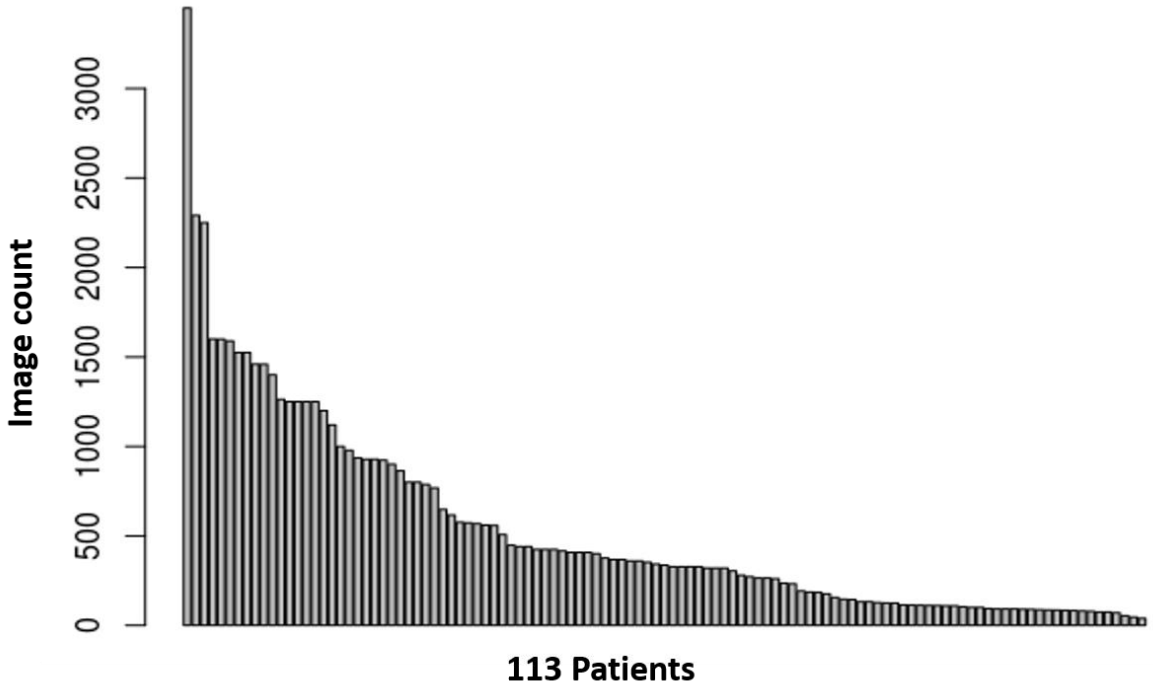


Figure 3-3: Distribution of image numbers by patients. There are three patients who have more than 2,000 images, so we excluded them to avoid potential bias. After deleting the images of the three patients, we were left with 52,238 images for 110 patients.

However, these images still have different image size (**Figure 3-4**) and image depth. Some preprocesses were applied to make these images comparable. We first zoomed the image size of all images to 64×64 pixels uniformly and then rescaled the pixel values to the range of 0 to 1 for meeting deep learning model requirements. After all, we have 52,238 images from 110 patients as samples and each sample has 64×64 pixels.

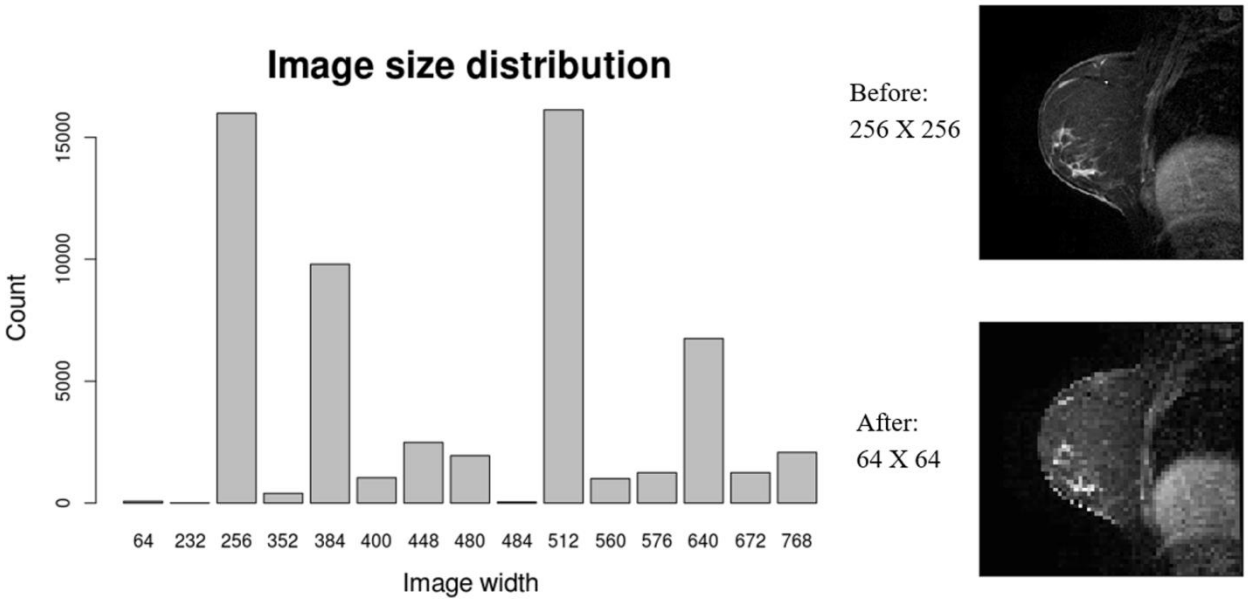


Figure 3-4: Size distribution of the raw MR images. The resolution of the images was lost after we reshaped them to a uniform size of 64×64.

3.1.2 Genomic data

The clinical and genomic (mRNA sequencing-based gene expressions) data of those 110 breast cancer patients who also have MRI data in TCIA were retrieved from TCGA by TCGA-Assembler[45]. For gene expression data, the sequencing, alignment, quality control, and quantification were done by the TCGA team previously[44]. The pipeline of these upstream processing analyses begins with the pre-alignment quality assessment using FASTQC tool[46]. Then the alignment is performed through a two-pass method with STAR2[47], which aligns each read group separately and then merges them into a final alignment using a splice junction

detection step proposed by the International Cancer Genome Consortium ICGC[44]. The alignment workflow outputs a BAM file, which contains both aligned and unaligned reads. Then a post-alignment quality assessment is performed using RNA-SeQC Tools[48]. Following that, the aligned BAM files would undergo an expression quantification workflow to generate quantified gene expression levels using HT-Seq-Count[49] and the GENCODE v22[50] is used as the reference genome for gene annotation. Although there is normalized data available for people to download through TCGA data portal, we decided to start with the raw gene count data and performed the filtering and normalization by ourselves so that different criteria could be tested.

Using the TCGA-Assembler tool[45], we excluded unexpressed genes with count per million (CPM) less than 1 in 3 or more patients. Normalization of the data using Upper Quartile Fragments per Kilobase of transcript per Million mapped reads (FPKM-UQ)[51] was also performed. We ended up with 16,197 (genes) × 110 (patients) for gene expression data.

3.1.3 Clinical data

Clinical data were carefully scanned. We extracted and binarized the clinical characteristics such as pathological TN status (Tumor size (T), Node metastasis (N)), estrogen receptor (ER) status, progesterone receptor (PR) status, human epidermal growth factor receptor 2 (HER2) status. Specifically, tumors with a size smaller than 2cm were assigned to the T-negative group, while those with size larger than 2cm were set to the T-positive group. Node metastasis was

coded as N-positive/N-negative simply according to whether there were lymph nodes invasion or not. The clinical data of the 110 patients are shown in **Figure 3-5.**

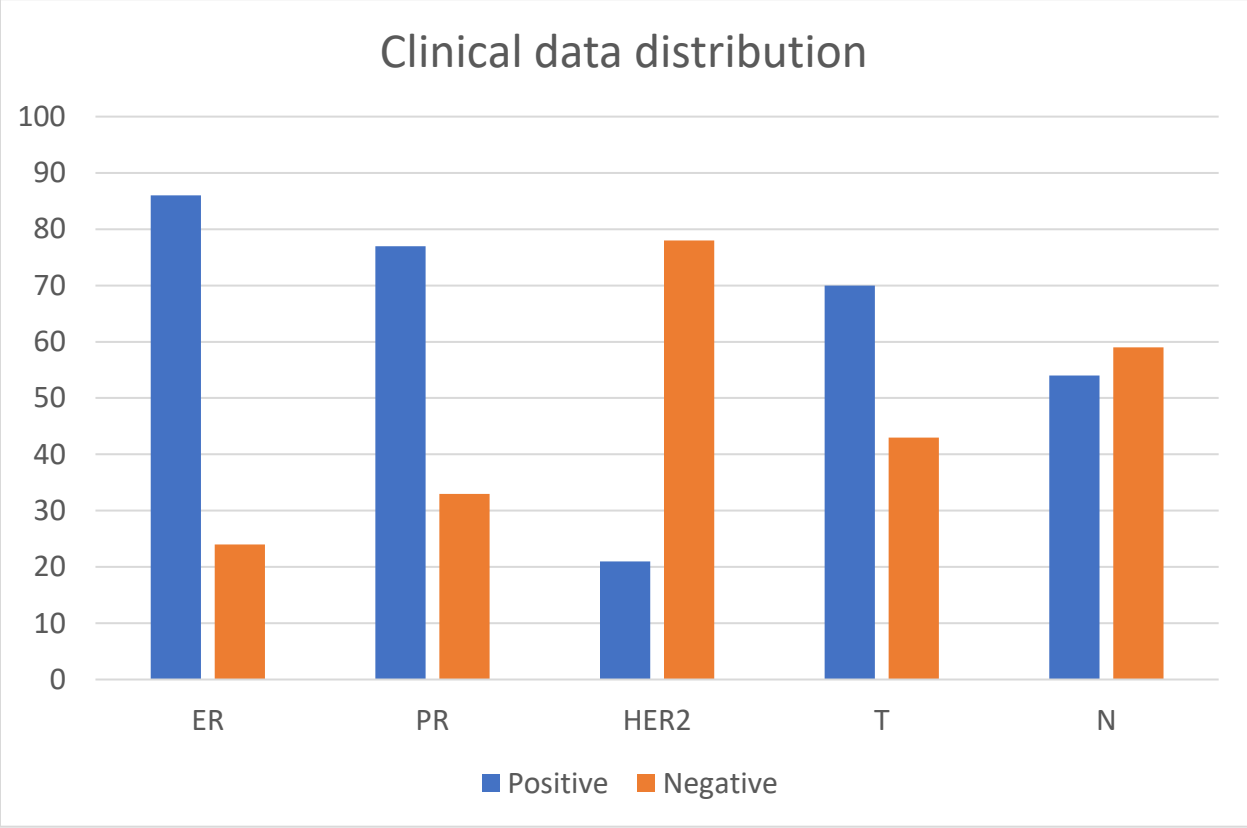


Figure 3-5: The 5 binarized clinical characteristics and their distribution in the 110 patients. The Y-axis is the number of patients and the X-axis is the clinical subgroups.

3.1.4 Breast cancer risk genes

Breast cancer is a complex and polygenetic disease. Identifying genes that are associated with the increased risk of breast cancer is important, not only because it is the foundation of gene-level subtyping, but also because the identified risk genes could be served as a shortlist of marks for further studies. Besides the well-known mutations in BRCA1, BRCA2, TP53 that have been accepted as high-risk factors of breast cancer onset, many other breast cancer risk genes

were identified by GWAS. Most recently, Baxter. et al. used Capture Hi-C to annotate 63 of the established breast cancer risk loci and identified CHi-C interaction peaks involving 110 putative target genes mapping to 33 loci[13]. Wu et al. performed a transcriptome-wide association study to evaluate associations of genetically predicted gene expression with breast cancer risk in 122,977 cases and 105,974 controls of European ancestry and identified 179 genes whose predicted expression was associated with breast cancer risk at FDR < 1.05×10^{-3} [14]. We used this list of 288 well-validated breast cancer risk genes identified from these two studies [13, 14] as the candidate breast cancer risk genes for further examination in our study which was a subset of our gene expression data with size of 288 (genes) ×110 (patients).

3.1.5 Breast cancer gene signatures

Based on different sets of risk genes (also called as gene signatures) and their performance in breast cancer recurrence or prognosis prediction, some patient-specific numeric risk scores have been designed and calculated for each of the gene signatures. These gene signature-specific risk scores quantified the risk of recurrence, risk of death, and/or effect of therapy. Their significance in predicting patient outcomes has been well-validated. What's more, some of the genetic testing based on these gene signatures have already been approved and commercialized[18].

In this study, we calculated the patients-specific risk scores for the 6 published and commercialized gene signatures (**Table 3-1**) using R package genefu [52]. These risk scores measure prognostic significance at patient level [52]. Oncotype DX is a recurrence risk score calculated from the expression of 16 cancer-related genes and 5 reference genes. It has been validated to have the ability of quantifying the distant recurrence in ER-positive, N-negative, and

tamoxifen-treated breast cancer[53]. EndoPredicte examines the expression of 8 cancer related genes and 3 reference genes, which has been validated as an independent predictor of distant recurrence in ER-positive, HER2-negative, endocrine-treated breast cancer patients[54]. Prosigna (rorS) is a gene expression signature estimating distant recurrence risk of ER-positive, PR-positive, hormone-treated, postmenopausal women with breast cancer. It is calculated from Prediction Analysis of Microarray (PAM50) gene set[55]. MammaPrint (GENE70) is offered in clinic as a gene expression signature for N-negative breast cancer women under the age of 61. It is calculated from the expression of 70 genes and could predict the benefit of adjuvant therapy[55][56]. GENIUS is a predicted prognostic risk score applicable for any subtype of breast cancer. It has a hyperparameter to decide the number of included genes[57]. PIK3CA-GS is derived from exon 20 (the kinase domain) mutations and is able to predict PIK3CA mutation status and tamoxifen sensitivity of ER-positive and HER2-negative breast cancer[58].

Table 3-1: The six gene signatures used in this study.

Gene signatures	Clinical usage	Comments	Genes
OncotypeDX	recurrent risk prediction of ER+/HER2- breast cancer	Approved by FDA. Available in USA, but not available in majority provinces in Canada. >$4,000 USD/test.	ESR1, PGR, BCL2, SCUBE2, Ki67, STK15, Survivin, CCNB1, MYL2, HER-2, GRB7, MMP11, CTSL2, GSTM1, CD68, BACG1
endoPredict	10 years recurrent risk prediction of ER+/HER2- breast cancer	More recent tool. More accurate than OncotypeDX. Not approved by FDA. Only available in Europe.	BIRC5, UBE2C, DHCR7, RBBP8, IL6ST, AZGP1, MGP, STC2, CALM2, OAZ1, RPL37A
rorS (Prosigna)	Prediction of distant (another	Approved by FDA and Europe. Will be	ACTR3B, ANLN, BAG1, BCL2, BIRC5, BLVRA, CCNB1, CCNE1, CDC20, CDC6, CDH3, CENPF,

	part of the body called metastatic cancer) recurrence risk.	available in Alberta. >$3,000USD/test	CEP55, CXXC5, EGFR, ERBB2, ESR1, EXO1, FGFR4, FOXA1, FOXC1, GPR160, GRB7, KIF2C, KRT14, KRT17, KRT5, MAPT, MDM2, MELK, MIA, MKI67, MLPH, MMP11, MYBL2, MYC, NAT1, NDC80, NUF2, ORC6L, PGR, PHGDH, PTTG1, RRM2, SFRP1, SLC39A6, TMEM45B, TYMS, UBE2C, UBE2T
gene70 (MammaPrint)	Prognosis and chemotherapy efficacy	70 genes	BBC3, EGLN1, BCL2, TGFB3 , ESM1, IGFBP5, SCUBE2, TGFB3, WISP1, FLT1, HRASLS, STK32B, RASSF7, DCK, MELK, EXT1, GNAZ, EBF4, MTDH, PITRM1, QSCN6L1, CCNE2, ECT2, CENPA, LIN9, KNTC2, MCM6, NUSAP1, ORC6L, TSPYL5, RUNDC1, PRC1, RFC4, RECQL5, CDCA7, DTL, COL4A2, GPR180, MMP9, GPR126, RTN4RL1, DIAPH3, CDC42BPA, PALM2, ALDH4A1, AYTL2, OXCT1, PECI, GMPS, GSTM3, SLC2A3, FGF18, COL4A2, EGLN1, MMP9, LOC100288906, C9orf30, ZNF533, C16orf61, SERF1A, C20orf46, LOC730018, LOC100131053, AA555029_RC, LGP2, NMU, UCHL5, JHDM1D, AP2B1, MS4-A7, RAB6B
GENIUS	prognostic risk	Hyperparameter to determine the size of gene set	
Pik3cags-GS	Tamoxifen monotherapy outcomes of ER+/HER2- breast cancer	PIK3CA mutations (exon 20 kinase domain)	PIK3CA

3.1.6 KEGG pathways

KEGG database is a collection of 182 pathways representing the existing knowledge on each of the three categories of the biological pathways[20], which include metabolism pathways, gene-regulation pathways and signaling pathways. We calculated the pathway activity scores of these 182 KEGG pathways from the expression of the 16,197 genes in the 110 patients using the Single Sample Gene Set Enrichment Analysis (ssGSEA) function[59] which was implemented in the GenePattern toolkit [60]. Eventually, we got an activity score matrix with size of 182(pathways) ×110(patients).

3.2 Deep Learning Model for Extracting Deep Radiomic Features

When applying an unsupervised deep learning model to extract features from data, what we want is not the output results of the interconnected layers of the mathematical architecture of the deep neural network. Instead, we are more interested in the information in the hidden layers in the middle of the model structure, which would be our deep features when they are exported from the well-trained model and would be passed on to the follow-up analysis. As we mentioned in the Chapter 1, we do not want to force our model to learn a set of features that only work well for predicting certain known information, but we wish the deep features to capture the intrinsic patterns of the data. Therefore, we won't give any known information as labels to the model. To do this, here we employ the autoencoder to solve the problem. The idea of the autoencoder is to use the input data itself as labels to train a special structured model that has a bottom neck in its architecture. The bottle neck means the model first encodes the raw input data to a lower dimensional space, then decodes the lower dimensional projection of the data back to the high dimensional space. It is easy to follow the idea because if the low dimensional projection of the raw data could be successfully decoded to the original high dimension format, and could be very similar to the raw data, then we could say the projection of the raw data in the low dimensional space is representative. However, there is a limitation using this kind of traditional autoencoder, which is the so-called identity learning problem, that is, when the bottle neck has the size larger than the raw data, the model might become an identity function which only returns exactly the same output as the input without any intermediate processing. This situation is likely to happen because the introducing of kernels would usually increase the dimension of the middle layers of the model. To solve the identity learning problem, we introduce random noise to the raw data to avoid the model having exactly the same input and output. This is the key principle of denoising

autoencoder and the idea is also easy to follow by thinking in the way that if a deep learning model could recover data from its corrupted version, then we could say the model is more intelligent and the features it has learned in the deepest layer are more representative.

We developed a stacked convolutional denoising autoencoder (DA) model using Keras[61] with Tensorflow[62] as backend to automatically extract image features (**Figure 3-6**). It has 5 layers in the encode phase and 5 layers in the decode phase.

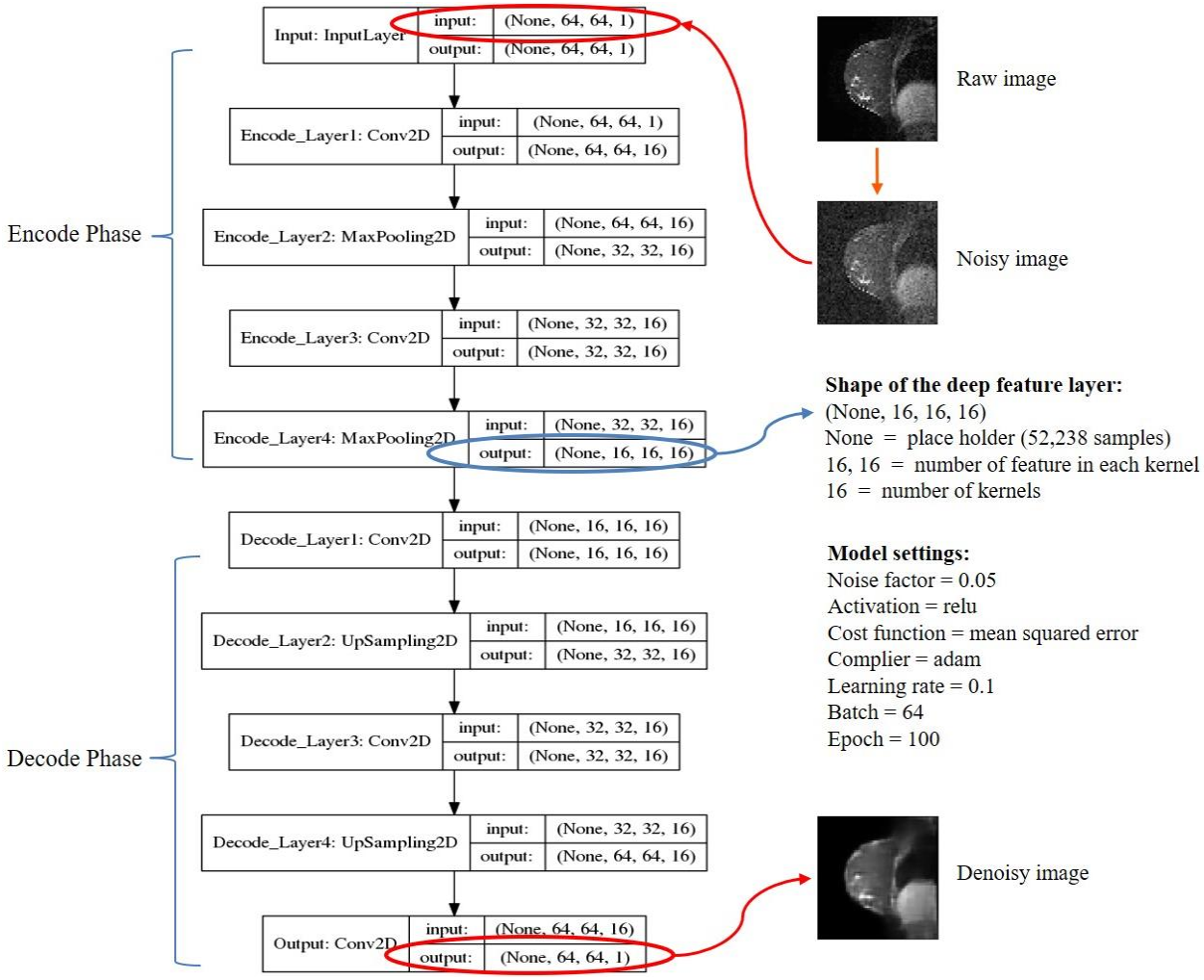


Figure 3-6: The DA model used in this study to extract deep radiomic features. There are two convolutional layers and two max-pooling layers in the encode phase, two convolutional layers, and two upsampling layers in the decode phase. For the (None, n, n, m), "None" is the place holding our 52,238 image samples, (n, n) in the middle indicates the number of features in each kernel. m is the number of kernels.

Before the raw data were inputted into the DA model, a 5% level of normally distributed random noise was added into the data using the following formula (**3-1**).

$$noisy\_data = raw\_data + noise\_level \times random.normal\ (0\ ,\ 1,\ size\ (raw\_data)) \quad \textbf{(3-1)}$$

Here *noise_level* was set to 0.05; *random.normal ()* is a function in Python NumPy package which was employed to generate random numbers normally distributed in a range of 0 to 1.

In the encoding phase, noised data were processed by 2 repeated convolutional blocks which contained one 2-dimensional convolutional layer followed by a 2×2 max-pooling layer. In this type of neural networks, convolutional operation could be described as using a feature detector/filter/kernel to slide across the input image, then store the sum of their dot product to a feature map. This operation could reduce the input to its essential features. Actually, in the real convolutional neural networks, several convolutional operations would be included and the model would determine the weights in the kernels in a backpropagation way during training[36]. Max-pooling summarizes a certain size of the neighboring values within a feature map to increase the abstraction level of the feature map and protects from overfitting[63]. In the decoding phase, upsampling was used to increase the size back to the same as the input by simply repeating the values[64].

Rectified Linear Unit (ReLU) was selected as the activation function to introduce interactions and non-linearities into the model. ReLU is the most commonly used activation function in deep learning[65]. It returns 0 if the input is negative, and returns the same value as the input if the input is positive. ReLU function could be illustrated by the following formula (**3-2**).

$$f(x) = max\ (0, x) \quad \textbf{(3-2)}$$

The cost function is a training parameter, which is also called loss function measuring how the prediction is similar to the ground truth. Actually, the goal of training is to minimise the cost function. We selected mean square error (MSE) as cost function in this study.

$$\boldsymbol{MSE = \frac{1}{N}\sum_{i=1}^{N}(yi - \tilde{y}i)} \quad \textbf{(3-3)}$$

Here y is the ground truth, $\tilde{y}$ is the prediction. In our case, y is the raw image data, $\tilde{y}$ is the denoised image data. N is the sample size. The goal of our training is to make the denoised images as close as the raw images.

Other hyperparameters required in the training step were also chosen carefully. Adam was used as optimizer[66], the original learning rate was set to 0.1, the batch size was 64, and the epoch was 100.

After the model was well trained, raw data were likely reconstructed from the noisy data, while the compact and efficient representations from the raw data were learned as well[41]. We extracted the output of last encode hidden layer as our radiomic features (**Figure 3-6** blue circle). The radiomic feature matrix, which has a dimension of 52,238×4,096, where 52,238 is the image size from 110 patients while 4,096 is the deep radiomic feature size from 16 kernels with kernel size of 16×16.

3.3 Preprocessing of the Learned Deep Features and Their Visualization

For the extracted deep radiomic features, a sample-level quantile normalization was then performed using the following algorithm to make the deep features comparable[67].

Algorithm 3.1: Quantile normalization

Input:

matrix X has dimension of $p \times n$

where n is the sample size, while p is the feature size.

Procedure:

1. Sort each column of X to form X_{sort}
2. Take the means across rows of X_{sort} and assign this mean to each element in the row to get X'_{sort}
3. Get $X_{normalized}$ by rearranging each column of X_{sort} to have the same ordering as original X.

Output:

normalized data matrix $X_{normalized}$

After quantile normalization was performed, the distributions of the unnormalized and normalized image-level deep features were visualized using kdeplot (kernel density estimation) function from Python package seaborn[68].

Unlike the semiauto handcraft radiomic features, deep radiomic features extracted by the deep learning model do not have predefined biological meaning. Therefore, visualization of the

learned deep radiomic features can be helpful for interpreting their potential biological meaning. According to the mathematical structure of our model, we know that the kernels created by the convolutional operations can be treated as different filtrations of an individual image, which means the kernels of the same image can highlight different information the image contains. Therefore, it would be more understandable if we plot the deep radiomic features from the same kernel into a map which has consistent position information as the original image. We employed the Heatmap function[68] to create the kernel-wise map for each of our 16 kernels for the 52,238 images with different colors representing the magnitude of the feature values while position associating to the original coordinates of the images.

3.4 Unsupervised Clustering and Visualization

Heatmap3 [69] and t-Distributed Stochastic Neighbor Embedding (t-SNE)[70] were performed to visualize and cluster the normalized deep radiomic features in an unsupervised way. Complete linkage function in the hierarchical clustering process and visual-guided criteria by analysis of the dendrogram were used to decide the number of clusters in the heatmap.

t-SNE is a prize-winning nonlinear dimensionality reduction technique. It has been validated to have better performance in capturing the structure of high-dimensional data[71]. Traditional linear dimensionality reduction methods such as PCA (principal component analysis) always focus on keeping the dissimilar data points far away from each other. But in high dimensional case, besides emphasizing the difference, we also want to keep the similar data points as close to each other as possible. Therefore, nonlinear dimensionality methods are required to deal with high dimensional data. Among all the existing nonlinear dimensionality reduction methods, t-SNE is an outstanding one because of its ability to capture both local and global structures simultaneously. To be brief, the algorithm first defines a group of neighbors, and then gives higher weights to the neighbors within the same group and lower weights to the points out of the group. There is a hyperparameter *Perp* (perplexity) in the algorithm, which is used to adjust the effect of the number of neighbors on the final results. Usually, the hyperparameter *Perp* is chosen from 5 to 50[71].

We performed t-SNE clustering of the deep radiomic features at both patient and image levels with the perplexity set as 5 and 50, respectively. Patient-level features were calculated as the mean of image-level features. Because we performed the clustering of the images and patients using the deep radiomic features in an unsupervised way, there were no colors representing the category differences of the images and patients in the original t-SNE maps.

However, after the patient-level t-SNE map was generated, colors were manually assigned to the clusters based on the patients' clinical characteristics. In this way, we assigned the same colors to the image-level t-SNE map.

3.5 Classification of Clinical Characteristics Using Deep Radiomic Features

We performed supervised classification analysis using the learned deep radiomic features to predict the status of the 5 clinical characteristics (ER status, PR status, HER status, T and N) at image level. Since there are 4,095 radiomic features as predictors in the classification model, overfitting is likely to occur. We built a least absolute shrinkage and selection operator (LASSO) regression model using the R packages biglasso[72]. LASSO is a regularization technique that can be added into the fitting process of a linear regression model to reduce the magnitude of coefficients so that overfitting could be avoided. The formula of the multiple linear regression is shown in the Equation **3-4**.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \qquad i=1, 2, \ldots, N \qquad \textbf{(3-4)}$$

Here *X*s are the 4,096 deep radiomic features, *Y* is a given clinical characteristic. *N* is the sample size, while *p* is the number of deep radiomic features. Generally speaking, to fit the multiple linear regression model, we need to find the best βs to minimize the sums of squares error between regressed values and the true values of *Y*. This is called least squares, which is the most commonly used approach to approximate the coefficients. Since there are too many predictors in the model, it's easy to overfit the model if we include all of them in the model. The idea of LASSO is to add a penalty term into the least squares as shown below.

$$\hat{\beta}^{lasso} = arg\min_{\beta} \frac{1}{N} \sum_{i=1}^{N} \left(-y_i \log(\beta_j x_i) - (1-y_i)\log(1-\beta_j x_i)\right) + \lambda \sum_{j=1}^{p} |\beta_j| \qquad \textbf{(3-5)}$$

Here x is a vector of radiomic features. *y* is a given clinical characteristic. *N* is the sample size. *p* is the number of the radiomic features in the feature vector. λ is a hyperparameter used to control the level of penalty. The consequence is that as λ increases, some of the βs would decrease to 0, which would result in some of the features being removed from the linear

regression model to avoid the overfitting issue and select the important features used to fit the model[72].

Models were trained on a randomly selected sample set with 70% of the total samples and performance was evaluated using a test set with the remaining 30% of the total samples. 100 λs were tried and the performances of the models with different λs were measured using a metric called AUC which is a short term of the area under the receiver operating characteristic (ROC) curve. To understand AUC and ROC better, we need to look at the confusion matrix (**Table 3-2**), which is one of the efficient methods to describe the performance of a classifier.

Table 3-2: Confusion matrix to describe classification performance

		True class	
		positive	negative
Predicted class	positive	True positive (TP)	False positive (FP)
	negative	False negative (FN)	True negative (TN)

Using the confusion matrix, some classification performance measures can be defined, which include sensitivity or true positive rate (TP/(TP+FN)), specificity (TN/(TN+FP)), and false positive rate (FPR=FP/the actual number of negative samples tested). ROC curve is a line plotted in a space with false positive rate (FPR) as x-axis and true positive rate (TPR) as y-axis, while AUC is the area under the ROC curve (**Figure 3-7**). An AUC close to 1 means the model performs very well in classifying the test samples, while a smaller AUC value usually indicates a failure model. The R packages ROCR[68] and MLmetrics[68] were used to calculate the AUC values for our LASSO models with different λs.

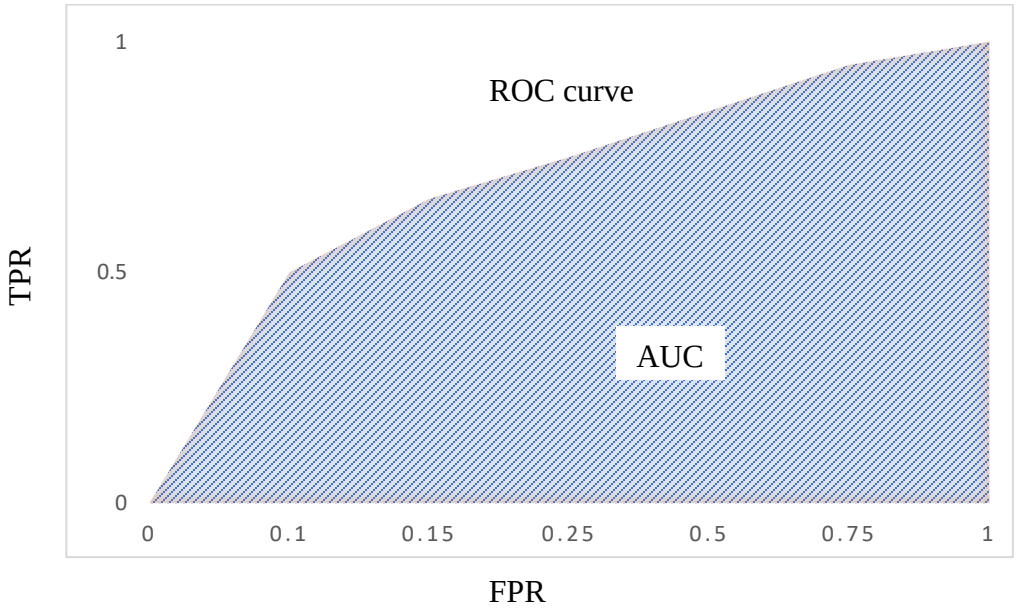


Figure 3-7: The relationship of ROC curve and AUC.

3.6 Association Analysis of Genomic Features and Deep Radiomic Features

To evaluate the relationship between the genomic features (breast cancer risk genes, breast cancer gene signatures and KEGG pathways) and the deep radiomic features from our learned deep learning model, we performed association analyses between each of the 4,096 deep radiomic features and each of those genomic features using a Linear Mixed Effect (LME) Model, which can model and analyze the complex and structured data with multi-levels [75]. In our case, multiple images can be obtained from each individual patient. We implemented the analysis using the R package nlme[76]. The formula of the LME model is as follows.

$$X_i = \beta_0 + \beta_1 G_i + \boldsymbol{\mu} Z_i \qquad i=1, 2, \ldots, N \qquad \textbf{(3-6)}$$

Here X is a given deep radiomic feature, G is a given genomic feature and i is the i^{th} images. Since our deep radiomic features are at image-level while each of the 110 patients has multiple images with a total of 52,238 images or samples. Hence, the deep radiomic features were not independent of each other. However, the genomic features are at patient-level, which has only 110 patients or samples **(Figure 3-8**).

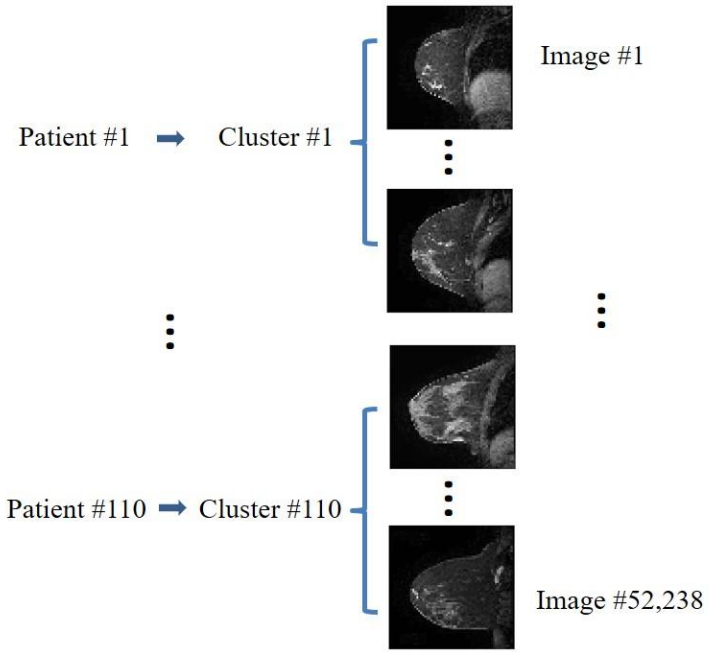


Figure 3-8: Non-independent issue for image-level deep radiomic features. Sample size doesn't match between genomic features (110 samples) and radiomic features (52,238 samples). The observations of image-level deep radiomic features are not independent, as within the same patient, these features are more similar.

In order to match the sample size of the response variable *X* and the predictor variable *G*, the genomic feature *G* have to be repeated multiple times based on the number of images a given patient has. Therefore, to take the repeated measurements of the genomic features into account and to address the effect caused by the dependence of the deep radiomic features, a random effect term $\boldsymbol{\mu Z}$ was introduced to the formula to simulate the variations coming from the patient differences where *Z* stands for patient ID.

In the association analysis, we usually need a statistical hypothesis: H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$. Through this hypothesis, we want to test whether the predictor variable can significantly explain the observed variability in response? To test the hypothesis, we would calculate a P-value to decide whether the null hypothesis H_0 should be rejected or not. Usually, a significance cutoff should be selected, and a P-value less than this cutoff means a rejection of the null hypothesis. In this case we could say the H1 hypothesis $\beta_1 \neq 0$ is true. In our case, if H_1 hypothesis is true, then the genomic feature being tested is associated with the given deep radiomic feature.

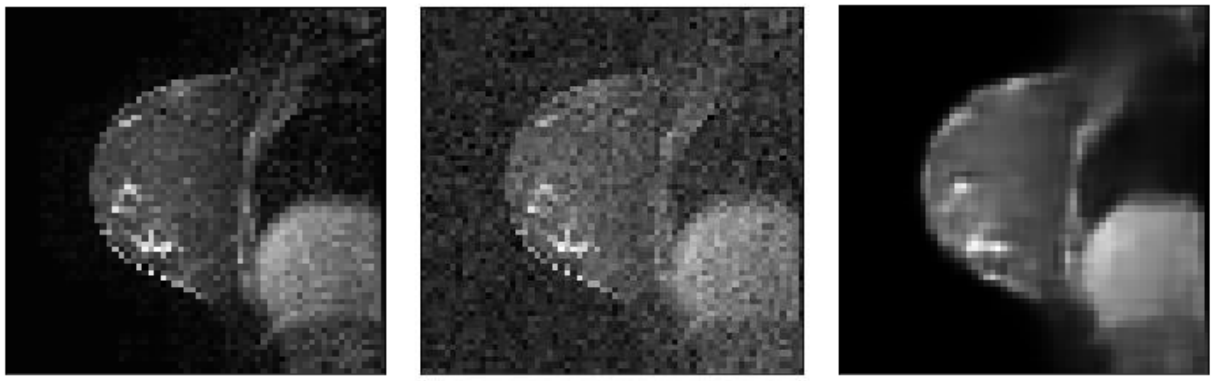
Since a large number of such hypotheses were tested, multiple testing correction was performed to calculate more strict P-values using Benjamin and Hochberg multiple testing procedure[77]. Significant genes, signatures and pathways were selected based on the adjusted P-values (false discovery rate < 0.05).

4 Chapter 4: Results and Discussions

4.1 Visualization and Unsupervised Clustering of Deep Radiomic Features

4.1.1 Generation of Deep Radiomic Features

At a noise level 5%, we trained the DA model and generated the denoised images. **Figure 4-1** shows one of the denoised images (**C**) and its raw (**A**) and noised images (**B**), respectively. The denoised image showed a better resolution than its original image, suggesting that the built DA model can learn deep radiomic features with better representation ability than original radiomic features.



A. raw image B. noised image C. denoised image

Figure 4-1: Image data after different processing steps. An example of reshaped 64×64 pixels MR image with noise (B) and without noise (A) as well as its denoised counterpart (C) after applying the proposed deep learning model.

For the learned deep radiomic features, we further processed them at sample-level using quantile normalization. **Figure 4-2** showed the density distribution of the 4,096 deep radiomic features in one randomly selected sample before (**A**) and after (**B**) the normalization. It can be

seen that before the sample-wise normalization, the distribution of the 4,096 feature values from the selected sample showed 2 or 3 peaks in the density plot (**Figure 4-2A**), while the distribution of the feature values didn't change much after the normalization, but still kept the similar shape in the density plot, especially in the low-value and high-value regions (**Figure 4-2B**). However, the normalization enhanced the contrast of the middle-value region of the learned deep radiomic features in the density plot, making it show a density distribution with three clear peaks that indicate the bright, grey and dark regions in the image. After the quantile normalization, the features are more comparable across samples[67].

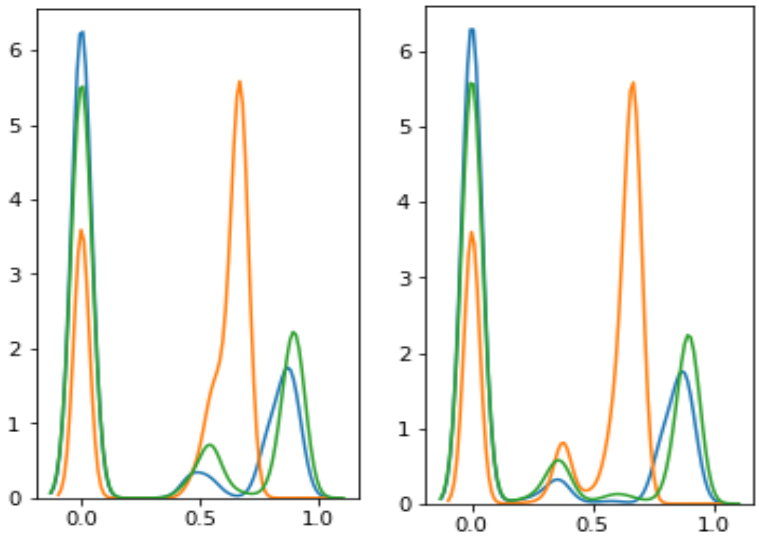
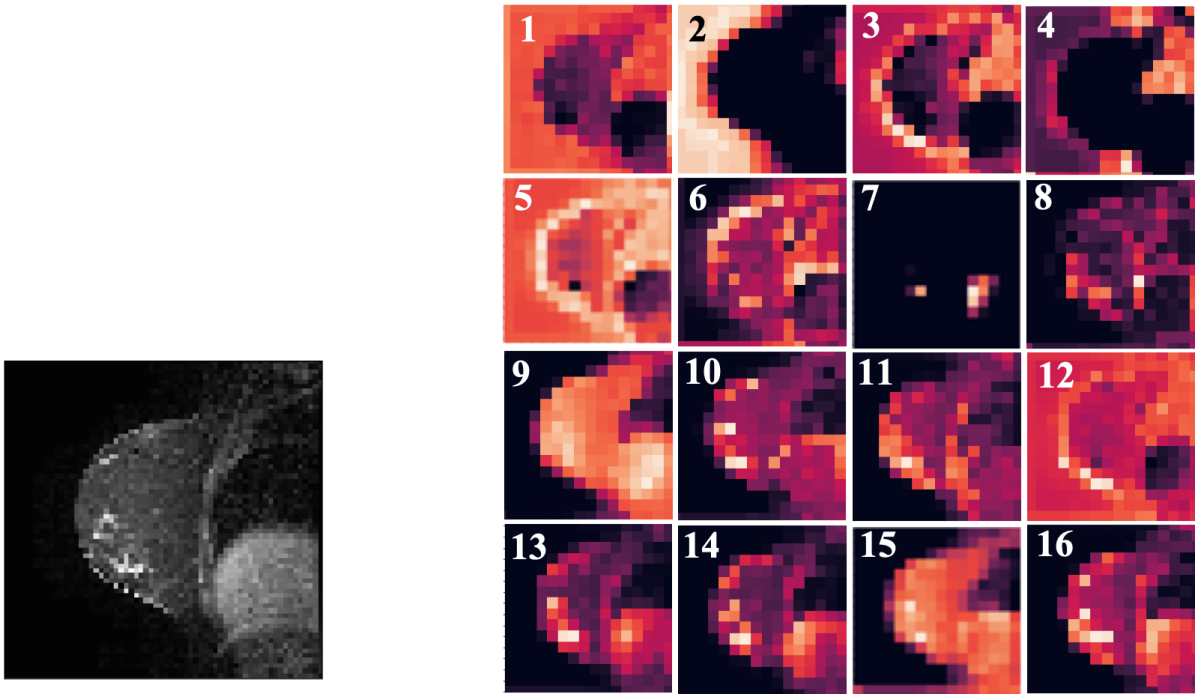


Figure 4-2: Density of auto-image features of the first 3 samples before and after quantile normalization. A: Density distribution before quantile normalization. B: Density distribution after quantile normalization.

4.1.2 Visualization of Deep Radiomic Features

The normalized deep radiomic features were plotted in a heatmap way so that they could be visualized at kernel-wise. **Figure 4-3** showed the kernel-wise normalized deep radiomic features of one randomly selected sample (**B**) and the sample's raw image (**A**). The 16 kernels have learned different information from the original image. Some kernels highlighted the tumor edge while others showed the high pixel value regions. For example, kernel #5 highlighted the edge of breast cancer, kernel #12 emphasized the breast edge close to the tumor region. Kernel #7 only showed the high-density regions (the tumor and diaphragm regions) of the raw image. More interestingly, almost half of the heatmaps (kernel # 9, 10, 11, 13, 14, 15, 16) emphasized the tumor regions. It has to be noticed that all these deep radiomic features are just numeric values. Therefore, if we just treat the values of these deep radiomic features as pixel values and plot them in a positional-based image, some may not have a biologically meaningful indication. However, as shown in the figure, we can still get some biological insight as to what these deep radiomic features represent. Further discussion about these radiomic features can be found in Section **4.3.**



A B

Figure 4-3: Potential biologic interpretation of kernel-level deep radiomic features. A: A randomly selected raw image as an example. This image is in sagittal view of the body. B: The kernel-wise heatmaps of the deep learning-based radiomic features of the same image shown in A.

4.1.3 Clustering of Deep Radiomic Features

The result of hierarchical clustering of the normalized deep radiomic features is shown in **Figure 4-4**. Patients are clustered into roughly 2 groups. The sizes of these two clusters are not balanced, in that one has only 14 patients while the other has 96 patients. However, according to the sidebar labels, these two clusters do not enrich any of the five clinical characteristics (Fisher's exact test, P-value>0.05).

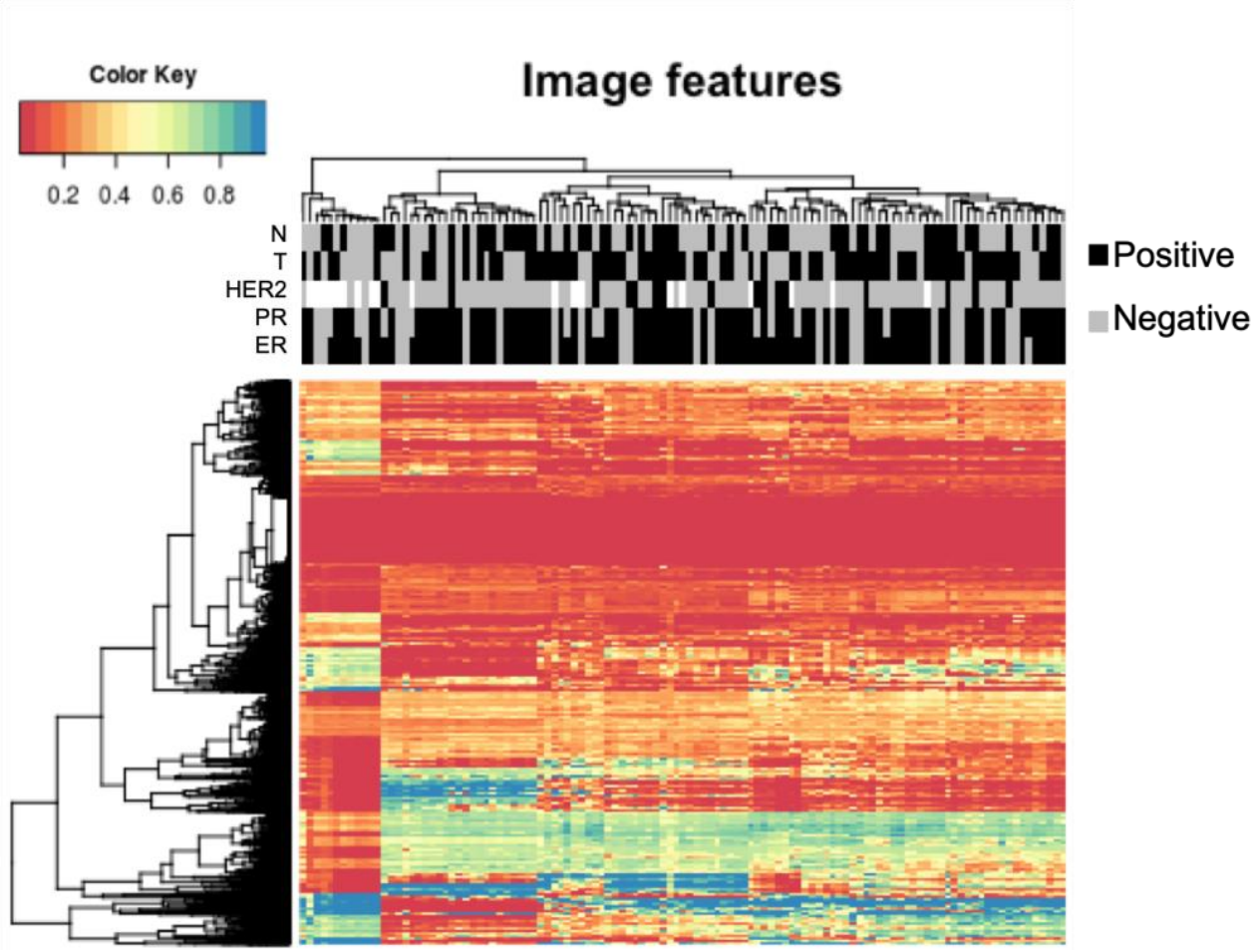
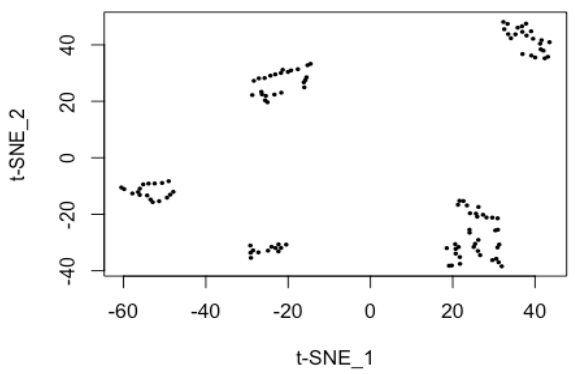
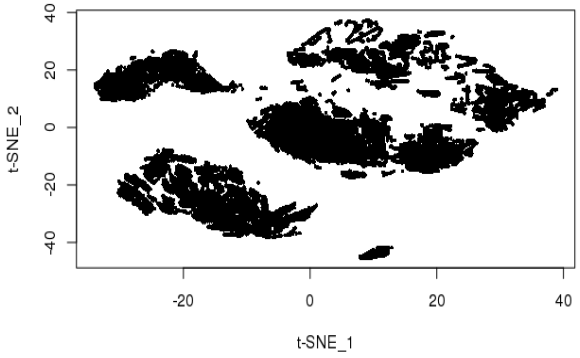


Figure 4-4: Unsupervised clustering of the deep radiomic features using hierarchical analysis function. Columns are the 110 patients; rows are the 4,096 deep radiomic features. Clinical information is shown in the sidebar. T refers to the tumor size. For breast tumors, bigger than 2cm are considered to be T-positive. N refers to node status, which is considered to be positive when the tumor cell spreads into lymph nodes. ER, PR, HER2 refer to estrogen receptor status, progesterone receptor status, and human epidermal growth factor receptor 2 status. Patients seem to be clustered into 2 groups, but these two groups have no obvious clinical difference.

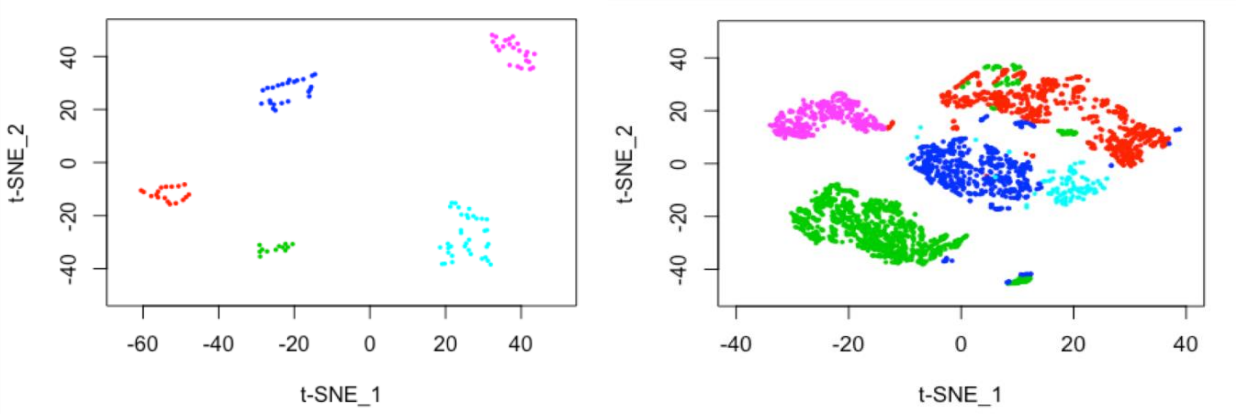
Besides the unsupervised hierarchical clustering, we also performed unsupervised t-SNE clustering of the normalized deep radiomic features at both patient-level (**Figure 4-5A**) and image-level (**Figure 4-5B**). In the patient-level t-SNE map, patients are clearly clustered into 5 groups, but these groups show no clinical difference (**Figure 4-6**). We first manually colored the 5 groups in the patient-level t-SNE map (**Figure 4-5C**), then tracked the dots in image-level t-SNE map to patient-level, and finally colored the dots in the image level map using the same colors as what we used in coloring patient-level t-SNE map (**Figure 4-5D**). The clustering patterns at both patient-level and image-level are consistent, indicating the robustness of the 5 patient clusters or groups. However, similar to the 2-groups pattern in the hierarchical clustering, the 5-groups do not have explainable clinical difference (**Figure 4-6**). The biological or clinical meaning of these clusters needs further investigation in other cohorts with deep phenotype information.



A



B



C D

Figure 4-5: t-SNE visualizes the deep radiomic features. A: Clustering results at patient level. Each dot is one patient. B: Clustering results at image level. Each dot is one image. C: Different colors are marked on different patient-level clusters manually. D: We first tracked the dots in image-level t-SNE map to patient-level, and then colored them using the same colors as what we used in coloring patient-level t-SNE map.

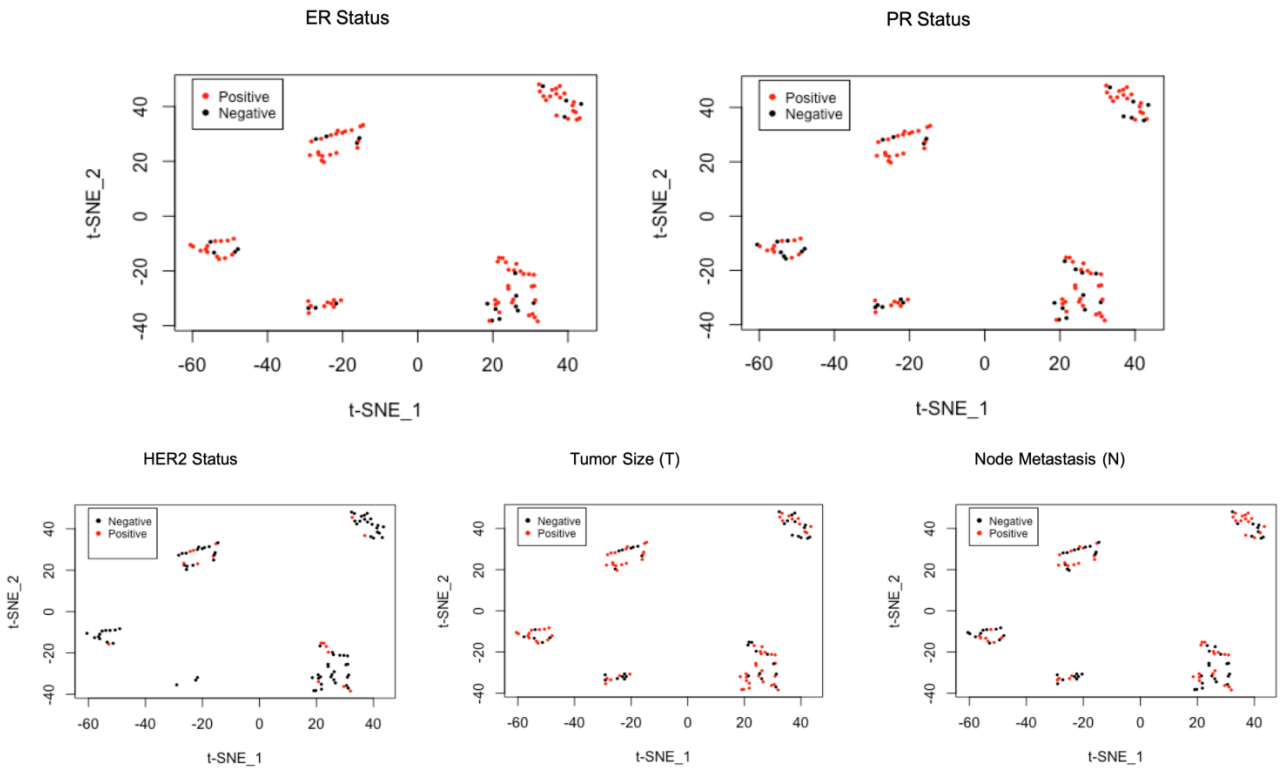


Figure 4-6: The t-SNE maps colored by different clinical characteristics. The 5 clusters showed no obvious correlation with clinical features.

4.2 Classification of Clinical Characteristics Using the Learned Deep Radiomic Features

As the hyperparameter λ increases, different numbers of deep radiomic features remain in the LASSO regression model to predict the clinical characteristics. **Figure 4-7** shows the prediction performance measured by AUC under different number of selected deep radiomic features. As we can see from the figure, using the LASSO models with proper λ, the deep radiomic features performed very well in predicting the 5 clinical characteristics. For example, the AUC could reach 90% or larger using 500 selected deep radiomic features in the LASSO models.

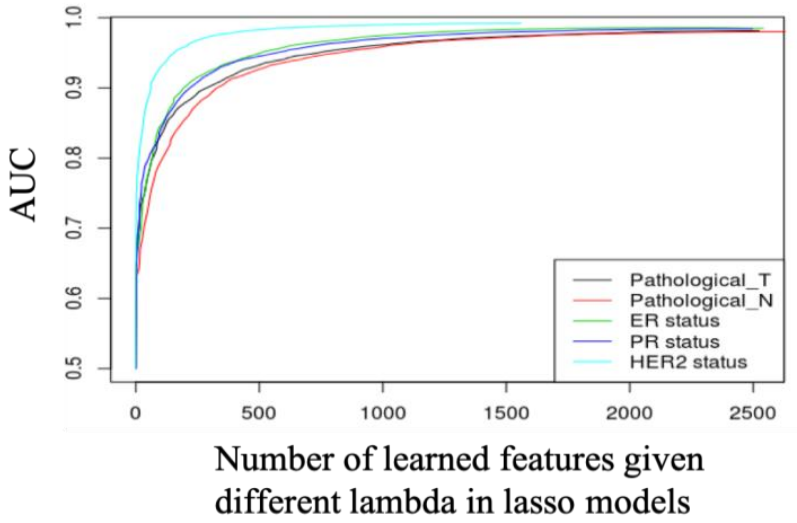


Figure 4-7: Performance of using deep radiomic features to predict clinical characteristics. Different colors represent different clinical characteristics. The x-axis represents the number of deep radiomics features given different $\boldsymbol{\lambda}$ in the LASSO models. The y-axis represents the corresponding area under the curve (AUC) which is a metric used to assess the performance of the prediction.

4.3 Evaluation Relationship of Genomic Features and Deep Radiomic Features

Using the LME model, we evaluated the association between each pair of the deep radiomic features and the genomic features and the results are shown in **Figure 4-8**. After multiple testing correction, 1,774 out of the 4,096 deep radiomic features are significantly associated with 213 of the 288 breast risk genes (**Figure 4-8A**). Two of the six gene expression signatures are significantly associated with the deep radiomic features. EndoPredict and Prosigna (rorS) scores are significantly associated with 848 and 1,395 deep radiomic features (**Figure 4-8B**), respectively. EndoPredict is a novel gene expression signature predicting the likelihood of distant recurrence in patients with estrogen ER-positive, HER2-negative breast cancer treated with adjuvant endocrine therapy, and it has been validated by clinical trials [78]. Moreover, EndoPredict has been approved by the US Food and Drug Administration (FDA) in 2014 and has been considered to be better than the commonly used Oncotype DX test in guiding chemotherapy decision-making[79]. Currently, the cost of a EndoPredict test is around1,500USD [79]. Prosigna (rorS) score is calculated from PAM50 intrinsic subtypes, correlation between molecular subtypes and a subset of proliferative genes, and it has also been cleared by the FDA in 2013 for marketing as a prognostic tool[80, 81]. The price is around 2,000USD for testing the intrinsic subtypes and Prosigna (rorS) score[17]. Furthermore, 1,739 out of the 4,096 deep radiomic features are significantly associated with 166 out of the 182 KEGG pathways (**Figure 4-8C**).

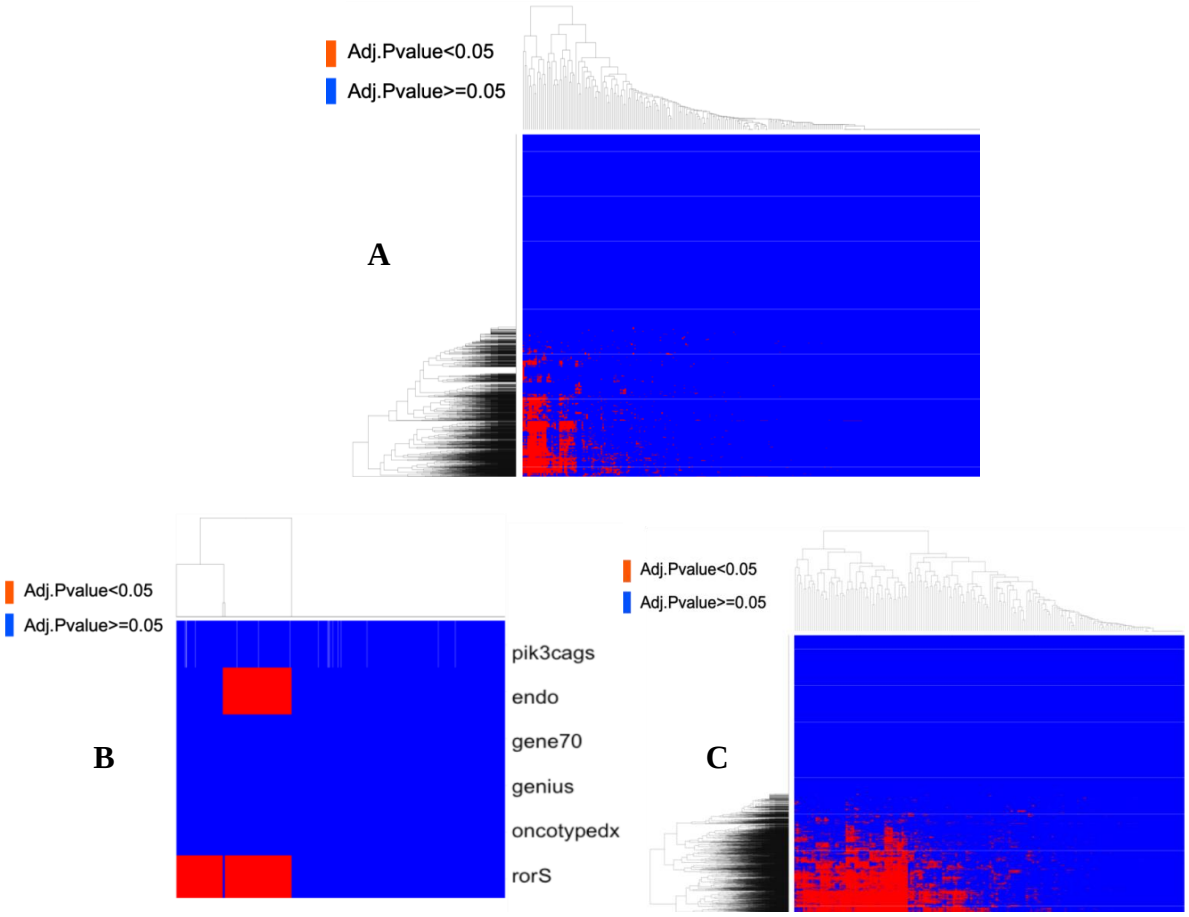


Figure 4-8: Adjusted P-values of the association analyses between deep radiomic features and genomic features. The X-axis is the 4,096 image features. The Y-axis is the genomic features. A: 288 breast cancer risk genes. B: 6 breast gene signatures. C: 182 KEGG pathway activity scores.

In total there are 2,028 of the 4,096 deep radiomic features significantly associated with 397 (213 risk genes + 2 gene signatures + 166 biological pathways = 381) genomic features. The details of the top 50 deep radiomic features are shown in the **Table 4-1**. Take the top 1 deep radiomic feature in the first row as an example, "fea_4043" is its feature ID, "55" in the first cell means that the radiomic feature "fea_4043" is significantly associated with 55 breast cancer risk genes. "2" in the second cell indicates that "fea_4043" is significantly associated with all the two

gene signatures. “89” in the third cell refers to 89 KEGG pathways are significantly associated with the radiomic feature “fea_4043”. Hence, the total number of significant genomic features that are associated with the “fea_4043” is 146. This means that the radiomic fea_4043 is associated with greatest number of genomic features analyzed in the current study. This radiomic feature belongs to kernel #16.

Table 4-1: The 50 most frequently genomic associated radiomic features. Rows are the top 50 deep radiomic features that are significantly associated with the one or more genomic features (ranked by the frequency of the associated genomic features). The first three columns are the number of significant associations between the given deep radiomic feature and the three levels of genomic features. The fourth column is the accumulated number of significant associations the given radiomic feature has. The last column is the kernel where the given radiomic feature is.

	risk_genes	gene_signatures	pathways	frequency	Kernel No.
fea_4043	55	2	89	146	16
fea_3782	54	2	87	143	15
fea_3494	53	2	87	142	14
fea_3750	49	2	91	142	15
fea_3238	52	2	87	141	13
fea_3488	53	2	86	141	14
fea_3760	53	2	86	141	15
fea_4015	50	2	89	141	16
fea_3510	50	2	88	140	14
fea_3222	48	2	89	139	13
fea_4002	49	2	88	139	16
fea_3232	47	2	89	138	13
fea_3451	45	2	90	137	14
fea_3723	44	2	90	136	15
fea_4034	48	2	86	136	16
fea_3254	47	2	86	135	13
fea_3520	48	2	85	135	14
fea_3766	46	2	87	135	15
fea_3478	46	2	86	134	14
fea_4031	46	2	86	134	16
fea_4047	46	2	86	134	16
fea_3216	44	2	87	133	13
fea_3240	46	2	85	133	13
fea_3504	46	2	85	133	14

fea_3526	47	2	84	133	14
fea_4012	45	2	86	133	16
fea_4026	47	2	84	133	16
fea_4063	46	2	85	133	16
fea_4058	47	2	83	132	16
fea_3270	48	2	81	131	13
fea_3464	45	2	84	131	14
fea_3752	41	2	88	131	15
fea_3794	39	2	90	131	15
fea_4010	40	2	89	131	16
fea_3462	46	2	82	130	14
fea_3472	45	2	83	130	14
fea_3480	43	2	85	130	14
fea_3744	47	2	81	130	15
fea_3756	45	2	83	130	15
fea_3776	48	2	80	130	15
fea_3802	49	2	79	130	15
fea_3999	36	2	92	130	16
fea_4042	46	2	82	130	16
fea_3206	42	2	85	129	13
fea_3500	43	2	84	129	14
fea_3754	42	2	85	129	15
fea_4018	44	2	83	129	16
fea_4050	41	2	86	129	16
fea_3248	43	2	83	128	13
fea_3264	49	2	77	128	13

As we can see from the **Table 4-1**, these deep radiomic features are mainly from kernel #13, 14, 15, and 16, which means they have more genomic associations. To make it clearer, we further calculated the frequency of the kernel-level deep radiomic features associated with the genomic features as shown in **Table 4-2**. This table was made based on all 2,028 significant deep radiomic features. A bar plot was also made to illustrate the frequency of the associations (**Figure 4-9**). As we can see, kernel #12, 13, 14, 15, and 16 are associated with the largest number of genomic features.

Table 4-2: Genomic association frequency of the kernel-level radiomic features. Rows are each radiomic feature kernel. The first three columns are the number of significant genomic features that are associated with the radiomic features mapped in the given kernel. The last column is the accumulated number of significant genomic features that are associated with the radiomic features within the given kernel. It should be noted that each kernel has 256 deep radiomic features.

kernel	risk_genes	gene_signatures	pathways	frequency
13	3251	227	7848	11326
14	3374	208	7423	11005
15	3039	190	6776	10005
16	3093	178	6650	9921
12	2510	203	6853	9566
11	2176	194	5934	8304
10	2045	185	5307	7537
9	1964	183	4706	6853
8	1852	163	3988	6003
7	1358	118	2752	4228
6	725	85	1601	2411
5	577	75	1399	2051
1	654	59	1158	1871
4	496	67	961	1524
2	358	56	890	1304
3	384	52	793	1229

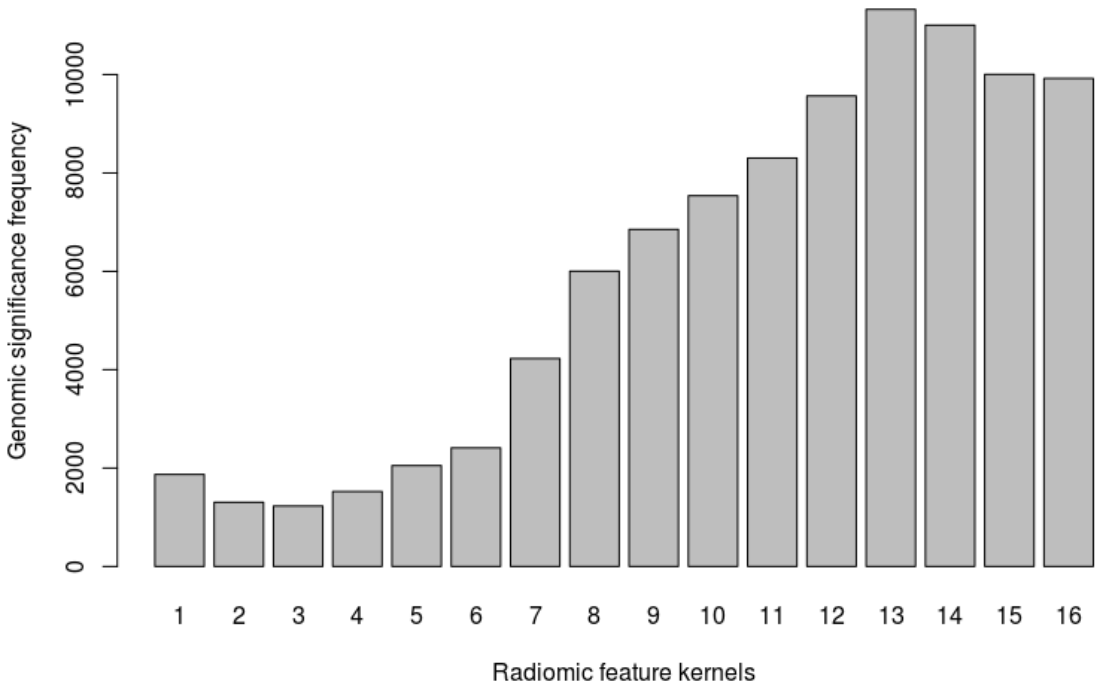


Figure 4-9: The genomic association frequency of the kernel-wise radiomic features

we also analysed the frequency of each of the genomic features associated with the number of deep radiomic features (**Table 4-3**). We did the analyses separately for the three different levels of genomic features, which included the top 5 significant breast cancer risk genes in the association tests (RP11-57H14.3, FIBP, ATP6AP1L, OVOL1, RP11-400F19.8), the 2 gene signatures (EndoPredict, Prosigna), and the top 5 significant KEGG pathways in the tests (KEGG_FATTY_ACID_METABOLISM, KEGG_INSULIN_SIGNALING_PATHWAY, KEGG_PHENYLALANINE_METABOLISM, KEGG_RNA_DEGRADATION, KEGG_TYROSINE_METABOLISM), which will be discussed in detail in next section.

Table 4-3: The genomic features that are associated with the largest number of radiomic features. We only reported the top five significant genomic features in each of the three categories (risk genes, gene signatures and pathways).

Gene signatures and pathways	genes	Number of significant radiomic features	Top 5 frequent radiomic feature kernels (ordered)
RP11-57H14.3		1118	13, 14, 15, 12, 16
FIBP		1050	13, 14, 15, 10, 11
ATP6AP1L		1038	13, 14, 16, 10, 15
OVOL1		1019	13, 14, 10, 16, 11
RP11-400F19.8		1017	13, 14, 16, 12, 15
EndoPredict	BIRC5, UBE2C, DHCR7, RBBP8, IL6ST, AZGP1, MGP, STC2, CALM2, OAZ1, RPL37A	848	13, 14, 12, 15, 11
Prosigna (rorS)	ACTR3B, ANLN, BAG1, BCL2, BIRC5, BLVRA, CCNB1, CCNE1, CDC20, CDC6, CDH3, CENPF, CEP55, CXXC5, EGFR, ERBB2, ESR1, EXO1, FGFR4, FOXA1, FOXC1, GPR160, GRB7, KIF2C, KRT14, KRT17, KRT5, MAPT, MDM2, MELK, MIA, MKI67, MLPH, MMP11, MYBL2, MYC, NAT1, NDC80, NUF2, ORC6L, PGR, PHGDH, PTTG1, RRM2, SFRP1, SLC39A6, TMEM45B, TYMS, UBE2C, UBE2T	1395	13, 14, 12, 11, 10
KEGG_FATTY_ACID_METABOLISM	ACAA1, ACAA2, ACADL, ACADM, ACADS, ACADSB, ACADVL, ACAT1, ACAT2, ACOX1, ACOX3, ACSL1, ACSL3, ACSL4, ACSL5, ACSL6, ADH1A, ADH1B, ADH1C, ADH4, ADH5, ADH6, ADH7, ALDH1B1, ALDH2, ALDH3A2, ALDH7A1, ALDH9A1, CPT1A, CPT1B, CPT1C, CPT2, CYP4A11, CYP4A22, ECHS1, ECI1, ECI2, EHHADH, GCDH, HADH, HADHA, HADHB	1269	12, 10, 11, 13, 9
KEGG_INSULIN_SIGNALING_PATHWAY	ACACA,ACACB,AKT1,AKT2,AKT3,ARAF,BAD,BRAF,CALM1,CALM2,CALM3,CALML3,CALML5,CALML6,CBL,CBLB,CBLC,CRK,CRKL,EIF4E,EIF4E1B,EIF4E2,EIF4EBP1,ELK1,EXOC7,FASN,FBP1,FBP2,FLOT1,FLOT2,FOXO1,G6PC,G6PC2,GCK,GRB2,GSK3B,GYS1,GYS2,HK1,HK2,HK3,HRAS,IKBKB,INPP5D,INPP5K,INS,INSR,IRS1,IRS2,IRS4,KRAS,LIPE,MAP2K1,MAP2K2,MAPK1,MAPK10,MAPK3,MAPK8,MAPK9,MKNK1,MKNK2,MTOR,NRAS,PCK1,PCK2,PDE3A,PDE3B,PDPK1,PHKA1,PHKA2,PHKB,PHKG1,PHKG2,PIK3CA,PIK3CB,PIK3CD,PIK3CG,PIK3R1,PIK3R2,PIK3R3,PIK3R5,PKLR,PPARGC1A,PPP1CA,PPP1CB,PPP1CC,PPP1R3A,PPP1R3B,PPP1R3C,PPP1R3D,PRKAA1,PRKAA2,PRKAB1,PRKAB2,PRKACA,PRKACB,PRKACG,PRKAG1,PRKAG2,PRKAG3,PRKAR1A,PRKAR1B,PRKAR2A,PRKAR2B,PRKCI,PRKCZ,PRKX,PTPN1,PTPRF,PYGB,PYGL,PYGM,RAF1,RAPGEF1,RHEB,RHOQ,RPS6,RPS6KB1,RPS6KB2,RPTOR,SH2B2,SHC1,SHC2,SHC3,SHC4,SLC2A4,SOCS1,SOCS2,SOCS3,SOCS4,SORBS1,SOS1,SOS2,SREBF1,TRIP10,TSC1,TSC2	1243	13, 12, 11, 14, 10
KEGG_PHENYLALANINE_METABOLISM	ALDH1A3,ALDH3A1,ALDH3B1,ALDH3B2,AOC2,AOC3,DDC,GOT1,GOT2,HPD,IL4I1,MAOA,MAOB,MIF,NAT6,PAH,PRDX6,TAT	1217	12, 13, 10, 9, 11
KEGG_RNA_DEGRADATION	C1D,C1DP2,C1DP3,CNOT1,CNOT10,CNOT2,CNOT3,CNOT4,CNOT6,CNOT6L,CNOT7,CNOT8,DCP1A,DCP1B,DCP2,DCPS,DDX6,DIS3,EDC3,EDC4,ENO1,ENO2,ENO3,EXOSC1,EXOSC10,EXOSC2,EXOSC3,EXOSC4,EXOSC5,EXOSC6,EXOSC7,EXOSC8,EXOSC9,HSPA9,HSPD1,LSM1,LSM2,LSM3,LSM4,LSM5,LSM6,LSM7,MPHOSPH6,NAA38,PAPD7,PAPOLA,PAPOLB,PAPOLG,PARN,PATL1,PNPT1,RQCD1,SKIV2L,SKIV2L2,TTC37,WDR61,XRN1,XRN2,ZCCHC7	1211	12, 13, 11, 14, 15
KEGG_TYROSINE_METABOLISM	ADH1A,ADH1B,ADH1C,ADH4,ADH5,ADH6,ADH7,ALDH1A3,ALDH3A1,ALDH3B1,ALDH3B2,AOC2,AOC3,AOX1,COMT,DBH,DCT,DDC,FAH,GOT1,GOT2,GSTZ1,HEMK1,HGD,HPD,IL4I1,LCMT1,LCMT2,MAOA,MAOB,METTL2B,METTL6,MIF,NAT6,PNMT,TAT,TH,TPO,TRMT11,TYR,TYRP1,WBSCR22	1205	12, 9, 10, 11, 13

4.4 Biological Explanations of the Associated Deep Radiomic Features

From **Table 4-2** and **Figure 4-9**, we can see that the radiomic feature kernel #12, 13, 14, 15, and 16 are associated with the largest number of genomic features. According to the selected heatmaps in **Figure 4-10**, kernel #12 is almost absolutely reversed the signals in the diaphragm and tumor regions(the tumor regions are bright in the raw image but are dark in kernel #12), but it has put a large weight on the bottle edge of the breast that is close to the tumor region, and there are some unclear patterns within the breast, chest, and lung regions. #13 emphasizes the tumor regions while lowers the values of other regions inside and outside the breast, but it has kept reasonable values for the diaphragm region. Kernels #14, 15, and 16 have the similar patterns as #13 but highlight different levels of the tumor regions. They both emphasize the tumor regions, but #15 puts the similar values to the breast and diaphragm and slight weaker values to the chest. Kernels #14 and #16 are almost the same as #13 with a dimming in breast region. Among these top 5 deep radiomic feature kernels, #12 is special as it emphasizes the edge information. While kernel #13 to #16 more focus on the tumors and tissues around the tumors. Kernel #13, 14, 15, and 16 are considered as information rich kernels. Comparing with the first several kernels (e.g. #1 to #12), they all learned more abstract and representative information from their original images, which captured the tumor regions, put different weights to the tissues around tumors, and partially kept the signals for other tissues far away from tumors. Interestingly, Kernel #7 is not standing out in our genomic-radiomic association analyses. However, from the kernel-level heatmaps, we can see it uniquely filtered out all of other parts but kept and highlighted the high-density regions, which fortunately contained the tumors. So, kernel #7 is actually a kind of a segment of the tumor. Since it has no significant associations for most of our genomic and deep radiogenomic features, this makes us believe that not only the

tumor regions matter, the tissues surround them also play a role in reflecting the genomic-level information in the images. Therefore, we have reasons to believe that the traditional segmentation-based feature extraction methods might result in losing a lot of meaningful molecular information. Kernel #5 is worthy of being discussed as well, because it puts a lot of weight on the edge of breast but generally ignores the signals in the tumor regions. Majority of the deep radiomic features in the kernel has no association with the genomic information. On the contrary, majority of the deep radiomic features in the kernel #12, which also lowers the signals of tumors and emphasizes the edge of breast, showed significant association with the genomic information. The difference is that kernel #5 emphasizes all the breast edge without any bias, while kernel #12 only emphasizes the edge close to the tumor regions. Hence, we believe that the edge of breast, which is really close to the tumor regions, might contain meaningful genomic information that is worthy of being further studied.

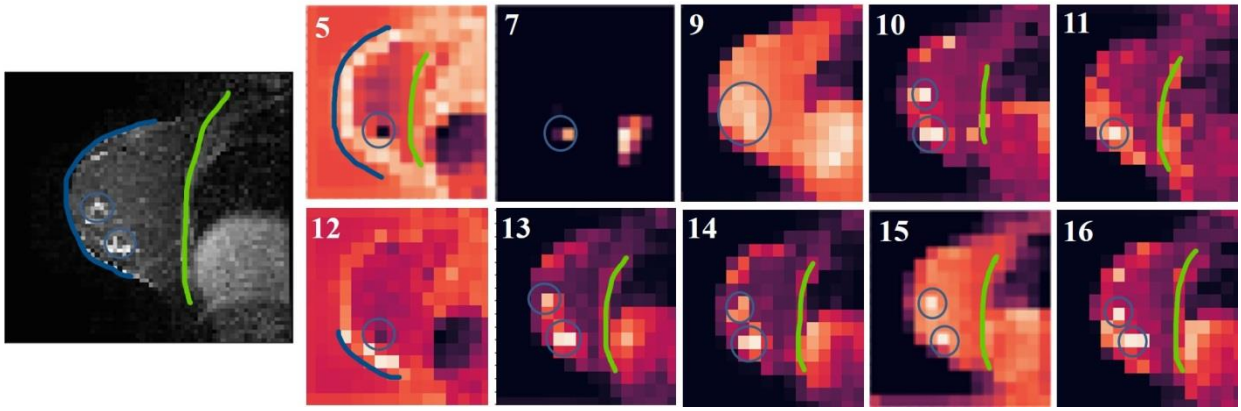


Figure 4-10: Selected kernel-level radiomic features. Kernel #5 lowers the signal of tumor regions (blue circle) but highlights the edge of the whole breast. The signal of chest (green curve) is a little bit emphasized as well in this kernel. The kernel #12 looks like the kernel #5, but it puts more weight on the breast edge that is close to the tumor regions. The Kernel #7 keeps only the high-density regions

including tumor regions. It is kind of a segmentation of tumors since the signals of the tumors surrounding tissues are all decreased to 0 (black). Kernel #9 is a smooth kernel with a blurred contour of tumor regions. Kernel #10, 11, 12, 14, and 16 are information-rich kernels. They all highlight the tumor regions and the chest. The kernel #15 looks like those information-rich kernels but with lower values for the chest.

As shown in the **Table 4-3**, 5 breast cancer risk genes are standing out, which are RP11-57H14.3, FIBP, ATP6AP1L, OVOL1, and RP11-400F19.8. The RP11-57H14.3 and the RP11-400F19.8 are classed as the processed transcript biotype. They do not code proteins and their biological functions are not clear[82, 83]. However, they were observed in several cancer related studies [84, 85]. Interestingly, these genes are highly associated with our top 5 radiomic feature kernels. FIBP, ATP6AP1L, and OVOL1 are protein-coding genes and their products are acidic fibroblast growth factor intracellular-binding protein[86], subunit ATPase[87], and a zinc finger protein[88] respectively. The acidic fibroblast growth factor, intracellular-binding protein and the subunit ATPase coded by FIBP and ATP6AP1L are related to cell metabolism and growth. The zinc finger protein coded by OVOL gene could influence cell proliferation and malignant transformation by adjusting the MYC transcription, which is a well-known oncogene[89]. Comparing with RP11-57H14.3 and RP11-400F19.8, kernel #12 is not associated with the two protein-coding genes, which makes kernel #12 even more special. Hence, we suspect that RP11-57H14.3 and RP11-400F19.8 are related to the morphological changes of the skin that is close to the tumor regions.

EndoPredict and Prosigna are gene signatures predicting breast cancer recurrence risk, but they were designed based on different gene sets. Deeper radiomic features are associated with Progsigna (1,395) than EndoPredict (848). These deep radiomic features are mapped

commonly in kernels #11, 12, 13, and 14. However, Progsigna has a specific most significant kernel #10 while EndoPredict also has a unique most significant kernel #15. Kernel #11 is more similar to other information rich kernels (#13, 14, 15, and 16) because they all give higher values to the chest and lower values to the breast. However, kernel #15 shows reversal of the values of the chest and breast. Therefore, according to the evidence from the deep radiomic features, we infer that EndoPredict and Prosigna can capture the similar radiomic information but Progsigna might do a better job.

Several metabolism pathways stand out in the radiomic feature-based association analysis, such as KEGG_FATTY_ACID_METABOLISM, KEGG_PHENYLALANINE_METABOLISM, and KEGG_TYROSINE_METABOLISM pathways. The radiomic features associated with these pathways are mapped to the common kernels #10, 11, 12, 13. These pathways are all reported in breast or human reproductive cancers related studies[90–92]. Kernel #9 looks very smooth, and it is included in three of these metabolism pathways, which might give us a clue that the genomic features in systemic function-level capture more heavily on the blurred and rough information of the breast cancer MR images.

Although there are a lot of interesting findings in this study, the mechanism by which those genomic features are associated with certain deep radiomic feature kernels is still unclear. Currently, there is no similar research on the association analysis between the genomic features and the MRI-based radiomic features extracted from the deep learning “black-box”. There is a lack of good visualization tools and references between deep radiomic features and molecular biological markers, which should be overcome in the future studies.

5 Chapter 5: Significance, Limitations and Future Directions

5.1 Significance

A novel deep learning model with denoising autoencoders technique was developed to extract the deep radiomic features from breast cancer enhanced T1-weighted MR images. These deep radiomic features were further mapped to the genomic features from three different angles. Their capability of predicting the clinical characteristics of breast cancer was also explored.

The study contributed to improving the deep learning algorithm itself by testing and tuning the architectures and parameters of the model. However, the more important point is that the identification of the biologically relevant auto-archived radiomic biomarkers would support and improve MRI as an economical and effective clinical assessment tool. For example, a functional application can be packaged based on this algorithm, which could be installed in the MRI workstation to provide radiologists a real-time and "one-stop for everything" tool to assess the likely genomic profile and clinical characteristics of the patients. Furthermore, this workstation could also generate radiogenomic signatures like those currently commercialized gene signatures. These radiogenomic signatures could be used to predict the patient diagnosis, prognosis, and therapeutic benefits as well. Compared to using gene signatures, the radiogenomic signatures would be faster, cheaper, as well as containing both genomic and radiomic information.

Unlike handcraft features which are based on radiologists' prior knowledge, deep radiomic features from current deep learning models are automatically generated. These models are usually treated as black boxes, and hence, the results from the models are difficult for us to interpret their biological and/or clinical meanings. In order to be readily adopted by real-world clinical practices, deep learning models must be interpretable without sacrificing their prediction

accuracy. In this study, we explored some visualization techniques for this purpose and provided some likely interpretations, but more work is still needed.

5.2 Limitations

One limitation of the study is that we do not have valuable survival data of the patients, all of the 110 patients with MRI data are alive. Therefore, survival analysis cannot be performed, which prevented us from evaluating the ability of the identified deep radiomic features to be associated with the patients' survival. High-quality databases and further studies are needed to explore the problem in the future.

Due to the computational complexity to analyze both of the MRI and genomic data, we did not use all the available information for the current analysis. To control the time and cost, we had to sacrifice some information by compressing the MR images to a lower resolution and stratifying the genome-wide level gene expression data into 3 higher-order levels so that the computing is possible to carry out. More advanced parallel techniques should be developed and added to the current deep learning and statistical analysis algorithms to make the computing faster.

The more challenging part of this study is the biological interpretation of the deep radiomic features and the identified relationship of these deep radiomic features with the genomic feature. Although many studies have identified the genes and their biological functions, the relationship of the imaging phenotypes, especially the deep radiomic phonotypes extracted by the deep learning models, and these genes and biological functions are highly lack of evidence. Furthermore, because of the complexity of the mathematic operations in extracting these deep radiomic features, there are no suitable visualization tools to clearly display the information contained in the features in a human-understandable way, which makes our further exploration even harder.

5.3 Future Direction

One future direction is to develop more advanced interpretable and extendable deep learning algorithms for modeling multimodal medical imaging and genomic data to identify clinically relevant biomarkers. The interpretability should be evidenced in both statistic aspect and biologic aspect.

Another direction goes to the development of an integrating framework for effectively utilizing multiple data sources (including both multiple imaging and genomic sources), which can be extended in the future to incorporate wider data modalities. For example, we can integrate MR images, X-ray images, ultrasound images together to generate super powerful radiomic features. We can also integrate gene expression data, copy number alteration data, mutation data, protein data, epidemiology data together to identify more valuable genomic features. By doing this, the relationship of the medical images and their molecular biology meaning would be clearer and more exhaustive. Hence, the patients would benefit more from the routine imaging examination.

New visualization tools are urgently needed to understand these deep radiomic features. Techniques like deconvolution could be considered in the future [93].

Reference

1. Fass L. Imaging and cancer: A review. Mol Oncol. 2008;2:115–52.

2. Runge VM. Contrast-enhanced clinical magnetic resonance imaging. University Press of Kentucky; 2015.

3. Brown RW, Cheng Y, N H, E M. Magnetic Resonance Imaging : Second Edition.

4. Weishaupt D, Marincek B. How does MRI work? An Introduction to the Physics and Function of Magnetic Resonance Imaging. 2008. papers3://publication/uuid/F2EC282D-FF52-4C88-8720-3E66239857DE.

5. Healy ME, Hesselink JR, Press GA, Middleton MS. Increased detection of intracranial metastases with intravenous Gd-DTPA. Radiology. 1987;165:619–24. doi:10.1148/radiology.165.3.3317496.

6. Buckley DL, Roberts C, Parker GJM, Logue JP, Hutchinson CE. Prostate Cancer: Evaluation of Vascular Characteristics with Dynamic Contrast-enhanced T1-weighted MR Imaging— Initial Experience. Radiology. 2004;233:709–15.

7. Böttcher J, Hansch A, Pfeil A, Schmidt P, Malich A, Schneeweiss A, et al. Detection and classification of different liver lesions: Comparison of Gd-EOB-DTPA-enhanced MRI versus multiphasic spiral CT in a clinical single centre investigation. Eur J Radiol. 2013;82:1860–9. doi:10.1016/j.ejrad.2013.06.013.

8. Van Goethem M, Tjalma W, Schelfout K, Verslegers I, Biltjes I, Parizel P. Magnetic resonance imaging in breast cancer. Eur J Surg Oncol. 2006;32:901–10.

9. Gross JM, Yee D. How does the estrogen receptor work? Jennifer. Ann Ital Chir. 2002;4:62–4.

10. Paleri V, Mehanna H, Wight RG. TNM classification of malignant tumours 7th edition: what's new for head and neck? Clin Otolaryngol. 2010;35:270–2. doi:10.1111/j.1749-

4486.2010.02141.x.

11. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001;98:10869–74. doi:10.1073/pnas.191367098.

12. Jiang G, Zhang S, Yazdanparast A, Li M, Pawar AV, Liu Y, et al. Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics. 2016;17 Suppl 7. doi:10.1186/s12864-016-2911-z.

13. Baxter JS, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V, et al. Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. Nat Commun. 2018.

14. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat Genet. 2018.

15. Prosigna T, Cancer B, Gene P, Assay S, Ii S, Score TP, et al. What is Prosigna ™ ? Indications for Use Order Information : Please refer to your Providence Regional Laboratory Physician Sales & Service representative for additional information . Stage II.

16. Slodkowska EA, Ross JS. MammaPrint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. Expert Rev Mol Diagn. 2009;9:417–22. doi:10.1586/erm.09.32.

17. Gyorffy B, Hatzis C, Sanft T, Hofstatter E, Aktas B, Pusztai L. Multigene prognostic tests in breast cancer: Past, present, future. Breast Cancer Res. 2015;17:1–7.

18. Ohnstad HO, Borgen E, Falk RS, Lien TG, Aaserud M, Sveli MAT, et al. Prognostic value of PAM50 and risk of recurrence score in patients with early-stage breast cancer with long-term follow-up. Breast Cancer Res. 2017;19:1–12.

19. Biological Pathways Fact Sheet | NHGRI. https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet. Accessed 28 Jun 2019.

20. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research. 1999;27:29–34.

21. Sangsoo Lim, Sangseon Lee, Inuk Jung SR and SK, Corresponding. Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. Bioscience. 2018;00 November:1–11. doi:10.1093/biosci/biy138.

22. Radhakrishna S, S A, Purvish MP, K K, Shikha P, Shelly S, et al. Role of magnetic resonance imaging in breast cancer management. South Asian J Cancer. 2018;7:69–71.

23. Veltman J, Stoutjesdijk M, Mann R, Huisman HJ, Barentsz JO, Blickman JG, et al. Contrast-enhanced magnetic resonance imaging of the breast: The value of pharmacokinetic parameters derived from fast dynamic imaging during initial enhancement in classifying lesions. Eur Radiol. 2008;18:1123–33.

24. Catalano R, Winett L, Wallack L, Satariano W. Evaluating a Campaign to Detect Early Stage Breast Tumors in the United States. Eur J Epidemiol. 2003;18:545–50.

25. Peters GN, Jones RC. New Horizons in the Diagnosis and Treatment of Breast Cancer Using Magnetic Resonance Imaging. 1993;166 December:749–55.

26. Warner E, Plewes DB, Hill KA, Causer PA, Jong RA, Cutrara MR, et al. Surveillance of BRCA1 and BRCA2 Mutation Carriers With Magnetic Resonance Imaging, Ultrasound, Mammography, and Clinical Breast Examination. Jama. 2004;292:1317–25.

27. Kriege M, Brekelmans CTM, Boetes C, Besnard PE, Zonderland HM, Obdeijn IM, et al. Efficacy of MRI and Mammography for Breast-Cancer Screening in Women with a Familial or Genetic Predisposition. N Engl J Med. 2004;351:427–37. doi:10.1056/NEJMoa031759.

28. Giger ML, Karssemeijer N, Schnabel JA. Breast Image Analysis for Risk Assessment, Detection, Diagnosis, and Treatment of Cancer. Annu Rev Biomed Eng. 2013;15:327–57. doi:10.1146/annurev-bioeng-071812-152416.

29. Lambina P, Rios-Velazqueza E, Leijenaara R, Carvalhoa S, Stiphouta RGPM van, Grantona P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. 2012;48:441–6.

30. Zhu Y, Li H, Guo W, Drukker K, Lan L, Giger ML, et al. Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma. Sci Rep. 2015;5. doi:10.1038/srep17787.

31. Bourgier C, Colinge J, Ailleres N, Fenoglietto P, Brengues M, Pelegrin A, et al. Radiomics: definition and clinical development. Cancer Radiother J la Soc Fr Radiother Oncol. 2015;19:532–7.

32. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. Phys Med Biol. 2016;61:R150–66. doi:10.1088/0031-9155/61/13/R150.

33. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, et al. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. Sci Rep. 2017;7:1–8. doi:10.1038/s41598-017-10649-8.

34. van ’t Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6. doi:10.1038/415530a.

35. Zhu Y, Li H, Guo W, Drukker K, Lan L, Giger ML, et al. Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma. Sci Rep. 2015;5:17787.

36. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Adv Neural Inf Process Syst. 2012;:1–9.

37. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Thakurta DG, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet. 2005;37:710–7.

38. Tan J, Ung M, Cheng C, Greene CS. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. In: Pacific Symposium on Biocomputing Co-Chairs. World Scientific; 2014. p. 132–43.

39. Guyon I, Elisseeff A. Feature Extraction, Foundations and Applications: An introduction to feature extraction. Stud Fuzziness Soft Comput. 2006;207:1–25. doi:10.1007/978-3-540-35488-8_1.

40. Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. Sci Rep. 2017;7:1–12. doi:10.1038/s41598-017-05848-2.

41. Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. ACM; 2008. p. 1096–103. http://www.iro.umontreal.ca/~lisa/publications2/index.php/publications/show/217.

42. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045–57.

43. Yang ZH, Feng F, Wang XY. A guide to technique of magnetic resonance imaging. 2007.

44. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An

immeasurable source of knowledge. Wspolczesna Onkologia. 2015;1A:A68–77.

45. Zhu Y, Qiu P, Ji Y. TCGA-assembler: Open-source software for retrieving and processing TCGA data. Nature Methods. 2014;11:599–600.

46. Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 19 Jun 2019.

47. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

48. Deluca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. 2012;28:1530–2.

49. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9.

50. The ENCODE Project Consortium Summary. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature. 2012;489:57–74.

51. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11.

52. Gendoo DMA, Ratanasirigulchai N, Schröder MS, Paré L, Parker JS, Prat A, et al. Genefu: An R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2016.

53. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. N Engl J Med. 2004;351:2817–26. doi:10.1056/NEJMoa041588.

54. Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, et al. A New Molecular Predictor of Distant Recurrence in ER-Positive, HER2-Negative Breast Cancer Adds Independent Information to Conventional Clinical Risk Factors. Clin Cancer Res. 2011;17:6012–20. doi:10.1158/1078-0432.CCR-11-0926.

55. Mullins M, Nobel AB, Parker JS, Leung S, Perou CM, Voduc D, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J Clin Oncol. 2009;27:1160–7.

56. van ’t Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6. doi:10.1038/415530a.

57. Bontempi G, Sotiriou C, Haibe-Kains B, Rothé F, Piccart M, Desmedt C. A fuzzy gene expression-based computational approach improves breast cancer prognostication. Genome Biol. 2010;11:R18.

58. Loi S, Haibe-Kains B, Majjaj S, Lallemand F, Durbecq V, Larsimont D, et al. PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor – positive breast cancer. Proc Natl Acad Sci U S A. 2010;107:10208–13.

59. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009.

60. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet. 2006;38:500.

61. Chollet F. Building Autoencoders in Keras. The Keras Blog. 2016;:1–14.

62. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI ’16). 2016. p.

265–84. doi:10.1038/nn.3331.

63. Boureau Y-L, Ponce J, Fr JP, Lecun Y. A Theoretical Analysis of Feature Pooling in Visual Recognition. Icml. 2010;:111–8. doi:citeulike-article-id:8496352.

64. Chollet F. Keras: The Python Deep Learning library. keras.io. 2015.

65. Nair V, Conference GH-P of the 27th international, 2010 U. Rectified Linear Units Improve Restricted Boltzmann Machines. CsTorontoEdu. 2010;:6421113. doi:10.1.1.165.6419.

66. Ruder S. An overview of gradient descent optimization algorithms. 2016;:1–14. http://arxiv.org/abs/1609.04747.

67. Bm B, Ra I, Astr, M STP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19:185–93.

68. Waskom M, Botvinnik O, Hobson P, Cole JB, Halchenko Y, Hoyer S, et al. seaborn: v0.5.0 (November 2014). 2014. doi:10.5281/ZENODO.12710.

69. Zhao S, Guo Y, Sheng Q, Shyr Y. Advanced Heat Map and Clustering Analysis Using Heatmap3. Biomed Res Int. 2014.

70. Van Der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579–605.

71. Laurens van der Maaten, Geoffrey E. H. Visualizing Data using t-SNE. J Mach Learn Res. 2008;164:10.

72. Zeng Y, Breheny P. The biglasso package: A memory-and computation-efficient solver for lasso model fitting with big data in r. arXiv Prepr arXiv170105936. 2017.

73. Fraley C, Raftery AE, Murphy TB, Scrucca L. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Tech Rep. 2012;:1–50.

74. Friedman J, Hastie T, Tibshirani R. glmnet: Lasso and elastic-net regularized generalized linear models. R Packag version. 2009;1.

75. Davidian M, Giltinan DM. Nonlinear models for repeated measurement data. 2017.

76. Cayuela L. Modelos lineales mixtos en R. 2010.

77. Hochberg B. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. J R Stat Soc. 1995.

78. Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, et al. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res. 2011.

79. EndoPredict gene e expression xpression profiling assa assay y for assessing risk of breast cancer recurrence. 2015;:1–42.

80. Bernard PS, Parker JS, Mullins M, Cheung MCU, Leung S, Voduc D, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.

81. Fayanju OM, Park KU, Lucci A. Molecular Genomic Testing for Breast Cancer: Utility for Surgeons. Ann Surg Oncol. 2018;25:512–9. doi:10.1245/s10434-017-6254-z.

82. Gene: RP11-400F19.8 (ENSG00000266929) - Summary - Homo sapiens - GRCh37 Archive browser 96. http://grch37.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000266929;r=17:40688528-40714080;t=ENST00000585572. Accessed 30 Jun 2019.

83. Gene: RP11-57H14.3 (ENSG00000225292) - Summary - Homo sapiens - GRCh37 Archive browser 96. http://grch37.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000225292;r=10:114648494-114665870;t=ENST00000428766. Accessed 30 Jun 2019.

84. Hoff AM, Johannessen B, Alagaratnam S, Zhao S, Nome T, Løvf M, et al. Novel RNA variants in colorectal cancers. Oncotarget. 2015;6.

85. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat Genet. 2018;50:968–78.

86. KOLPAKOVA E, WIĘDŁOCHA A, STENMARK H, KLINGENBERG O, FALNES PØ, OLSNES S. Cloning of an intracellular protein that binds selectively to mitogenic acidic fibroblast growth factor. Biochem J. 2015;336:213–22.

87. Chen E, Zollo M, Mazzarella R, Ciccodicola A, Chen CN, Zuo L, et al. Long-range sequence analysis in Xq28: thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD loci. Hum Mol Genet. 1996;5:659–68. doi:10.1093/hmg/5.5.659.

88. Nair M, Teng A, Bilanchone V, Agrawal A, Li B, Dai X. Ovol1 regulates the growth arrest of embryonic epidermal progenitor cells and represses c-myc transcription. J Cell Biol. 2006;173:253–64.

89. Seeger RC, Brodeur GM, Sather H, Dalton A, Siegel SE, Wong KY, et al. Association of Multiple Copies of the N-myc Oncogene with Rapid Progression of Neuroblastomas. N Engl J Med. 1985;313:1111–6. doi:10.1056/NEJM198510313131802.

90. Myers JS. Florida State University Libraries Investigation of Human Prostate Cancer Through Experimental and Bioinformatics Study of Gene and Protein Expression. 2016.

91. Sugimoto M, Wong DT, Hirayama A, Soga T, Tomita M. Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. Metabolomics. 2010;6:78–95.

92. Cioce M, Valerio M, Casadei L, Pulito C, Sacconi A, Mori F, et al. Metformin-induced

metabolic reprogramming of chemoresistant ALDH^{bright} breast cancer cells. Oncotarget. 2014;5.

93. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. arXiv Prepr arXiv150606579. 2015.