Texture Analysis Platform for Imaging Biomarker Research

by

Sara Ranjbar

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2017 by the
Graduate Supervisory Committee:

David Kaufman, Co-Chair
J. Ross Mitchell, Co-Chair
George Runger

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

The rate of progress in improving survival of patients with solid tumors is slow due to late stage diagnosis and poor tumor characterization processes that fail to effectively reflect the nature of tumor before treatment or the subsequent change in its dynamics because of treatment. Further advancement of targeted therapies relies on advancements in biomarker research. In the context of solid tumors, bio-specimen samples such as biopsies serve as the main source of biomarkers used in the treatment and monitoring of cancer, even though biopsy samples are susceptible to sampling error and more importantly, are local and offer a narrow temporal scope.

Because of its established role in cancer care and its non-invasive nature imaging offers the potential to complement the findings of cancer biology. Over the past decade, a compelling body of literature has emerged suggesting a more pivotal role for imaging in the diagnosis, prognosis, and monitoring of diseases. These advances have facilitated the rise of an emerging practice known as Radiomics: the extraction and analysis of large numbers of quantitative features from medical images to improve disease characterization and prediction of outcome. It has been suggested that radiomics can contribute to biomarker discovery by detecting imaging traits that are complementary or interchangeable with other markers.

This thesis seeks further advancement of imaging biomarker discovery. This research unfolds over two aims: I) developing a comprehensive methodological pipeline for converting diagnostic imaging data into mineable sources of information, and II) investigating the utility of imaging data in clinical diagnostic applications. Four validation studies were conducted using the radiomics pipeline developed in aim I. These studies had the following goals: (1 distinguishing between benign and malignant head and neck lesions (2) differentiating benign and malignant breast cancers, (3) predicting the status of Human Papillomavirus in head and neck cancers, and (4) predicting neuropsychological performances as they relate to Alzheimer's disease progression. The long-term objective of this thesis is to improve patient

outcome and survival by facilitating incorporation of routine care imaging data into decision making processes.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Ross Mitchell, for his support and guidance through the course of my PhD. Dr. Mitchell, your guidance, mentorship, and encouragement were the light that guided me throughout the difficult times of my journey. Thank you for all the opportunities that you created to advance my learning. Dr. David Kaufman, your guidance and suggestions were instrumental to my course of study and I wish to offer you my sincerest gratitude. A special thanks to Dr. George Runger, your continuous support as the chair of biomedical informatics department and guidance and suggestions as a committee member have been extremely helpful to my learning and research training.

This dissertation was supported by the Arizona State University/Mayo Clinic seed grant program. To doctors Joe Hoxworth, Bhavika Patel, Leland Hu, and Yonas Geda at the Mayo Clinic, thank you for offering me opportunities to collaborate with you and to advance my work. Also, thank you, Dr. Jing Li for your guidance in devising the machine learning experiments of the seed grant studies. Dr. Christine Zwart, thank you for your support during the first two years of my PhD. Dr. Kristin Swanson, thank you for your support and guidance in more recent years.

The staff at the department of biomedical informatics are gratefully acknowledged for their help and support. Lauren Madjidi, thank you for the efforts that you put into advising me. Maria Hanlin and Patty Hutton thank you for your continuous help and support.

Special thanks to my dearest family and friends, both near and far, for their love and support. To name a few: Sahar and Kosar Ranjbar, Aazam Ranjbar, Hassan Bagherebadian, Daniel Furnish, Stephanie Furniss, Salvi Asefi, Pooya Khodaparast, Sangdi Lin, Wenzhe Xue, Stefanie Velgos, Héléne Hilaire, Amol Bhalla, Nima Tajbakhsh, Garen Minassian, and Maral Tabrizi. I would also like to thank friends back home whose support and encouragement during graduate school application process allowed me to be here: Hadi and Mehdi Farzaneh, Khashayar Pourkarami and Nima Ashrafi.

DEDICATION

To my wonderful mother, Äzar,
for her loving support and encouragement.

CONTRIBUTIONS

The main contribution of this thesis is a comprehensive methodological pipeline for converting diagnostic images to quantitative and mineable data. Second contribution of this thesis is several publications that demonstrated the utility of this data. Chapter 4 was published in the American Journal of Neuroradiology in 2017 [1]. Chapter 5 is pending acceptance. Chapter 6 has been accepted by the Journal of Computed Tomography in 2017, and chapter 7 is pending submission. My contributions to these studies are: 1) devising technical design of the studies, 2) implementing the analysis described in "Method" sections, both the image analysis and feature extraction processes, 3) implementing the machine learning and part of the statistical analysis processes presented in "Methods" sections, 4) producing the "Results" of the studies and 5) writing and editing manuscripts. I am the first author or co-first author of these manuscripts.

The content of chapters 1 and 2 is selected from a book chapter that I wrote for the Elsevier book titled *Biomedical Texture Analysis: Fundamentals, Applications, Tools, and Challenges Ahead*. The book aims to provide technical guidance to future students and researchers about the field of biomedical texture analysis. My chapter, chapter 8, is titled "An Introduction to Radiomics: An Evolving Cornerstone of Precision Medicine" [2] and is intended to introduce clinicians and future researchers to the utility of this field. All chapters are peer-reviewed by authors of other chapters as well as the editors to ensure validity of content and precision of language.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

MOTIVATION & PROBLEM STATEMENT

Cancer is the second leading cause of death in the United States after heart disease and is the leading cause of death in 22 states, and in Hispanic and Asian Americans [3]. Each year, the American Cancer Society estimates the number of new cancer cases and deaths for the current year. For 2017, 1,688,780 new cases and 600,920 cancer deaths are projected to occur in the United States. These estimates are the equivalents of more than 4,600 new cancer diagnoses and about 1,650 cancer deaths per day [3].

Over the past 3 decades, the 5-year relative survival rate for all cancers combined has increased >20% as a result of advances in early detections and improved treatments [3]. However, the progress in cancer care including the decline in number of incidences and increase in survival rate, has had almost no effect on cost of care. Cost of cancer care increases annually by a 2% rise of cost in the initial and last phases of care, for a total of 174$ billion for projected cost in 2020. This estimate represents a 39% increase from 2010 [4].

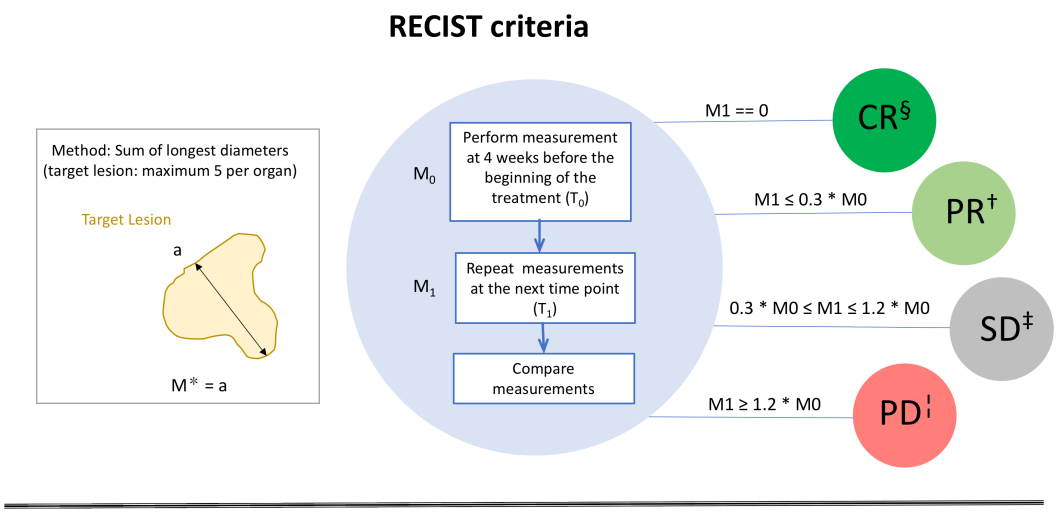
The rate of progress in improving patient survival for some cancers has been remarkable. Hematopoietic and lymphoid malignancies for instance have experienced an improvement of 5-year relative survival rate of 71% for acute lymphocytic leukemia and 66% for chronic myeloid leukemia [3]. For solid tumors, the steady increase in survival rate for most cancers has been much slower (i.e. 18% and 8% for lung and pancreatic cancer). The slower rate of progress for solid tumors has been linked to late stage diagnosis and poor tumor characterization processes that do not characterize the nature of these tumors effectively or do not reflect the aggressive behavior of these cancers [3].

Cancer care of solid tumors involves the detection, diagnosis, characterization, treatment, and monitoring of the disease. Detection begins when, upon an initial indication, a radiology

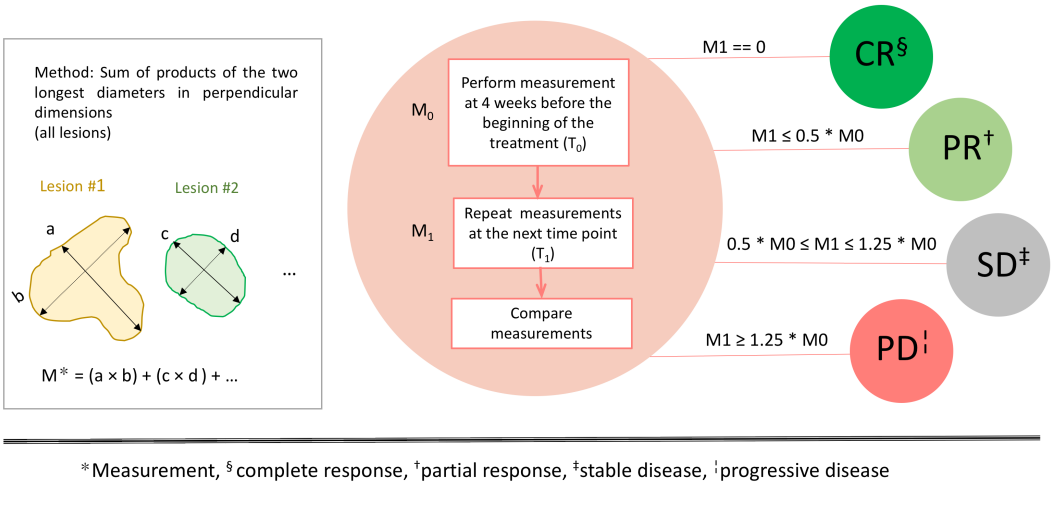
exam is requested for a patient. A radiologist then prompts an exam by determining the content of the order based on the clinical context for the initial request including the appropriate imaging modality (Computed Tomography or CT, Magnetic Resonance Imaging or MRI, etc.), body part (breast, lung, brain, etc.), and image acquisition protocol (with/without injection of contrast agent) [5]. Upon receiving the exam request, a radiology session is scheduled and the images are acquired. A radiologist then reviews the images and prepares a report containing the image findings. Image findings typically contain detailed information regarding the key attributes of the lesion(s) including anatomic location, qualitative and quantitative description of the lesion. The reported qualitative attributes of the lesion(s) include the radiologist's subjective visual description of the lesion such as lesion shape, margin, and border or density.

Quantitative attributes of lesions are measures of tumor burden estimation and are rarely reported in standard of care practice at initial detection [5] and are only presented during the treatment planning phase as part of the treatment response assessment process. The content of a radiology report at these stages includes the above-mentioned attributes, as well as an estimation of comparison of lesion attributes (qualitative and quantitative) over time. The criteria for measuring the lesion's quantitative attributes are defined in Response Evaluation criteria In Solid Tumors (RECIST), and/or Wold Health Organization (WHO). As shown in Figure 1.1, estimation of tumor burden in clinical practice relies on human measurement of one- or two-dimensional descriptors of tumor size [6, 7].

Following the initial diagnosis, the next step is characterization of the detected lesion. The gold standard for characterizing lesions is the histopathologic analysis of tissue samples acquired through surgical biopsy or resection [8]. Initial diagnosis (benign or malignant, as well as grading) and genetic status of solid tumors are determined from tissue samples obtained via a biopsy. Treatment of solid tumors may involve combination of tumor resection through surgery, radiation therapy, and chemotherapy. Radiotherapy is a major part of the



(a)



(b)

Figure 1.1: Measurement of tumor burden via RECIST and WHO response assessment criteria. Measurements rely on (a) sum of longest diameters of target lesion in RECIST, and (b) sum of products of the two longest axes in WHO criteria. In both criteria, measurements are repeated over time, leading to an estimation of disease status into four categories: complete response, partial response, stable disease, and progressive disease.

treatment process for roughly half of cancer patients and is particularly useful for cancers of the head and neck, lung, bladder and prostate. Radiotherapy is also used as a pre-surgery treatment to shrink the size of tumor in an attempt to control the spread of the disease (i.e. in breast cancer) [9]. Radiotherapy is the only treatment option for locally advanced cancers where the location of the lesion renders surgery an unsuitable option (i.e. cancer of the larynx) [9]. The prescribed radiation dose depends on the type of cancer, the tolerance of the surrounding normal tissue, and the amount of radiotherapy toxicity up to five years past treatment [9].

Treatment of cancer via chemotherapy heavily depends on the information provided via the acquired tissue samples. For decades, chemotherapy treatments of malignant lesions have involved targeting rapidly dividing cancer cells using intravenous cytotoxic chemotherapy. Due to the blind nature of this type of treatment many rapidly dividing normal cells (e.g., hair, gastrointestinal epithelium, bone marrow) are also affected leading to classic toxicities for the patients (i.e. alopecia, gastrointestinal symptoms, and myelosuppression) [10]. Cancer treatment has experienced a dramatic change as a result of the new advances towards targeted therapies. In the age of targeted therapies particular monoclonal antibodies and small molecule inhibitors are targeted to disrupt distinct mechanisms of tumor cells. Elements such as the lesion cellular environment, including presence/absence of certain enzymes and protein expression, cell type, presence/absence of gene alterations, level of oxygenation within the cell, and growth factors all affect the course of treatment [11].

1.1 Motivation

Invasive biopsy or surgery procedures are the sole means for the provision of bio-specimen samples that provide insight into tumor composition and molecular context. These invasive procedures cannot be performed with an aggressive frequency; in fact, oncologists normally rely on the information provided via two or three biopsies for their decisions during the whole course of treatment.

An important factor to consider is that intra-tumor heterogeneity can vary among cancer types and subtypes [12]. Because biopsy samples are extracted from a single (or two) location(s) within the lesion, they can be narrow in scope and therefore fail to reflect the variability of molecular structure of tumors [13]. In fact, they can only serve as a snapshot of the tumor's molecular structure at the time and location of biopsy sampling. This sets a limitation on the utility of biopsy samples given that the heterogeneity of solid tumors has been established at the phenotypic, physiologic, and genomics levels [14–16]. Another drawback of biopsies is the narrow temporal scope of the information that they can provide. Given that cancer is a phenotypically and molecularly progressive disease, the molecular context of the disease changes over time. Hence, a single biopsy at a single time point may not be able to convey the changes in the molecular context of the lesion [17].

Biopsy procedures are also susceptible to sampling error. Closed biopsies (with a needle, endoscopy, the preferred biopsy procedure in complicated cases) have been shown to have a lower rate of accuracy compared to open biopsies (with an incision), especially for soft-tissue tumors [18]. The histopathology reports may contain errors as well. Nguyen et al. [19] found an error rate of 44% in prostate pathology reports. The authors found that in 10% of the pathology reports the errors were significant enough to change the course of treatment. Another study [20] found that 7.8% of the breast cancer pathology reports included errors significant enough to affect the choice between a lumpectomy and mastectomy for the patients.

The high prevalence of cancer has prompted significant promotion in biomarker (defined in Table 1.1) research, even though the utility of biomarkers in clinical domain doesn't start or end with cancer. This promotion is due to the fact that further advancement of targeted therapies goes hand in hand with advancements of biomarker research. Biomarkers are introduced at various stages of cancer to help identify cancer type and stage, forecast the effect of treatment, monitor progress of the disease, and predict outcome (e.g. time to relapse, survival) [21]. Thus, biomarkers offer many benefits for cancer care including faster,

cheaper, more efficient diagnosis and treatment processes [8, 22], as well as smaller sample sizes for clinical trials [21].

Table 1.1: Biomarker terminology (courtesy of [23])

Term	Definition
Biological Marker (Biomarker)	A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacological responses to a therapeutic intervention.
Clinical endpoint	A characteristic or variable that reflects how a patient feels or functions, or how long a patient survives.
Surrogate endpoint	A biomarker intended to substitute for a clinical endpoint. A clinical investigator uses epidemiologic, therapeutic, pathophysiologic or other scientific evidence to select a surrogate endpoint that is expected to predict clinical benefit, harm or lack of benefit.
Useful Biomarker	(1) Informs risk/benefit ratio when decisions are to be made. (2) Does so in a better/safer/faster/earlier/cheaper way than existing approaches. (3) Generally applicable: sample and technology must be available/accessible. (4) Has a known identity.

An ideal cancer biomarker should follow the changes of the molecular context in which they find themselves. Most new potential treatments fail to obtain regulatory approval partly due to clinical trial biomarkers that are poor predictors of treatment effectiveness. Within the context of solid tumors, bio-specimen samples such as biopsies serve as the main source of biomarkers used in the treatment and monitoring of cancer. However, bio-specimen biomarkers inherit the limitations of their source: biopsy samples. They offer limited *longitudinal granularity*: A quality that is critical for assessing the impact of therapeutic interventions on the molecular context of the disease [24]. To address this shortcoming imaging biomarkers have been investigated. An imaging biomarker refers to "an imaged characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or a response to a therapeutic intervention" [25]. Lesion size remains the only recognized imaging biomarker extracted from routine care imaging

to evaluate the response to treatment. This is despite that lesion size doesn't reflect the complexity of lesion context and changes in its behavior [26]. Changes in tumor dynamics are known to start much earlier than the any change in lesion size [26], meaning current measurement criteria rely on a rather distal indicator of treatment effectiveness. Another limitation of using lesion size as the primary imaging biomarker is that lesion size does not predict patient prognosis such as overall or progression-free survival [5, 27]. Inter-reader and intra-reader size measurement variability [28] along with inconsistency between readers in selection of target and non-target lesions are among other limitations of this process [5]. These issues are so widely known that the pharmaceutical companies often employ "independent review panels" to standardize tumor response measurements in clinical trials [7]. Based on a report by Thiesse and colleagues [29] on a large multicenter trial in oncology, the independent review panels and the case radiologists were found to have major disagreement in 40% of the cases and minor disagreements in a further 10.5% of cases. Qualitative descriptions are also reported in radiology reports of cancers. To name a few, characteristics such as lesion margin, shape, border, and intensity patterns (enhancement pattern, vascularity in breast lesions) are among parameters that are reported qualitatively and therefore suffer from subjectivity and measurement variability across readers.

1.2 Objective & Aims

The rise of high-throughput tools capable of massive processing of genomics data has led to exciting advances in the field of computational biology. The ability to translocate gene expression profiles of various tissues has provided us with an unprecedented snapshot of disease signatures on a molecular and cellular level [30]. These advances are slowly influencing how we view other data sources in health care. Imaging has traditionally been viewed as a largely qualitative diagnostic tool; a means to provide important anatomical and morphological information about the clinical question at hand. The vast amount of available high quality imaging data is an excellent candidate for the analytical high-throughput

approach. Recently, an emerging practice known as "radiomics" has been introduced to investigate this opportunity. Radiomics refers to extraction and analysis of large numbers of quantitative features from diagnostic images to improve disease characterization and prediction of outcome [13, 26, 30, 31]. The findings of recent radiomics studies suggest the potential for imaging data in quantification of tumor phenotypic characteristics to improve biomarker research.

Radiomics research falls directly in the field of biomedical imaging informatics by pursuing the effective use of imaging data, information, and knowledge for scientific inquiry, problem solving and decision making. It also investigates how to reason on these data in order to effect beneficial change in the healthcare enterprise. This thesis, therefore, focuses on contributing to the body of research in the field of radiomics. The ultimate goal of this thesis is to improve integration of imaging data in clinical decision making processes to improve patient outcome. The intermediate goal is providing evidence of the utility of imaging data in the context of other sources of knowledge such as clinical data to promote imaging biomarker discovery. The research, therefore, unfolds over the following two aims.

AIM I Making imaging-derived information accessible:

The usefulness of images in developing decision support systems in dependent on the availability of quantitative imaging data. To this end, an analytical informatics pipeline is implemented for characterization of multiscale imaging data. This tool is a comprehensive methodological pipeline for conversion of diagnostic images to quantitative and mineable data. When pointed to an anatomical region with clinical relevance, it extracts large numbers of quantitative features to objectively describe the spatial and spectral appearance of the region on a pixel-by-pixel level. Images from several modalities (Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Contrast Enhanced Spectral Mammography (CESM)) can be processed with the pipeline.

AIM II Bridging the gap of imaging-derived information and other biomedical data:

The integration of quantitative features from biomedical images with other clinical and genomic data is an ongoing area of research. Image-drived data can be used for integration and construction of models that yield new insight about diseases, or can be used for development of decision support tools to improve clinical practice. For this aim, I focused on linking the imaging features with the context of diseases in which imaging biomarkers have an established presence: cancer and Alzheimer's disease. Four retrospective validation studies were conducted to assess the utility of imaging data in diagnostic applications, and if possible and relevant, to assess if using this data can lead to improved diagnostic accuracy compared with that of clinical practice.

1.3 Overview of Chapters

Chapter 2 describes the background of the radiomics approach, focusing on the motivations of radiomics, the literature in this field, and summary of a generic radiomics workflow. Chapter 3 presents technical details of the theoretical and methodological pipeline proposed in AIM I. This chapter provides a description of preprocessing steps including image normalization and segmentation of regions of interest. Analysis methods used for extraction of quantitative features are described next followed by the summary of the analysis processes used for assessing clinical utility of imaging-derived features.

Chapter 4 to 7 presents the validation studies conducted for fulfillment of AIM II. In respective order, these chapters review studies related to sinonasal cancer (chapter 4), breast cancer (chapter 5), human papillomavirus (chapter 6) and finally, the Alzheimer's disease (chapter 7). These chapters begin by a summary section in which the goal of the study, details of the publication, and my contributions to the study are highlighted. Each chapter follows the format required for publication: abstract, introduction, material and method, results, discussion and conclusion.

Some of the challenge in imaging biomarker discovery is resolving the technical and standardization barriers to providing actionable evidence of imaging data. Chapter 8 highlights the challenges related to feature extraction, selection, and analysis processes and their respective impact on the result. This chapter also discusses the future of radiomics research and its impact in advancement of image-based diagnosis and biomedical imaging informatics. The limitations of the pipeline presented in this thesis are discussed next. Chapter 9 presents the conclusion of this research.

Chapter 2

BACKGROUND LITERATURE

2.1 Introduction

In this chapter, I describe a recent paradigm for quantitative image analysis known as radiomics. My aim for this chapter is to clarify what radiomics is and where it stands in the spectrum of cancer research. First, the motivations and potential impacts of radiomics is discussed particularly in relation to advancement of solid tumor cancer research. This section is followed by a review of the findings of the latest literature. Next, high-level workflow of radiomics is described including the input imaging data, feature extraction methods, and challenges of data analysis.

2.2 Radiomics

Quantitative image analysis has a well-established footprint in cancer research. Imaging traits have been used for characterization of healthy and pathologic tissues, for differentiation of tumor regions, as well as for tumor grading. Some examples in this regard are the imaging studies in brain [32–35], breast [36–41], lung [42, 43], and prostate [44] cancers. Imaging qualities have also been used to predict prognosis markers such as time to progression and survival rate of brain tumors [45], non-small cell lung cancer [46], colorectal cancer [47–49], and renal cell cancer [50].

There are similarities between the radiomics approach and conventional quantitative image analysis. They both explore imaging data for potential utility in treatment, monitoring, outcome prediction, or biomarker discovery. However, radiomics differs from quantitative image analysis in the high throughput approach to quantification of tumor phenotypes. By facilitating extraction of large quantities of information from diagnostic imaging radiomics

not only leads to comprehensive assessment of the clinical utility of the imaging features, but also serves as a bench mark to standardize feature extraction and analysis processes.

Radiomics data has shown potential in reflecting the complexity of tumor composition and behavior. By promptly detecting small and early focal responses this data can lead to enhanced tumor characterization and response assessment. Advocates of radiomics are motivated by the fact that imaging is nearly non-invasive (completely when no contrast agent is injected) and can shed light on the lesion state in its entirety even in complicated cases where biopsy procedures are risky. Moreover, the extracted imaging information may prove to be complementary or interchangeable with other sources, i.e. demographics, pathology, blood biomarkers, genomics [13]. Hence, radiomics can lead to earlier termination of ineffective treatment plans leading to lower cost of treatment and enhanced quality of life during treatment for the patient.

Growth of radiomics research can potentially lead to the progress of imaging biomarker research. Alizadeh and colleagues [51] identify blood-borne biomarkers as the main non-invasive physical source for gaining insight into the molecular composition of tumors. Cell-free DNA and circulating tumor cells can be accessed through blood samples to provide insight into the molecular composition of tumors. However, blood samples have a fundamental limitation, as they discuss: "It is unclear whether primary tumors or metastases contribute more to the pool of circulating cancer material. It seems clear, however, that even if circulating material is found to faithfully reflect the tumor itself, there is still a need for more efficient ways of isolating the cells and nucleic acids from the blood and for data analysis tools that can more faithfully reconstruct the parent tumor". Upon validation, radiomics imaging traits can breed non-invasive biomarkers that can be identified as surrogates for important clinical outcomes [52]. These biomarkers can be utilized in various stages of cancer care such as detection, diagnosis, assessment of prognosis, prediction of response to treatment, and monitoring [31].

2.2.1 Radiogenomics

One of the emerging areas in this area of study is known as *Radiogenomics*. This term has been used in the literature to describe two quite different concepts: (1) the practice of analyzing whole genome to enhance our understanding of molecular pathogenesis and risk factors of variation in radiotherapy toxicity [9, 53, 54], (2) establishing association maps between radiomics data and genomics (or other -omics) data. To clarify, here we refer to the latter definition which has also been described as 'Imagenomics' in the literature [55].

Radiogenomics focuses on mapping the quantitative imaging input to the findings of computational biology. A major motivation behind radiogenomics studies is that imaging is already part of routine diagnostics, an advantage that gene expression profiling is only beginning to develop [13, 30]. In contrast to genomics profiling, virtually all patients go through imaging multiple times during their treatment [31] and, given the non-invasive nature of the procedure, imaging can theoretically be repeated as often as required. Therefore, by establishing association maps between imaging traits and previously determined treatment response or gene-expression findings, radiogenomics can lead to non-invasive imaging signatures for the genomics pathways and a shift towards individualized medicine in routine care.

Given that genomics profiling is not available for all the patients who go through cancer treatment, radiogenomics can potentially offer two services for advancing targeted therapies towards a more genetically informed paradigm. Firstly, radiogenomics offers the possibility of mapping the distribution of genomics markers across the entirety of lesion. Genomics markers are detected from local and temporal tissue samples. By building association maps between these markers and imaging traits, radiogenomics can introduce *radiophenotypes*, specific image phenotypes that associate with presence/absence of genes [52]. These characteristics can serve as surrogates for gene expression signatures [30]. If such association maps are established and validated, radiogenomics can detect where in other parts of the lesion those

imaging traits are present and therefore prompt biopsy procedures to target more specific areas of tumors [56]. Secondly, radiogenomics can serve as an independent source of additional information to inform on how a biological process is reflected in images [52]. Such information could be merged with genomics data leading to enhanced diagnostic and prognostic power [31].

2.3 Examples of Radiomics Literature

Aerts and colleagues [27] showed that it is possible to identify imaging traits that are consistent across multiple datasets and associate them with tumor prognosis. They conducted a study in which 440 features were extracted to quantify tumor intensity, shape and texture on the CT images of patients with lung or head-and-neck cancer. Features were found to have significant association with tumor histology, and to have prognostic power in seven independent datasets. The authors also found associations between radiomics features and the gene-expression profiles of lung cancer patients in one dataset. Along these lines, Parmar and colleagues [57] investigated cancer-specific radiomic feature clusters in four independent Lung and head-and-neck cancer cohorts. Several feature clusters, extracted from the pre-treatment CT images, exhibited stability across the datasets. The authors found strong associations of these features with lung prognosis and staging of head and neck lesions.

Coroller and colleagues [58] extracted 635 radiomic features from pre-treatment CT-scans of lung adenocarcinoma patients. 35 features were predictive of distant metastasis (DM), and 12 features showed correlation with survival. The authors concluded that, given the strong power of their radiomics signature in predicting DM, the result could be used as a prognostic biomarker for factors such as distant metastasis.

Hu and colleagues [17] highlighted how radiomics can serve as a noninvasive means to potentially predict histopathology. They extracted 900+ texture features from various MRI contrasts of Glioblastoma Multiforme (GBM) lesions at the biopsy locations. The authors

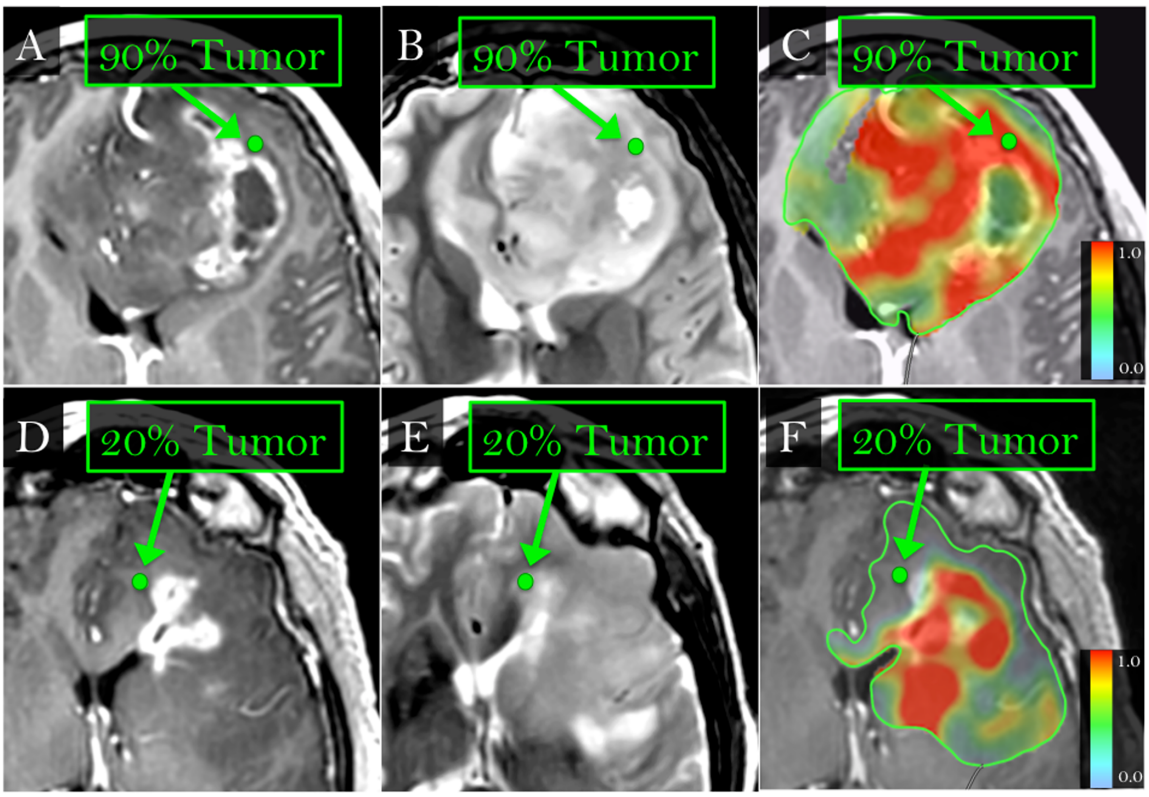


Figure 2.1: Prediction of tumor content in Glioblastoma Multiforme (GBM) using radiomics data (courtesy of Hu et al. [17]). Authors generated a model to bridge the gap betweem radiomics features extracted from MRI images and tumor content of biopsy samples. This figure shows the success of their model on two previously unseen cases in the validation set. (A, B, C, E) Biopsy locations within the non-enhancing (green dots, arrows) on T1+C (A, D) and T2W (B, E) images corresponding with high-tumor (>80% tumor nuclei) and low-tumor (<80% tumor nuclei) tissue samples on histologic analysis. Colormaps (C, F) with manual tracings (green) show the probability (range 0-1) of tumor-rich (red) vs tumor-poor (green/blue) content across the tumor for the two cases.

generated a classification model that predicted the percentage of tumor content from the texture features. Figure 2.1 shows the colormap overlay of their predictions across the entire lesions in two validation cases. Hu and colleagues [56] also predicted local gene expression data through multivariate analysis of copy number variant (CNV) status, multi-parametric MRI, and texture analysis. This study proposed that a radiomics model with an acceptable level of success in the prediction of the presence of a certain gene can provide invaluable information to guide further biopsy procedures. Figure 2.2 shows the proposed tree model of imaging features for prediction of PDGFRA amplification (Part A), along with colormap overlay of prediction (Part C) on 2 stereotactic biopsies (Part B) for a validation case.

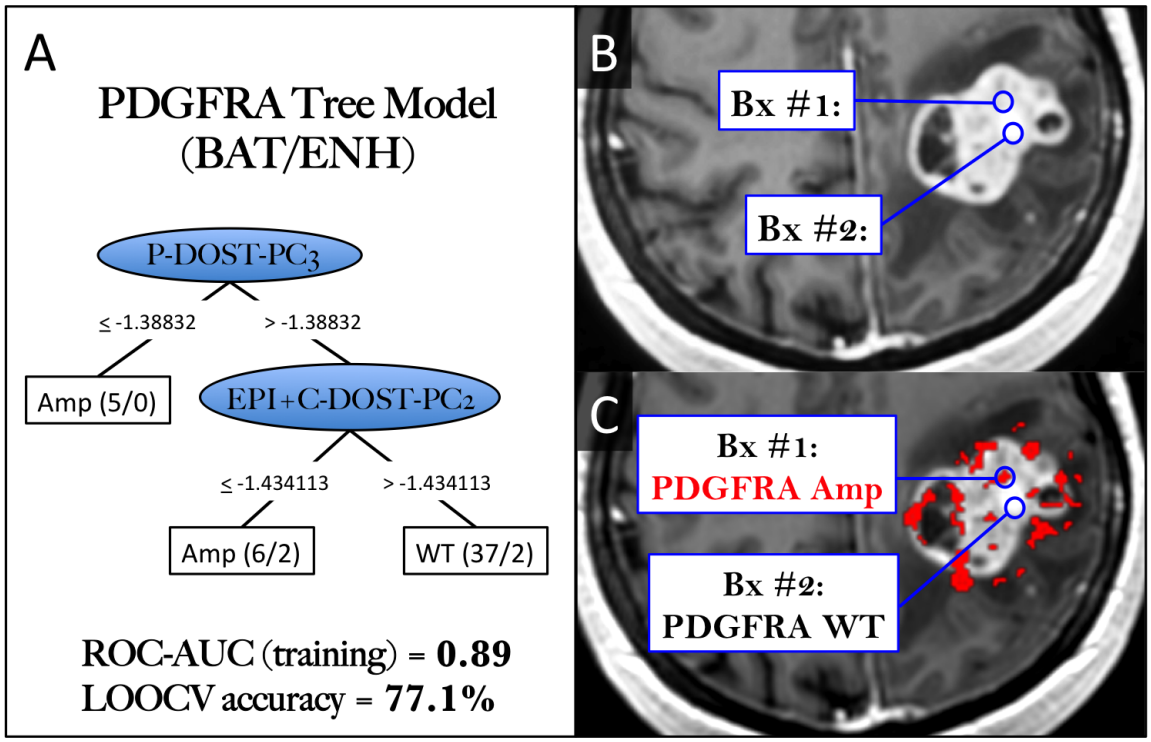


Figure 2.2: Radiogenomics model for GBM (Figure courtesy of [56]). (A) (P= isotropic diffusion; EPI+C=T2*W signal loss; DOST=discrete orthonormal Stockwell transform; PC3, PC2 derived from principal component analysis (PCA), AUC, area under curve). (B) the locations of 2 stereotactic biopsies (Bx#1, Bx#2) on CE-MRI. DNA CNVs demonstrated amplification (Amp, Bx#1) and diploid/wild-type status (WT, Bx#2). (C) regions (ROIs) of predicted PDGFRA amplification (red voxels) using tree model classification.

Zinn and others [59] identified and corroborated genomic correlates of edema/cellular invasion in GBM with MRI phenotypes. Gevaert and others [60] extracted quantitative image features from the enhancing necrotic part of GBM tumor and peritumoral edema on MR images. Their radiogenomics map linked 56% of the 50+ imaging features with biologic processes, while 3 and 7 imaging features were significantly correlated with survival and molecular subgroups, respectively. Focusing on GBMs, Diehn et al. [61] showed initial proof that a radiogenomics approach could potentially be used in selecting patients for targeted therapies. They identified an imaging biomarker of EGFR expression from MR images; an image trait termed "contrast necrosis" highly associated with EGFR gene expression. They effectively utilized this image trait to select GBM patients with over-expressed EGFR protein.

In a clear cell renal cell carcinoma application, Karlo and others [62] found associations between known genetic mutations and imaging traits describing the tumor margin, size, vascularity, and enhancement on pretreatment CT images. Segal and colleagues [63] associated distinctive imaging traits in CT scans with global gene expression profiles of liver cancer related to angiogenesis (such as vascular endothelial growth factor, VEGF). Their results showed association maps between 28 features and 78% of the global gene expression profiles. In a preliminary breast cancer radiogenomics study, Yamamoto and colleagues [64] provided an association map linking MR image phenotypes to underlying global gene expression patterns. They found 21 (out of 26) imaging traits of breast MRI as globally correlated with 71% of the total genes measured in patients with breast cancer.

2.4 Workflow of Radiomics

Implementation of the radiomics approach (in the traditional learning scheme) involves the following discrete processes: (a) image acquisition, (b) segmentation, (c) feature extraction, (d) feature analysis and model generation. In this section, we briefly describe each of these steps. Figure 2.3 shows the simplified workflow of the training and application of the radiomics approach. Step 1 shows the training phase of the workflow in a conventional learning scheme

which is bound by preparation of feature sets. In this learning scheme, a high number of features are extracted from diagnostic images via a radiomics engine (part a). These features are then analyzed in the prediction engine (part b) using machine learning techniques. The output of the prediction engine divides the high dimensional feature space of radiomics to a much simpler classification space (part c) to assess the clinical hypothesis of the dataset. Step 2 shows how this pipeline can be applied for prediction of the result in a previously unseen case.

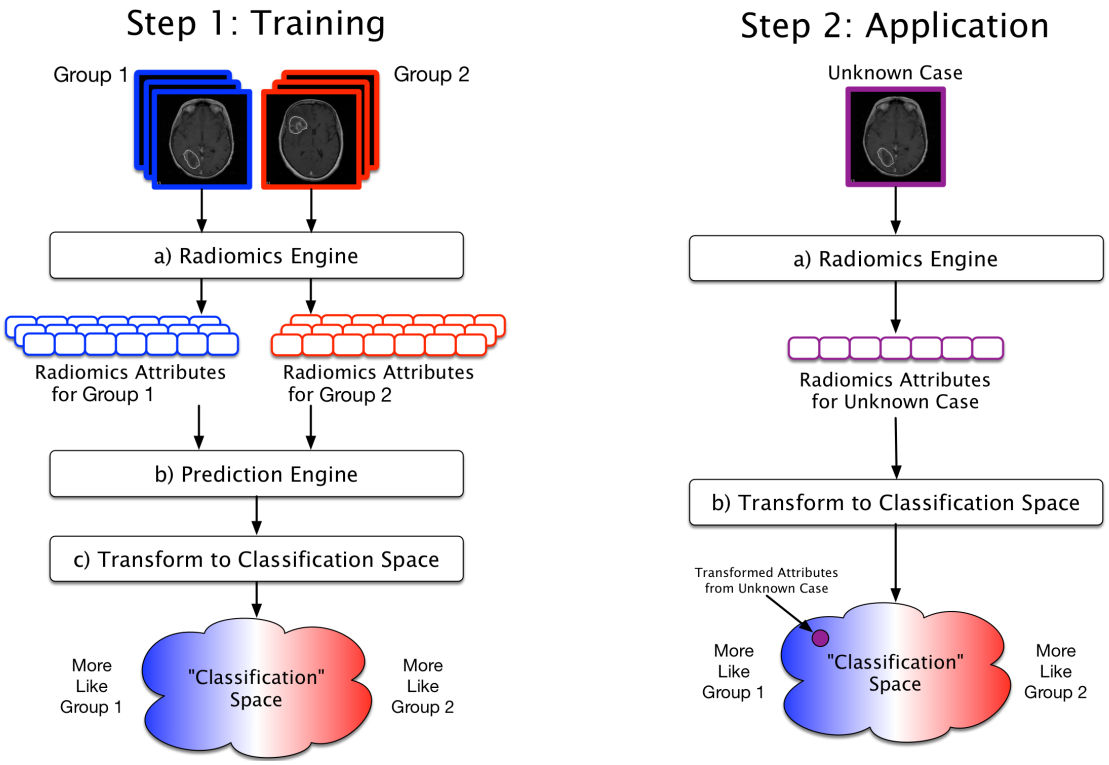


Figure 2.3: A simplified radiomics workflow. In training phase (Step 1) a high number of features are extracted from diagnostic images via a radiomics engine and are used to generate a low dimensional classification space from the training data. The classification model is used on unknown cases in Step 2.

2.5 Inputs of the Radiomics Engine

Just as cancer therapy research has advanced, so too has cancer imaging, resulting in images with higher temporal resolution (number of frames per unit time), spatial resolution (number of image pixels per spatial unit), and contrast resolution (number of bits per pixel). Depending upon the cancerous organ, the location of metastasis, and patient tolerance to the procedure, various imaging modalities are ordered. Some well-known image modalities are Computed Tomography (CT), Magnetic Resonance Image (MRI), or Positron Emission Tomography (PET).

CT scans can provide high resolution structural information about the location and shape of lesions [65]. When combined with contrast injection, CT scans exhibit strong contrast reflecting the presence of underlying disease or injury in the organs, blood vessels and/or tissue types [27]. CT perfusion shows which areas are perfused adequately with blood and provides detailed functional information on delivery of blood or blood flow to specific organs.

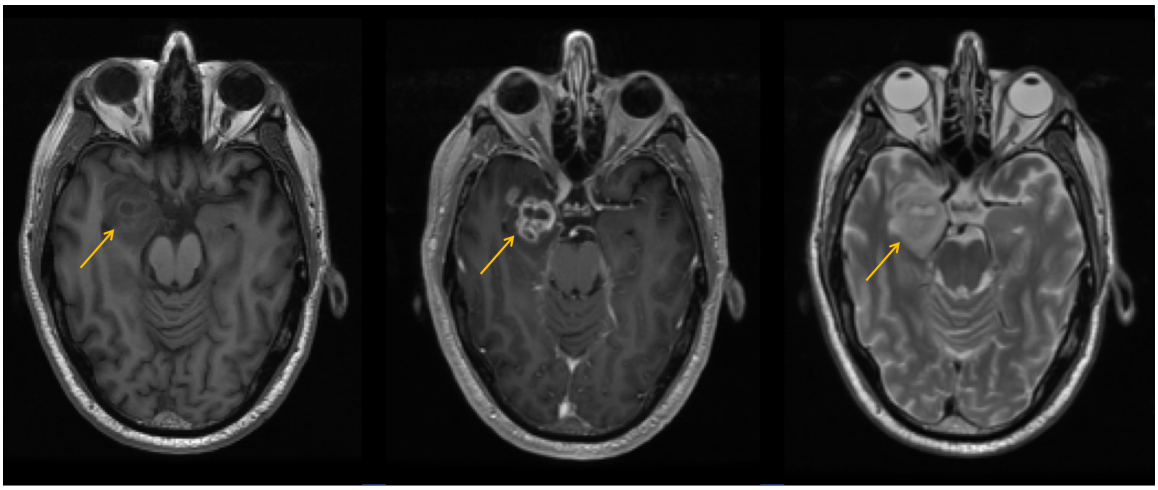


Figure 2.4: MRI images of brain with tumor viewed in axial plane. The arrows show the location of glioblastoma multiforme (GBM) on (a) standard pre-contrast T1-weighted, (b) post-contrast T1-weighted, and (c) pre-contrast T2-weighted sequences.

Tumor functional and molecular characteristics are assessed using Magnetic Resonance

Imaging (MRI). The use of different spin-echo sequences for capturing MRI results in exhibition of different functionalities of lesions. MRI contrast may be weighted to demonstrate different anatomical structures or pathologies (Figure 2.4). For instance, T1-weighed MRI is used for obtaining morphological information while MR diffusion weighted, Flair sequences and T2-weighted are useful as measures of cellularity, edema and inflammation. MR diffusion has been reported to detect early changes that correlate with tumor response [66, 67]. Lesion perfusion is captured using dynamic contrast enhanced-magnetic resonance imaging or DCE-MRI, providing information on lesion blood flow, permeability, and angiogenesis [68]. Diffusion-weighted MRI (DW-MRI) holds promise for use as a cancer treatment response biomarker as it is sensitive to macromolecular and micro structural changes which can occur at the cellular level earlier than anatomical changes during therapy [69].

Tumor metabolism is assessed using a type of nuclear medicine imaging known as Positron Emission Tomography (PET) imaging. PET imaging provides important information regarding blood flow, oxygen use, and sugar metabolism of lesions. However, PET has limited spatial resolution. More recent developments in imaging have resulted in devices that perform imaging via dual sources such as PET/CT and PET/MR [70, 71]. However, the standard-of-care imaging modalities such as CT and MRI are the most explored imaging modalities for incorporation in radiomics workflows due to the potential impact in routine care [72–74].

Regardless of imaging modality, the next step prior to the extraction of radiomics features is segmentation of the region(s) of interest. Lesion segmentation is performed to precisely define the spatial extent of the lesion on the images. The validity of the segmentation result is of utmost importance because of its direct impact on the radiomics features and the utility of their content. Prior to segmentation, several image preprocessing steps are requisite, including spatial registration of different image sequences and modalities, noise reduction, and intensity calibration to account for the differences in the anatomical and intensity variations of the images.

Manual outlining by an expert is considered the *ground truth* in tumor segmentation. However, the result of manual segmentation can be unreliable due to inter and intra reader segmentation variability [75]. Moreover, outlining a 3D volumetric object manually is a tedious and time consuming task rendering it impractical. In comparison, semi-automatic methods involve a minimal amount of input from an expert user (e.g., a seed point to initialize the segmentation or manual editing of the results). Region growing methods, level set methods, graph cut methods, active contour algorithms and semi-automatic segmentations such as livewires are among the most widely used segmentation approaches [26]. Automatic and semi-automatic segmentation methods are the preferred method of segmentation in radiomics worflows due to their robustness and significantly higher levels of reproducibility [76].

2.6 Feature Extraction

Radiomics features cover a wide range of quantitative attributes including shape, margin, boundary, intensity, texture, etc. Lesion shape, boundary, margin, and location are established prognostic measures, and are already a part of the radiologists' lexicons for description of lesions [31]. In regard to this group of features, the goal of the radiomics approach is to introduce quantitative measurements to reduce variations of image interpretation due to human error and inter-observer variability [39, 77].

Biomedical texture information relates to the micro- and macro- structural properties of biomedical tissue. Texture descriptors are generally divided into three major categories: statistical, structural, and spectral. Statistical texture features are non-deterministic descriptors of the distribution of gray-level intensities in a region of interest. First order statistical features (i.e. mean, max, min, variance, kurtosis) are histogram-based descriptors of intensity distribution and can only provide global information regarding the target region intensity pattern. Second order statistical features describe inter-relationships of gray-level values within a region and are the most widely used feature extraction method in medical pattern

recognition tasks. gray level co-occurrence matrices (GLCM) [78] and gray level run length matrices (GLRLM)[79] are two instances of this group. In GLCM, a variety of features are used to describe a matrix generated from the observed frequency of gray-level combinations within neighboring pixels, while in GLRLM the focus is on describing coarseness of texture in predefined directions. The utility and efficacy of GLRLM in comparison with GLCM or spectral methods were debated in the early stages [80, 81], however it has been increasingly investigated for clinical applications in the last decade [82–84]. Grey-scale and rotationally invariant Local Binary Patterns (LBP) [85] is another statistical descriptor of texture focusing on the patterns of intensity transition within the sub regions of an area of interest. The radius of the sub regions determines the scale of the texture described by the LBP features.

Structural descriptors view texture in terms of texture primitives such as micro and macro textures. Model-based features describe the region of interest through mathematical models such as fractals or stochastic models [86]. There are fewer reported radiomics applications for structural and model-based methods (an example: [87]), as they are viewed as more suitable for synthesis than analysis.

Spectral descriptors rely on the frequency or scale domain representation of the region of interest for calculation of texture features. By adjusting the window/filter expansion in space or frequency, these transform-based methods allow for multi-resolution, multi-scale representation of texture. The wavelet transform (WT) [88], the Riesz transform [89], the Stockwell transform (ST) [90], the Discrete Orthonormal Stockwell Transform (DOST) [91] and the Polar Stockwell Transform (PST) [92], Gabor filter banks (GFB) [93], and Laplacian of Gaussian Histograms (LoGHist) [94] are only some examples of this group.

2.7 Predictive Engine

The predictive engine is in charge of extracting clinically relevant information from the large amount of data generated using the radiomics approach. Dimensionality reduction and

feature selection are two of the fundamental building blocks of this engine and are responsible for selecting and/or generating valuable features from the large number of often-redundant radiomics features. Selected features can be used to build predictive models of clinical endpoints such as patient survival and tumor progress or models that predict diagnostic information such as tumor stage and tumor diagnosis. Predictive modeling generally involves a type of supervised learning scheme in which the labels of samples (i.e. tumor stage, diagnosis, or pathology) are presented to the classifier for a subset of the available data known as training set. The model is later tested by predicting the label of previously unseen data (2.3).

A certain classification approach might be better suited for a type of data based on the characteristics of the data and the manner in which the classification approach is utilized [95]. Designing the best classification approach that can lead to stable and reproducible models requires a substantial amount of knowledge in machine-learning and statistical inference. The interested reader is referred to [96–98] for more information.

Chapter 3

METHODOLOGICAL PIPELINE: A RADIOMICS APPROACH TO IMAGE ANALYSIS

3.1 Summary

This chapter describes the methodological pipeline that has been proposed in fulfillment of specific AIM I. Beginning with the overall description of the platform, this chapter provides the steps involved in preprocessing of biomedical images and extraction of region of interest. Next, feature descriptors are explained in detail including intensity-based, statistical, and spectral texture features. To finalize, machine learning techniques and statistical analysis used along with this pipeline are presented.

3.2 Proposed Platform

The pipeline presented here follows the genreric radiomics workflow described in chapter 2. This pipeline provides a path for quantitative analysis of image texture and intensity along with a selected number of morphological features (shape descriptors). Although a universally agreed-upon definition of texture does not exist among the experts, it is often described as the patterns perceived by humans that reflect the spatial organization or repeated sub-patterns of the image. Several feature extraction algorithms are included in this pipeline that examine brightness distributions or spectral representation within the region, searching for spatial relations between pixels that have particular gray levels. The pipeline computes numerical values that correspond to intrinsic properties of texture and intensity.

Texture can not be detected in a single pixel (or voxel). These patterns can only be characterized within sufficiently large regions that can contain several repeated patterns or a texture fragment which is representative for a whole texture. Similarly, the whole image cannot be the input for texture feature computation because the features won't correctly

reflect characteristics of particular objects or meaningful areas. A region of interest therefore is required to limit the area from which the features are computed. These regions are provided to the pipeline as input files.

There are two modes for feature computation in this package. The first mode computes features from a user defined region of interest to provide a numeric representation of the region as a whole. In this mode, the pipeline provides a close-to-optimal description of the region by computing features that describe incremental degrees of texture scales. The second mode computes features from a sliding small window placed around every image pixel within a large region of interest, capturing the variability of small scale texture features across a large region.

The inputs of the pipeline are 2D images or slices and a file that provides the coordinates of the region of interest boundary (with *'.txt'* or *'.roi'* extensions). Features extraction algorithms included in the pipeline are listed in Table 3.1. The output of the pipeline is a spreadsheet in which columns are image modality, patient Identification number, slice number, size of the region of interest, and the name of the numeric features. The rows represent the values of columns for each patient. The script for this pipeline is written in the platform-independent and open source Python[99] programming language.

3.2.1 Preprocessing

Figure 3.1 shows the steps involved in preprocessing of images. Inputs to the pipeline are the address of a root folder where images and the region of interest (ROI) files physically exist, as well as the format of images (DICOM, or generic image formats such as PNG/JPEG/TIFF) and the format of ROI files. To ensure similarity of images, several image properties are assessed and matched across the cohort including image dimensions (image size), image pixel spacings (physical size of a pixel in x and y directions), and dynamic range of images (range of intensity levels). Image resampling is performed when inconsistencies are observed in pixel

spacings and image dimensions of images. These preprocessing steps are necessary to ensure that the difference among subsequent texture features are not the result of imaging protocol or image acquisition settings.

Table 3.1: Texture Descriptors Included in the Platform

Algo.	Description	Category	default # features for a region size: 8x8	size:16x16	size:32x32
-	Raw Intensity Descriptors	Statistical	3	3	3
GLCM	Gray Level Co-occurence Matrices	Statistical	26	26	26
LBP	Local Binary Patterns	Statistical	12	12	12
GFB	Gabor Filter Banks	Spectral	8	8	8
DOST	Discrete Orthonormal Stockwell Transform	Spectral	6	10	15
LoGH	Laplacian of Gaussian Histograms	Spectral	18	18	18

Segmentation of the ROIs is the next preprocessing step, in which the pipeline loads the ROI files and select the regions of interest from the images. The ROI files provide image coordinates of the boundary of a selected region with a clinical relevance. The ROIs are the result of manual outlining by an expert and can include either coordinates of a selected rectangular area or coordinates of a closed free-form boundary. Given that a larger area can result in a larger number of features, at this step the pipeline compares the sizes of the regions of interest and calculates the size of the largest square-shaped region that can be extracted from all cases, while remaining within the boundaries of the ROI file. This leads to extraction of sub-images which are input of feature extraction engine. One last step before feature extraction is intensity normalization of sub-images to a common dynamic range to ensure that the subsequent features characterize the sub-image texture exclusively

and do not depend on any global imaging characteristics such as the overall brightness or illumination of images.

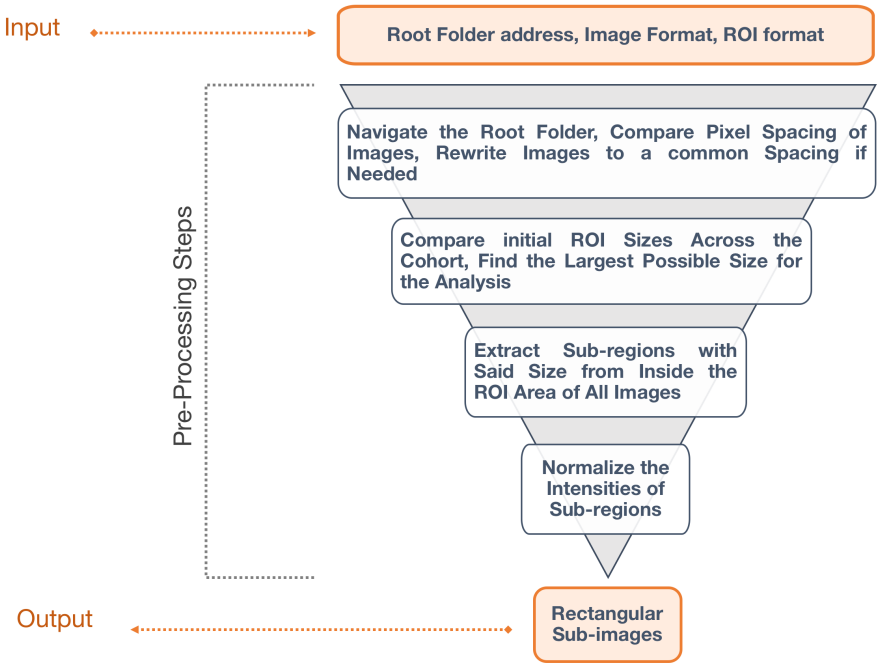


Figure 3.1: Image Preprocessing Steps. Provided with a the address of a root folder, images and ROI files are processed to select a common size of region of interest for analysis.

3.2.2 Feature Extraction

The Feature extraction module is the heart of the pipeline that generates two categories of features: texture descriptors, and shape descriptors. Texture descriptors can be further divided into three categories: Intensity-based, statistics-based, and transform-based measures.

Intensity-based features are first-order statistics of the histogram of pixel intensities (Figure 3.2). Features computed in this method are simply statistical descriptors of the histogram distribution (pdf) such as: mean brightness, variance, skewness, kurtosis and percentiles. Because these features disregard the spatial relations between the pixels within

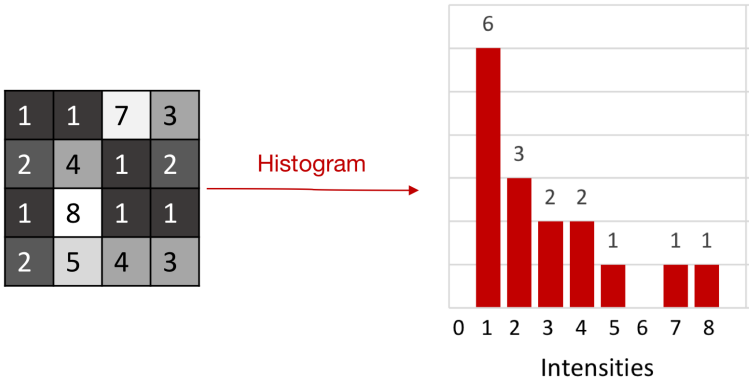


Figure 3.2: Intensity-based features are first-order statistics of the histogram of pixel intensities. Examples are mean brightness, variance, skewness, kurtosis and percentiles.

the image some disagreement exists in the literature about whether or not they should be considered texture features. Nonetheless, they are descriptors of the visual appearance of the image as a whole and are generally reported along with established texture features. This pipeline includes mean, range, and standard deviation of the gray-level intensities from this category. These features are computed before the final preprocessing step (dynamic range normalization) to report the visual appearance of the original image.

The Next group is statistical texture descriptors which attempt to provide insight into local intensity distributions within the region. The module includes two statistical texture descriptors: Gray Level Cooccurrence Matrices (GLCM) and Local Binary Patterns (LBP). The gray-level co-occurrence matrix (GLCM) [78] is a second-order histogram, computed from the intensities of pairs of pixels. The popularity of GLCM lies in its simplicity. Each element in the co-occurrence matrix shows how often a pair of intensity levels is seen in a neighborhood defined with a certain angle and distance (Figure 3.3). A GLCM matrix is computed per one direction/angle at a time. In order to produce unidirectional and rotationally-invariant features, several matrices are computed and merged by either regrouping the counts of operators over all directions in a shared matrix, or by averaging scalar texture measurements

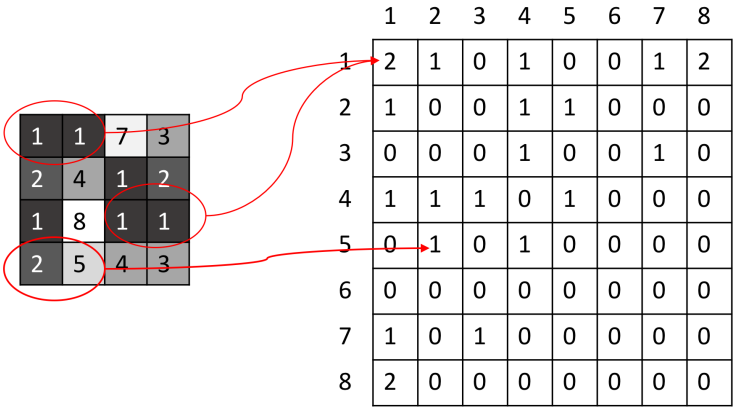


Figure 3.3: GLCM matrices are calculated per a distance and an angle. The matrix in right is calculated for the image in the left for distance of 1 and the horizontal direction. The element (1, 1) in the matrix contains the value of 2 because there are two instance in the input image where two horizontally adjacent pixels have the values 1 and 1. The element (5, 2) in the matrix contains the value of 1 because there is only one instance where two horizontally adjacent pixels have the values 5 and 2. This process is continued, scanning the image for other pixel pairs (i, j) for remaining directions (and potentially radii).

from the matrices obtained with different directions. In this pipeline 4 directions are accounted for 0°, 45°, 90°, and 135° to produce isotropic texture features. The radius of 1 is used as default to describe spatial relationships between immediately neighboring pixels. The radius can be adjusted to capture these relationships across larger neighboring regions. A set of 13 features are calculated using this method, examples of which are presented in Table 3.2. This pipeline computes the mean and range of the set of features for a total of 26 features.

Rotationally-invariant local binary patterns (LBP) [85] focus on intensity distributions of neighboring areas inside the region. Given a radius, this method places a patch with that radius around every pixel in the region. The choice of the neighborhood radius defines the scale of features described by this method. In each round, the intensities of the pixels around a patch are converted to binary values based on whether intensities are greater or lesser than that of the center point (Figure 3.4). The binary levels for each patch are aggregated

Table 3.2: Texture Features from GLCM. The reader is referred to [78] for the complete list of features.

Feature	Formula	Description
Angular Second Moment	$\sum_{i,j=0}^{N-1} \left(P_{ij}\right)^2$	Also known as uniformity and homogeneity, ASM is high when image is homogeneous.
Contrast	$\sqrt{\sum_{i,j=0}^{N-1} \left(P_{ij}\right)^2}$	Square root of ASM.
Entropy	$\sum_{i,j=0}^{N-1} -ln\left(P_{ij}\right) P_{ij}$	measure of randomness of matrices. Inhomogeneous scenes have low first order entropy.
Variance	$\sum_{i,j=0}^{N-1} P_{ij} \left(i-\mu\right)^2$	Variance or sum of squares. This feature puts relatively high weights on the elements that differ from the average intensity value.
Correlation	$\sum_{i,j=0}^{N-1} \frac{\left(i-\mu_i\right)\left(j-\mu_j\right) P_{ij}}{\sigma_i \sigma j}$	Measure of gray-tone linear dependencies.
Inverse Difference Moment	$\sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+\left\|i-j\right\|^2}$	IDM is the local homogeneity. It is high when local gray level is uniform.

P_{ij} = Element i,j of the normalized symmetrical GLCM

$\mu_i = \sum_{i=0}^{N-1} iP_{ij}$, $\mu_j = \sum_{j=0}^{N-1} jP_{ij}$, $\sigma_i^2 = \sum_{i,j=0}^{N-1} P_{ij}\left(i-\mu\right)^2$, $\sigma_j^2 = \sum_{i,j=0}^{N-1} P_{ij}\left(j-\mu\right)^2$

and converted to a binary code which corresponds to a decimal value. Aggregated decimal values for the whole region are organized into a histogram from which the LBP features are extracted. These features can be either normalized counts of histogram bins or descriptors of histogram shape such as mean, variance, skewness and kurtosis. Default setting calculates local binary patterns for the radius of 3. Normalized bin counts of a 12-bin histogram are reported as the set of LBP features.

Transform-based texture descriptors are computed using 3 methods: Discrete Orthonormal Stockwell Transform (DOST), Gabor Filter Banks (GFB), and Laplacian of Gaussian

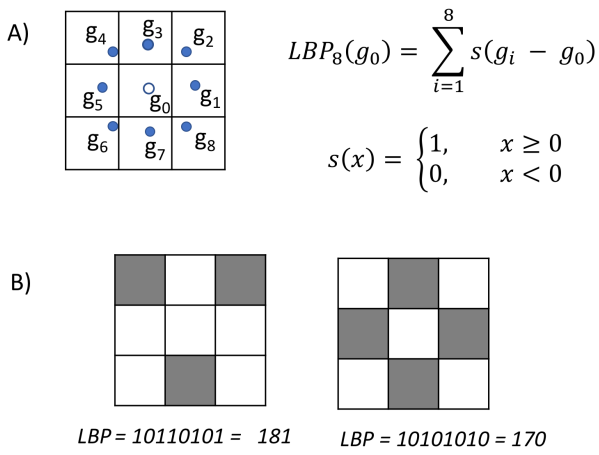


Figure 3.4: The circularly symmetric neighbor set of eight pixels in a 3x3 neighborhood. A) The gray-level intensity of the center pixel (g_0) is compared with the gray values of the eight surrounding pixels of the circularly symmetric neighborhood. The result is a binary code that defines the transition of pixel intensities across the neighborhood. B) Two examples of a 3x3 region along with the LBP binary codes and the final digits.

Histograms (LoGHist). The methods use convolution as the type of linear operation used to apply the texture operator to the image. A convolution in the spatial domain corresponds to a multiplication in the Fourier domain.

Discrete Orthonormal Stockwell Transform (DOST) [91] provides a multi-resolution spatial-frequency representation of an image. The process of calculating DOST features for a 2D image begins with defining the forward 2D Fourier Transform (FT) of a discrete function $f(x, y)$ which is assumed to have sampling interval of one in the x- and y- directions:

$$H[m, n] = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} h[x, y] \; e^{-2\pi i \left(\frac{mx}{M} + \frac{ny}{N}\right)}$$

And the inverse 2D-FT:

$$h[x, y] = \frac{1}{MN} \sum_{m=-M/2}^{M/2-1} \sum_{n=-N/2}^{N/2-1} H[m, n] \; e^{2\pi i \left(\frac{mx}{M} + \frac{ny}{N}\right)}$$

Next a dyadic sampling scheme (orders 0, 1, 2, ..., Log N-1) is used to partition the 2D FT space into non-overlapping sections. The 2D-DOST of a NxN image h[x, y] is therefore

calculated by partitioning the 2D-FT of the image, H[m,n], multiplying by the square root of the number of points in the partition and performing an inverse 2D-FT,

$$S\left[x\prime, y\prime, \nu_x, \nu_y\right] = \frac{1}{\sqrt{2^{p_x+p_y-2}}} \sum_{m=-2^{p_x-2}}^{2^{p_x-2}-1} \sum_{n=-2^{p_y-2}}^{2^{p_y-2}-1} H\left[m+\nu_x, n+\nu_y\right]\ e^{2\pi i\left(\frac{mx\prime}{2^{p_x-1}}+\frac{ny\prime}{2^{p_y-1}}\right)}$$

where $v_x = 2^{p_x-1} + 2^{p_x-2}$ and $v_y = 2^{p_y-1} + 2^{p_y-2}$ are the horizontal and vertical "voice frequencies". The spectrum is partitioned such that the wave numbers (ν_{x0}, ν_{y0}) are shifted to the zero wave number point, a 2^{p_x-1} x 2^{p_y-1} inverse FFT is performed, resulting in a rectangular (in general) voice image of 2^{p_x-1} x 2^{p_y-1} points. In this package rotationally invariant features are calculated from the DOST image by averaging the magnitude of the horizontal and vertical frequency values for each order.

Gabor filter banks [93] is another spectral method included in this pipeline. This methods captures frequency content of the ROI through filtering the ROI with 2D Gabor filters. Gabor filters are described by their orientation (θ), central frequency (f) and bandwidths in horizontal and vertical directions $(\sigma_x,\ \sigma_y)$. The equation of Gabor filter in the frequency domain is:

$$G(u,\nu,f,\theta) = \frac{1}{2\pi\sigma_u\sigma_\nu}\left[\exp(-\frac{1}{2}\left(\frac{(u_\theta-u_0)^2}{\sigma_u^2}+\frac{(\nu_\theta-\nu_0)^2}{\sigma_\nu^2}\right)+\exp(-\frac{1}{2}\left(\frac{(u_\theta+u_0)^2}{\sigma_u^2}+\frac{(\nu_\theta+\nu_0)^2}{\sigma_\nu^2}\right)\right]$$

$$\begin{cases} u_\theta = ucos\theta + \nu sin\theta \quad \nu_\theta = -usin\theta + \nu cos\theta \\ u_0 = \frac{2\pi cos\theta}{f} \qquad\qquad \nu_0 = \frac{2\pi sin\theta}{f} \\ \sigma_{u,\nu} = \frac{1}{2\pi\sigma_{x,y}} \end{cases}$$

Gabor filters are handcrafted and reflect upon only a section of the frequency representation of the ROI. Therefore, their success in capturing the true frequency content of the ROI depends on the choice of filter shape, frequency, and bandwidths. In our experience with these filters, $\sigma_x = \sigma_y = (1.0, 3.0)$ and $f = (0.6, 1.0)$ showed successful for small ROIs of sizes 8x8 - 16x16. For larger ROIs, the pipeline computes features for a wide range of Gabor filter

shapes and sizes (as reported in chapter 4). Mean and standard deviation of each filtered image are reported as Gabor features.

Laplacian of Gaussian [94] or LoG is the last convolutional texture processing approach included in the pipeline. LoG uses a radial second-order derivative of a 2-dimensional Gaussian filter as:

$$g_{\sigma_i}(x) = -\frac{1}{\pi\sigma_i^2}\left(1 - \frac{\|x\|^2}{2\sigma_i^2}\right) e^{-\frac{\|x\|^2}{2\sigma_i^2}}$$

where σ_i defines the scale of the Gaussian filter. All structures at scales much smaller than a particular σ are blurred leaving only textural information of a particular scale. Isotropic multi-scale texture measures can be obtained by reporting the absolute values or the energies of the band-pass filtered maps. Default setting for computation of LoGhist features in this pipeline is set for Gaussian filters of $\sigma = (2.0, 4.0, 6.0)$ to cover a range of fine to medium scale textures. The pipeline then calculates statistical features (i.e. kurtosis, skewness, standard deviation, mean, entropy, and uniformity) for the normalized histogram of the LoGs. The said features are reported for each level of σ, counting up to a total of 18 LogHist features.

Four shape features are also included in the package. These features are extracted from cases where the region outline has clinical relevance, for instance breast cancer in which irregular shape of the lesion is an indication of malignancy. There are currently 4 shape irregularity descriptors, described in [100], included in the pipeline. These features quantify the similarity of the lesion boundary to a circle.

3.2.3 Feature Analysis

The number of features computed by this pipeline is large, reaching 73 features for an 8x8 region. Consequently an analysis engine is required to provide meaningful information related to the clinical utility of these features. However, the choice of the analysis approach depends highly on the data at hand and the question of the study. Therefore the analysis engine of

this pipeline is not hard-coded. Nonetheless, a number of frequently-used techniques for statistical analysis, dimensionality reduction, feature selection, and classification have been scripted in the pipeline. These scripts borrow from python-based open source libraries such as Scipy[101] and Scikit-Learn[102].

Hypothesis testing is implemented using 2-sided tests with statistical significance level defined as $P < 0.05$. 2-sample t test are performed for univariate comparison of features between different groups before the application of machine-learning methodology. P-values are corrected for multiple comparisons using the false discovery rate [103].

Given that high number of features can cause classifier overfitting, dimensionality reduction is an important pre-classification step. In this step the original feature space is transformed into a new space of lower dimensionality to increases the chance of capturing underlying relationships presumably present in the data and increase predictive performance of the models on unseen data. This pipeline includes dimensionality reduction technique using Principle Component Analysis (PCA) [104]. PCs are linear combinations of features and are uncorrelated and ordered accordingly to their standard deviations. PCA is performed separately for each texture algorithm and imaging source. Imaging source in this context refers to any factor that affect the content of the image, examples are image modality (CT, MRI, or Mammography), Imaging sequence (T1, T2, or post-contrast T1 for MRI), and image view (i.e. two common views of image acquisition in mammography imaging). The PCs can contain a certain level of the variability within a feature set (i.e. 80% - 95%). These PCs are used for further processing. PCs are computed on features of all patients regardless of diagnosis.

A critical step in modeling and analysis is feature selection. Feature selection is performed to avoid overfitting to noise and to remove redundant features. Given that in the radiomics approach the focus is on being *thorough*, it often computes a large number of features, many of which are either highly correlated with each other or are redundant. There are two

general strategies to feature selection. Filter strategy estimates the value of each attribute individually, resulting in low computational complexity. These methods can be applied prior to classification, possibly in conjunction with a random sampling approach, to serve as another dimensionality reduction technique. However, filter-based feature selection strategy may not be able to reveal the actual discriminative power as it views each feature independent of the others. An alternative is the wrapper approach in which utilizes a search method to search through the whole feature space and identify a relevant and non-redundant feature subset [105].

This pipeline includes Sequential Forward Feature Selection (SFFS)[104] method. SFFS starts with an empty set of features estimating discriminative power of individual features. Features are gradually added to the feature with highest power so that all combinations of feature pairs are examined. This method offers minimal redundancy but maximal relevance of features. SFFS can also or be wrapped in model generation and classification. The criterion that we used for adding features in SFFS is accuracy of the feature set, meaning features were added to the set as long as they contributed to the accuracy of leave one out cross validation beyond 1% or 5%. Threshold of 5% was used when the available data sample size didn't allow for separate train and test sets to reduce the possibility of overfitting.

Several classification (individual and ensemble algorithms) approaches have also been scripted in the pipeline and are available for use with minimal modification. Examples are Support Vector Machines (SVM), random forests, and Diagonal Quadratic Discriminant Analysis (DQDA). These methods were particularly selected since they are among supervised classification methods that have elicited good performance for high dimensional data with small sample sizes. Leave one out cross validation and splitting samples are implemented to assess the model performance on trainset. ROC curves statistics such as sensitivity, specificity, accuracy, area under the ROC curve, confidence intervals of the area under the ROC curve, and model significane are reported as the metrics of model evaluation. The

classification models is also stored for future validation studies where generalizability of the models can be assessed by testing them against previously unseen data.

Chapter 4

VALIDATION STUDY I: MRI-BASED TEXTURE ANALYSIS TO DISTINGUISH SINONASAL SQUAMOUS CELL CARCINOMA FROM INVERTED PAPILLOMA

4.1 Summary

This chapter reports on a retrospective pilot study to bridge between imaging-derived features and pathology results. In this study we evaluated the utility of imaging-derived features extracted from MRI images in detecting signature appearance of sinonasal tumors on MRI images. This work was published in American Journal of Neuroradiology in 2017 [1] and was presented at the Annual Meeting of the North American Skull Base Society, February 12-14, 2016, Scottsdale, Arizona. My contributions to this manuscript are: 1) devising technical design of the study, 2) implementing the the image analysis process, 3) implementing the machine learning process, 4) writing and editing the manuscript.

4.2 Abstract

Background and Purpose: Because sinonasal inverted papilloma can harbor squamous cell carcinoma, differentiating these tumors is relevant. The objectives of this study were to determine whether MR imaging-based texture analysis can accurately classify cases of noncoexistent squamous cell carcinoma and inverted papilloma and to compare this classification performance with neuroradiologists' review.

Materials and Methods: Adult patients who had inverted papilloma or squamous cell carcinoma resected were eligible (coexistent inverted papilloma and squamous cell carcinoma were excluded). Inclusion required tumor size of >1.5 cm and a preoperative MR imaging with axial T1, axial T2, and axial T1 postcontrast sequences. Five well-established texture analysis algorithms were applied to an ROI from the largest tumor cross-section. For a

training dataset, machine-learning algorithms were used to identify the most accurate model, and performance was also evaluated in a validation dataset. On the basis of 3 separate blinded reviews of the ROI, isolated tumor, and entire images, 2 neuroradiologists predicted tumor type in consensus.

Results: The inverted papilloma ($n = 24$) and squamous cell carcinoma ($n = 22$) cohorts were matched for age and sex, while squamouscell carcinoma tumor volume was larger ($P = .001$). The best classification model achieved similar accuracies for training (17S squamous cell carcinomas, 16 inverted papillomas) and validation (7 squamous cell carcinomas, 6 inverted papillomas) datasets of 90.9% and 84.6%, respectively ($P = .537$). The machine-learning accuracy for the entire cohort (89.1%) was better than that of the neuroradiologists' ROI review (56.5%, $P = .0004$) but not significantly different from the neuroradiologists' review of the tumors (73.9%, $P = .060$) or entire images (87.0%, $P = .748$).

Conclusion: MR imaging-based texture analysis has the potential to differentiate squamous cell carcinoma from inverted papilloma and may, in the future, provide incremental information to the neuroradiologist.

4.3 Introduction

Inverted papilloma (IP) is an uncommon sinonasal tumor of ectodermal origin that most commonly arises from the lateral nasal wall [106, 107]. In addition to its pattern of locally aggressive behavior and propensity for post-operative recurrence, there is a tangible association with malignancy, mostly squamous cell carcinoma (SCC). Although reports vary widely in frequency, the rate of carcinoma is on the order of 10-15% and approximately 60-70% of these are synchronous [108, 109]. Although office-based endoscopic incisional biopsy is safe, the sensitivity for the diagnosis of malignancy has been called into question due to sampling error [110].

It can be useful to preoperatively identify SCC when coexistent with IP to guide biopsy,

expedite surgery, and plan an oncologically sound resection. Although bone thinning and remodeling without large areas of erosion on CT are more characteristic of IP than SCC, this finding is imperfect because IP may also aggressively destroy bone, particularly when contacting the walls of the sphenoid sinuses and floor of the anterior cranial fossa [111, 112]. Not surprising, FDG-PET has shown a higher mean standard uptake value for SCC compared with IP, but significant overlap limits the clinical utility of PET [113]. To date, MR imaging has the most promise in differentiating SCC from IP. Although early work initially found no signature appearance on MR imaging for IP, alternating hypointense and hyperintense bands on T2-weighted and contrast-enhanced T1-weighted sequences have become recognized as a distinctive feature of IP and have been described as a convoluted cerebriform pattern [111, 114–116]. As with any subjective finding, interpretive errors can occur, especially when tumors are small or incompletely express the convoluted cerebriform pattern. Hence, a more objective form of image analysis is desirable.

Texture analysis is a form of image processing that seeks to characterize complex visual patterns by quantitatively identifying simpler but characteristic subpatterns. Within the field of head and neck radiology, texture analysis has shown applicability in predicting the p53 status of SCC, classifying SCC as human papilloma virus-related, predicting treatment response in head and neck cancer, differentiating benign from malignant thyroid nodules, and characterizing parotid tumors and structural changes after radiation therapy [117–125]. Because a large volume of data is generated when multiple texture analysis algorithms are applied to MR imaging sequences, the statistical comparison of individual texture features is of limited practical value; instead, a multifactorial data-driven analysis is necessary (ie, "radiomics"). Therefore, the objective of this study was to determine whether MR imaging-based texture analysis can differentiate sinonasal SCC from IP by using a multiparametric machine-learning model. Model performance was additionally compared against qualitative neuroradiologists' interpretation to determine its potential for added clinical value.

4.4 Methods

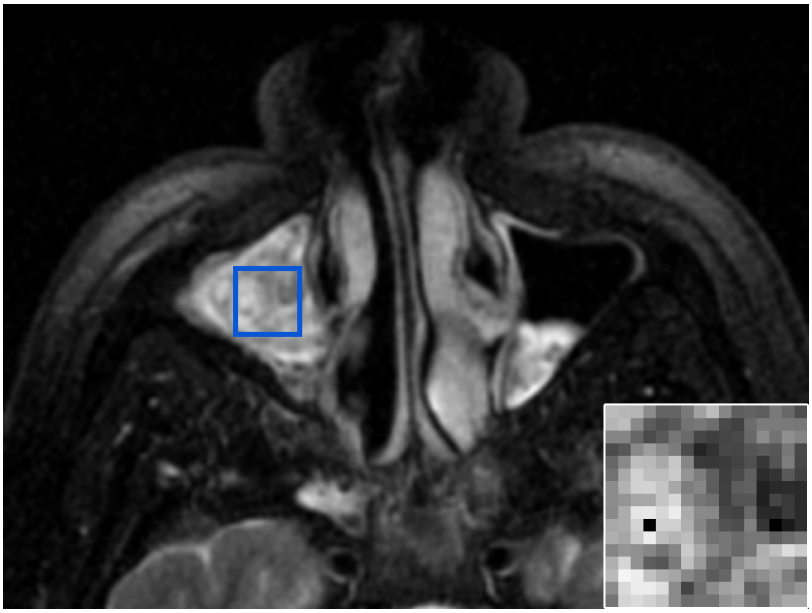


Figure 4.1: ROI placement. A 51-year-old man with an IP involving the right maxillary sinus. Axial T2-weighted fat-suppressed MR imaging pulse sequence demonstrates the manual placement of the largest rectangular ROI that would fit within the tumor margins on the axial image with the greatest tumor cross-sectional area. The inset image in the lower right corner is representative of the final 16x16 matrix that was derived from the ROI isocenter and served as the input for texture analysis.

4.4.1 Subject Enrollment

The institutional review board at the authors' institution approved this retrospective study, and the need for informed con- sent was waived. The pathology data base was queried to identify adult patients (18 years of age or older) who underwent resection of sinonasal IP or SCC. Subjects enrolled from January 1, 2009, to December 31, 2014, were included in the training dataset for model development, while those enrolled between January 1, 2015, and July 1, 2016, composed the validation dataset. To ensure that only a single histologic tumor type would be used for texture analysis, we excluded cases of coexistent IP and SCC. Potential subjects were screened to determine which of them had a preoperative face MR imaging available for review. The MRIs, which were performed on numerous scanners within

the authors' institution and at external facilities, had to be of diagnostic image quality. At a minimum, the imaging had to include an axial T1-weighted MRI pulse sequence (T1), an axial T2-weighted with frequency-selective fat suppression sequence (T2), and an axial T1-weighted postcontrast MRI pulse sequence with frequency-selective fat suppression (T1C) for texture analysis, with a section thickness of $\leq 5mm$, an FOV $\leq 22cm$, and a matrix size of at least 256x192. No restrictions on additional MR imaging technical parameters or type of gadolinium-based intravenous contrast were imposed, and studies were included whether they were performed at 1.5T or 3T field strength. The electronic medical record was reviewed for each potential case, and subjects were excluded if they had an intervention for the sinonasal tumor, including biopsy, surgery, chemotherapy, or radiation therapy before imaging. Subjects were further eliminated if the tumor did not have orthogonal transaxial dimensions greater than 1.5x1.5 cm on at least 1 axial image.

4.4.2 Image Preparation and Texture Analysis

DICOM files containing the T1, T2, and T1C pulse sequences (also referred to as "contrasts" for the purpose of texture analysis) were anonymized and encoded so that all subsequent image analysis was blinded. To ensure uniformity for texture analysis, we performed resampling and/or zero padding to generate images with an 18-cm FOV and a 256x256 pixel array and normalized image intensities to a dynamic range of 0-255. The studies were then reviewed by a board-certified neuroradiologist with OsiriX (Version 6.5; `http://www.osirix-viewer.com`). The borders of the tumor were manually traced on all T1C images on which tumor was visible to generate an ROI-based cross-sectional area for each image and an estimated tumor size by using the ROI volume function in OsiriX. On the axial image with the greatest tumor cross-sectional area, the neuroradiologist inserted the largest possible rectangular ROI that would fit within the tumor for all 3 sequences (Figure 4.1). To prevent the 2D texture analysis from being biased by tumor size, a computer script determined the maximal square ROI that could fit within all manually drawn rectangular

ROIs across all subjects and automatically positioned this smallest common square ROI at the isocenter of each of the rectangular ROIs. The contents of this square ROI, with 16x16 pixels, served as the input for texture analysis.

Texture analysis of each ROI consisted of 3 first-order intensity-based features (mean, standard deviation, and range of gray-level intensities) and features computed by using 5 widely available texture algorithms (all implemented in Python 2.7 programming language [99], by using either custom-written code based on publications or open-source libraries as noted):

1. Gray-Level co-occurrence Matrices (GLCM) is a widely applied method that uses second-order statistics to assess the arrangement of similar gray-scale intensities within an ROI [126]. GLCM evaluates how frequently a pair of intensity levels is identified in an orientation based on a specified angle and radius. In the current study, the co-occurrence matrix was determined for a distance of 1 pixel over 4 angular directions (0°, 45°, 90°, and 135°). The mean and range for 13 rotationally invariant features (including measures of homogeneity, entropy, angular/s moment, correlation, and dissimilarity) were computed at each ROI for each MR imaging contrast [126].
2. Local Binary Patterns [127] evaluates the set of points within a fixed radius of a specified voxel to determine in a binary fashion whether they are higher or lower in intensity than neighboring voxels. Depending on the number of bitwise transitions across this interrogated region, the local binary patterns can be classified as uniform or nonuniform, and histograms of these data provide a measure of ROI uniformity. A 3-pixel radius was selected to complement the smaller scale patterns already assessed by GLCM. A 12-bin histogram was used, resulting in 12 local binary pattern texture features being calculated at each ROI for each MR imaging contrast.
3. Discrete Orthonormal Stockwell Transform (DOST) provides a rotationally invariant multiresolution spatial-frequency representation of an image based on dyadic sampling

of the Fourier representation of the image [91]. Ten DOST features were calculated at each ROI for each MR imaging contrast.

4. Laplacian of Gaussian Histogram (LoGHist) is a convolution-based method to capture the spectral composition of an image in intermediate scales not achievable with first- and second-order statistics. Through the use of varying sizes of bandpass filters, different scales of texture ranging from fine to coarse are highlighted [94].Gaussians with 3 different values of (2.0, 4.0, and 6.0) were used to cover the range of fine-to-medium-scale textures, and 18 LoGHist features were generated at each ROI for each MR imaging contrast.
5. The Gabor filter banks (GFB) technique uses localized and linear filters to capture details in various frequency resolutions [93]. Four different Gabor filters were rendered by using 2 sigma levels (1.0 and 3.0) and 2 frequency levels (0.6 and 1.0). By calculating the mean and standard deviation of the filtered ROI, we computed 8 GFB features at each ROI for each MR imaging contrast.

Table 4.1: Patient demographic characteristics and tumor features

Study Group	Sample Size	Sex (Female/Male)	Age (yr)	Tumor Volume (cm^3)	Tumor Stage[a]			
					T1	T2	T3	T4
IP training	16	4:12	58.0 ± 12.1	21.2 ± 17.7	1	3	10	2
IP validation	6	1:5	58.2 ± 15.3	22.0 ± 6.9	1	1	3	1
IP combined	22	5:17[c]	58.1 ± 13.1[d]	21.4 ± 15.5[e]	2	4	13	3
SCC training	17	4:13	54.0 ± 13.5	55.8 ± 40.5	0	1	4	12
SCC validation	7	1:6	54.6 ± 9.4	43.5 ± 27.9	0	1	2	4
SCC combined	24	5:19[c]	54.2 ± 12.5[d]	52.2 ± 37.7[e]	0	2	6	16

[a] Data are presented separately for the training and validation sets and also as a single combined cohort for each tumor.
[b] Tumor stage represents the Krouse staging system [128] for IP and the American Joint Committee on Cancer staging [129] for SCC.
[c] Fisher's exact test, P = .578.
[d] Two-sample t test, P = .317.
[e] Two-sample t test, P = .001.

4.4.3 Neuroradiologists' Review

Using OsiriX, 2 neuroradiologists with 25 and 28 years of experience, respectively, performed a blinded review to reach a consensus diagnosis of IP or SCC for each case. This was performed during 3 separate rounds of image review, each of which was randomized and completed in the following order:

1. ROI: For the T1, T2, and T1C series, the neuroradiologists exclusively reviewed the 16x16 square ROIs that had been used for texture analysis.
2. Tumor: On all images in the T1, T2, and T1C series, the data outside the tumor margins were zero-filled so that the neuroradiologists could only base their assessment on the intrinsic appearance of the tumor without information regarding tumor location and invasive behavior.
3. Image: The neuroradiologists were able to review the unaltered T1, T2, and T1C imaging datasets in their entirety.

4.4.4 Machine Learning and Statistical Analysis

Open-source R statistical and computing software (`http://www.r- project.org`) was used to perform the analyses and classification. Hypothesis tests were 2-sided, and statistical significance was defined as $P < 05$. The comparison of subject demographics and tumor size between IP and SCC was performed by using a 2-sample t test for subject age and tumor volume and a Fisher exact test for sex. The 2-sample t test was used for a univariate comparison of texture features between IP and SCC before the application of machine-learning methodology, and P values were corrected for multiple comparisons by using the false discovery rate [103].

A total of 231 texture features were calculated for each case (77 texture features per MR imaging contrast). To reduce the dimensionality of the texture features and increase the

generalizability of the predictive model for the training dataset, we used principle component (PC) analysis [130–132]. PCs, which are linear combinations of features, were identified separately for each texture algorithm and MR imaging contrast. Those PCs that sufficiently accounted for 90% of the texture feature variability were selected for further processing. Three commonly described classification algorithms, Diagonal Linear Discriminate Analysis, Support Vector Machines, and Diagonal Quadratic Discriminate Analysis, were conducted on the basis of the selected PCs in an attempt to differentiate SCC from IP [35, 133, 134]. Support vector machine model was build using a radial basis function kernel with coefficient value of 1 and gamma parameter equal to $1/numberoffeatures$. The setting used for quadratic discriminant analysis was no priors on classes and threshold of 0.0001 for rank estimation. Sequential forward feature selection identified the image-based PCs that yielded the greatest accuracy [130, 131]. In developing the classification model, we initially selected the PC with the largest discriminatory power and incorporated additional PCs that improved model accuracy in an iterative fashion until incremental gains in accuracy were 1%.

Classification accuracy was determined by using leave-one-out cross-validation, in which all samples except for 1 were used, while the left-out sample served as the test case with which to assess classification accuracy [45]. This process was repeated until all samples in the training dataset had served as the test case, and the overall cross-validation accuracy was the averaged accuracy. The most accurate classification model was applied in a blinded fashion to the validation dataset, and the diagnostic performance of the model was assessed. Model performance accuracies between the training and validation datasets and between the best classification model and neuroradiologists' review were compared by using a 2-tailed test of population proportion.

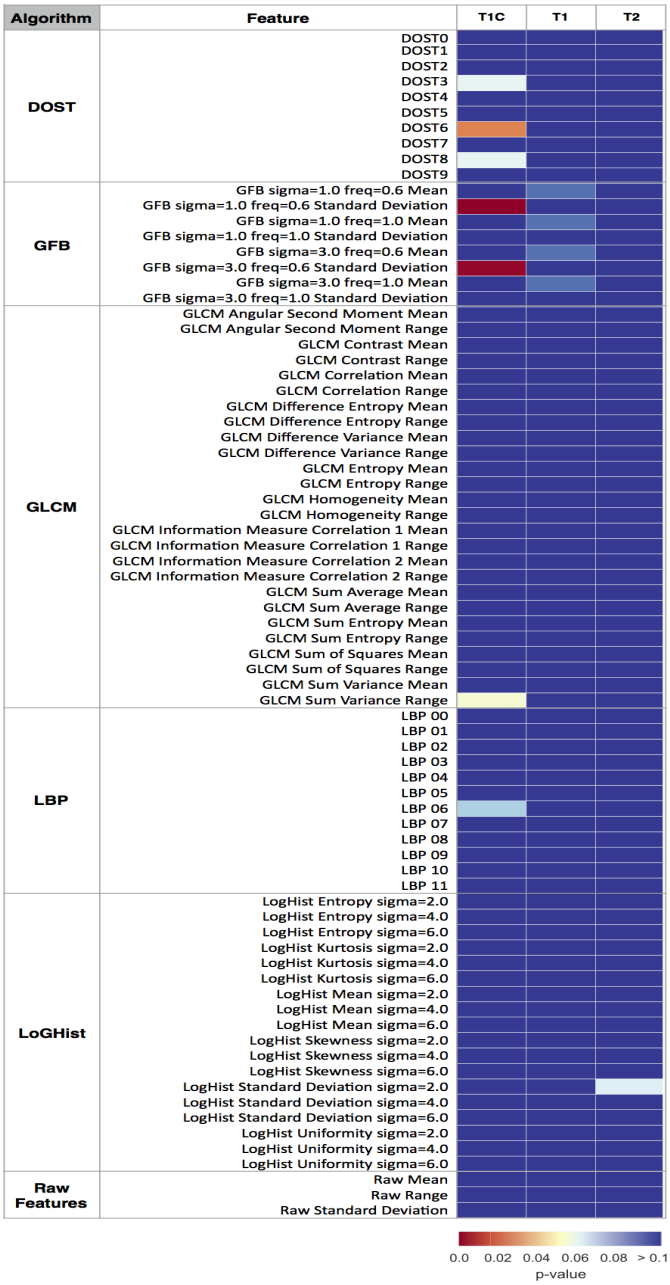


Figure 4.2: Heat map showing MRI texture feature significance in distinguishing tumor type. Univariate analysis compared pathology status (SCC vs. IP) with MRI-texture features. Color maps show the false discovery rate-adjusted P-values of two-sample t test. MRI contrasts (pulse sequences) are listed above the columns, and MRI-based texture features are listed in rows. DOST features 0 to 9 correspond with low to high frequency patterns. LBP 0 to 11 are the normalized bin counts in the LBP histogram. The reader is referred to the Methods section for additional details about the features.

4.5 Results

Thirty-three subjects were included in the training set, 16 IPs and 17 SCCs, while the validation set consisted of 6 IPs and 7 SCCs (Table 4.1). The patients were similarly matched for age and male-predominant sex. Mean tumor volume was larger for SCC than IP ($P < .001$).

Individual features derived from the 5 different texture analyses across all 3 MR imaging contrasts (T1, T2, T1C) were initially evaluated in a univariate fashion to look for significant differences between the IP and SCC groups (Figure 4.2). The greatest number of texture features showing statistically significant differences were derived from the DOST and GFB texture analyses.

Model performance for the training and validation datasets is presented in Table 4.2. Following PC analysis and machine-learning classification, the predictive classifier with the best classification result was using support vector machine, yielding 90.9% accuracy for the training dataset. The 84.6% accuracy of the validation dataset did not significantly differ from that achieved in the training dataset ($P = .537$). When we combined the training and validation cohorts ($n = 46$), the accuracy achieved by texture analysis (89.1%) was significantly better than that of the ROI-based neuroradiologists' review (Table 4.3, 56.5%, $P = .0004$) and showed a trend toward improved accuracy over neuroradiologists' review of the entire tumor (73.9%, $P = .060$). Texture-analysis accuracy was not significantly different from that of the neuroradiologists' reviewing the entire unaltered images (87.0%, $P = .748$).

Relative contributions to model accuracy from each texture analysis algorithm and MR imaging contrast are presented (Figure 4.3). The most significant texture features were derived from T1C-GFB, T1-GLCM, and T1-DOST (Figure 4.4).

Table 4.2: Diagnostic performance of machine-learning classification on training and validation datasets

Model Prediction	Tumor Type (Pathologic Diagnosis)			
for Training Dataset	SCC	IP		
SCC	16	2	Accuracy	90.9%[a]
			Sensitivity	94.1%
IP	1	14	Specificity	87.5%
			PPV	88.9%
Total	24	22	NPV	93.3%
Model Prediction for Validation Dataset				
SCC	6	1	Accuracy	84.6%[a]
			Sensitivity	85.7%
IP	1	5	Specificity	83.3%
			PPV	85.7%
Total	7	6	NPV	83.3%
Model Prediction for Entire Cohort				
SCC	22	3	Accuracy	89.1%
			Sensitivity	91.7%
IP	2	19	Specificity	86.4%
			PPV	88.0%
Total	24	22	NPV	90.5%

Note:–NPV and PPV indicate negative and positive predictive values, respectively.
[a]With a 2-tailed test of population proportion, the accuracies for the training and validation datasets were not significantly different (P = .537).

4.6 Discussion

MR imaging has long been recognized as the most useful technique with which to distinguish sinonasal SCC from IP. Most of the prior work focused on a qualitative imaging appearance known as the "convoluted cerebriform pattern" [111, 114–116, 135, 136]. Although this pattern has a high level of sensitivity for IP, it is not entirely specific. As an example, Jeon et al. [113] evaluated the performance of the convoluted cerebriform pattern in 30 patients with IP relative to 128 patients with sinonasal malignancies and reported a sensitivity of 100%, specificity of 87%, positive predictive value of 64%, negative predictive value of 100%, and accuracy of 89%.

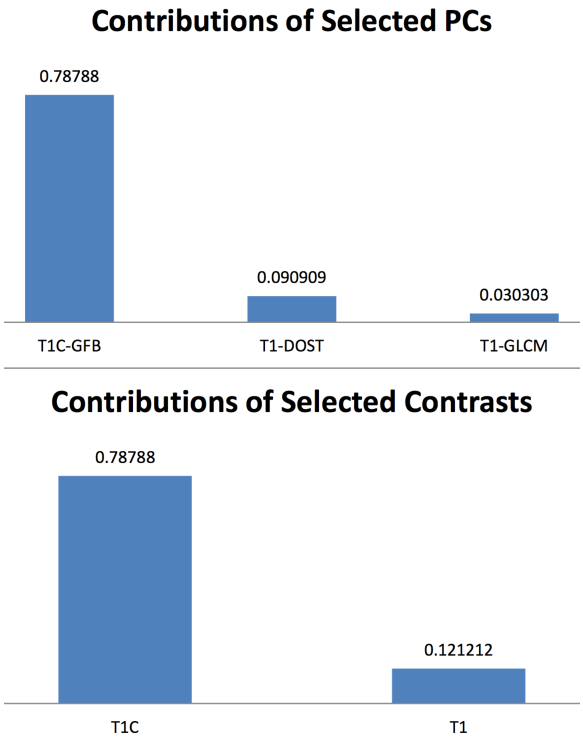


Figure 4.3: Relative contributions to model accuracy. The bar graph demonstrates the accuracy attributable to PCs derived from T1C-GFB, T1-DOST, and T1-GLCM (upper pane) to the 90.91% overall model accuracy. Across all texture algorithms, the contribution to total model accuracy was derived predominantly from T1C, with minor contributions from T1 and no input from T2 (lower pane).

Texture analysis integrated into a machine-learning model was able to classify SCC and IP with an accuracy on par with the previously published results based on the convoluted cerebriform pattern. It is also similar to the best consensus neuroradiologists' interpretation in the current study. However, this technology is meant to supplement a neuroradiologist's interpretive skills rather than compete with them. In clinical practice, a diagnosis is rendered by synthesizing all available data that include not only intrinsic tumor appearance but also other imaging features such as site of origin, tumor size, ex-trasinonasal extension, and tumor margins. Indeed, the current results support a neuroradiologist's accuracy improving for differentiating sinonasal IP and SCC as more imaging information is made available. On the basis of a 16x16 ROI, the texture-based machine-learning model outperformed the accuracy of the neuroradiologists (P = .0004). In terms of assessing the intrinsic tumor

Table 4.3: Diagnostic performance of texture analysis with machine learning compared with neuroradiologists' review for the differentiation of SCC from IP [a]

Analysis Method	Accuracy[b]	Sensitivity	Specificity	PPV	NPV
Texture analysis with machine learning	89.1%	91.7%	86.4%	88.0%	90.5%
Neuroradiologist's review, ROI	56.5% (P = .0004)	54.2%	59.1%	59.1%	54.2%
Neuroradiologist's review, tumor	73.9% (P = .060)	75.0%	72.7%	75.0%	72.7%
Neuroradiologist's review, image	87.0% (P = .748)	91.7%	81.8%	84.6%	90.0

[a] Results are shown for the entire cohort (22 IPs, 24 SCCs) and reflect the best classification model. The labels for the neuroradiologists' assessment indicate whether they reviewed the 16x16 ROI (ROI), tumor alone (tumor), or entire images (image).
[b] P values represent comparison of texture analysis with machine learning against each neuroradiologist's review using a 2-tailed test of population proportion.

appearance, texture analysis stands to provide incremental benefit when human pattern recognition becomes most limited, and this can occur with a small tumor. For example, Maroldi et al. [111] found it more challenging to recognize the convoluted cerebriform pattern on T2-weighted images for tumors of 2 cm. While tumors smaller than 1.5x1.5 cm were excluded from enrollment in the current study, the final processed ROIs were only 1.125x1.125 cm.

Because small noninvasive sinonasal tumors are not universally imaged with MR imaging, the greatest potential benefit for texture analysis might be in detecting a small focus of SCC within a larger IP to expedite patient management. Accurately assessing small regions would be a prerequisite for the detection of such tumor heterogeneity. The potential for interpretive error is greatest when a small focus of SCC exists within a much larger IP and goes unrecognized because a convoluted cerebriform pattern is still present. Indeed, this pattern of a "partial" convoluted cerebriform pattern has been described [111, 114, 137]. Likewise, necrosis, recognized as nonenhancing tissue on contrast-enhanced MR imaging, is associated with SCC but may not be apparent when a small focus of SCC coexists with an

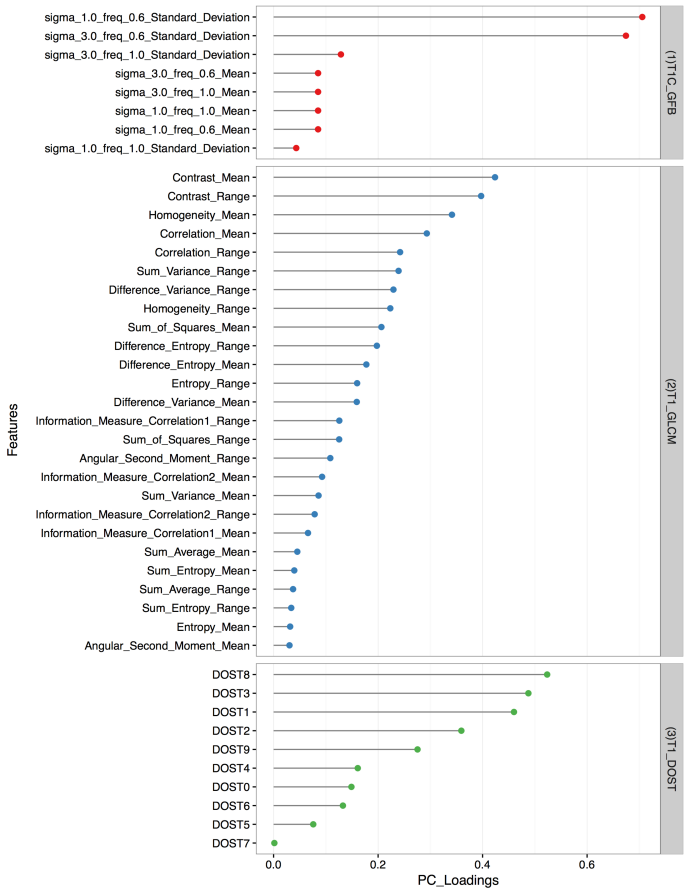


Figure 4.4: PC loading. The model with the greatest accuracy for discriminating SCC from IP was derived from T1C-GFB, T1-GLCM, and T1-DOST texture features (right). For the individually specified texture features (left), PC loadings are graphically represented, and larger values in the PC loading indicate greater significance in the final model.

IP [114, 115, 137]. A future goal for texture analysis of a mixed tumor containing both IP and SCC is to assist with interpretation by highlighting areas that are most suspicious for SCC.

Texture analysis can also extract useful features from images that have been traditionally neglected by the human eye. The convoluted cerebriform pattern has been historically described on T1-weighted postcontrast and T2-weighted sequences [111, 114–116, 135–137]. However, noncontrast T1-weighted MR imaging has received no attention to date, to our knowledge. The texture analysis in the current study found more significant features for

T1 than T2 on a univariate basis (Figure 4.2). Although T1-DOST and T1-GLCM made a minority contribution to the final model, no T2 features contributed to final model accuracy (Figure 4.3).

For MR imaging, it has been suggested that texture analysis models may not effectively translate across different imaging protocols and scanner platforms [138, 139]. Certainly, this possibility would make such results clinically meaningless because a new model would have to be created for each scanner running a unique protocol. Fruehwald-Pallamar et al. [140] concluded that texture analysis is not practical for differentiating malignant and benign tumors of the head and neck when using different protocols on different MR imaging scanners. However, their cohort was very heterogeneous, containing numerous types of benign and malignant lesions. The subjects for the current study were accrued during a long period and were not imaged with a common scanner and protocol. Nevertheless, an accurate texture-based model was achieved that performed similarly in the training and validation datasets. At least for the context of sinonasal IP and SCC, this outcome holds promise for reproducibility across scanner platforms.

The current study is limited, given its retrospective nature and small sample size. Hence, the high accuracy for the differentiation of SCC from IP with texture analysis is not meant to represent the performance of an established diagnostic imaging test. Instead, these results merely confirm the feasibility of this technique for distinguishing these 2 tumor types. In showing proof of concept, a 2D ROI-based analysis was used to confirm discriminatory ability with a limited data sample. Moreover, because SCC tends to be a larger tumor than IP on average, this approach eliminated the potential for falsely finding texture differences on the basis of relative oversampling of a larger tumor. Future directions will include the refinement of the texture analysis pipeline into a volumetric tool with the objective of highlighting foci of SCC when it is coexistent with IP. This will need to be studied prospectively to ensure that the histopathologic analysis can be accurately coregistered to MR imaging.

4.7 Conclusion

With MR imaging-based texture analysis, a machine-learning model for the differentiation of sinonasal SCC and IP achieved accuracy comparable with both neuroradiologists' interpretation and previously published reports on the convoluted cerebriform pattern. Because the classification model was significantly more accurate than the neuroradiologists' interpretation for a small ROI, texture analysis has the potential to provide incremental benefit to the neuroradiologists' interpretation, particularly in cases of small or heterogeneous tumors.

4.8 Acknowledgement

This work was supported by the Arizona State University/Mayo Clinic Seed Grant Program.

Chapter 5

VALIDATION STUDY II: COMPUTER-AIDED DIAGNOSIS OF CONTRAST-ENHANCED SPECTRAL MAMMOGRAPHY – A FEASIBILITY STUDY

5.1 Summary

This chapter presents a retrospective study to investigate the utility of quantitative analysis of contrast-enhanced spectral mammography images in improving the accuracy of breast cancer screening. This work has been accepted for presentation at RSNA 2017. The manuscript is currently in the second round of review by European Journal of Radiology. I contributed to this manuscript in 1) devising technical design of the study, 2) conducting the image analysis processes in order to convert images to mineable features, 3)implementing the machine-learning and data analysis techniques, and 4) writing and editing the manuscript.

Authors Sara Ranjbar[1*], Bhavika K. Patel[2*], Adrian M. Miller[2], Teresa Wu[3], Jing Li[3], Marc Lobbes[4], Barbara Pockaj[2], and J Ross Mitchell[5] (*authors contributed equally).

Affiliations [1]Department of Biomedical Informatics, Arizona State University, Tempe, AZ, [2]Division of Breast Imaging, Department of Radiology, Mayo Clinic, Phoenix, AZ, [3]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, [4]Department of Radiology and Nuclear Medicine, Maastricht University Medical Center, Maastricht, The Netherlands, [5]Department of Research, Mayo Clinic, Phoenix, AZ.

5.2 Abstract

Objective: To evaluate whether the use of a computer-aided diagnosis contrast-enhanced spectral mammography (CAD-CESM) tool can further increase the diagnostic performance of CESM compared with that of experienced radiologists.

Materials and methods: This IRB-approved retrospective study analyzed 50 lesions described on CESM from August 2014 to December 2015. Histopathologic analyses, used as the criterion standard, revealed 24 benign and 26 malignant lesions. An expert breast radiologist manually outlined lesion boundaries on the different views. A set of morphologic and textural features were then extracted from the low-energy and recombined images. Machine-learning algorithms with feature selection were used along with statistical analysis to reduce, select, and combine features. Selected features were then used to construct a predictive model using a support vector machine (SVM) classification method in a leave-one-out cross-validation approach. The classification performance was compared against the diagnostic predictions of 2 breast radiologists with access to the same CESM cases.

Results: Based on the SVM classification, CAD-CESM correctly identifies 45 of 50 studies in the cohort, resulting in overall accuracy of 90.0%. Detection rate for the malignant group was 88.4% (3 FN cases) and 91.6% for the benign group (2 FP cases). Compared with the model, radiologist 1 had an overall accuracy of 78.0% and detection rate of 92.3% (2 FN cases) and 62.5% (9 FP cases) for the malignant and benign groups, respectively. Radiologist 2 had an overall accuracy of 86.0%, with detection rate of 100.0% and 70.8% (7 FP cases) for the malignant and benign groups, respectively.

Conclusions: The results of our feasibility study suggest that a CAD-CESM tool can provide complementary information to radiologists, mainly by reducing the number of false-positive findings.

Keywords: breast cancer, contrast enhanced digital mammography (CEDM), contrast

enhanced spectral mammography (CESM), quantitative image analysis (QIA), texture analysis, computer aided diagnosis (CAD)

5.3 Introduction

The primary goal of breast cancer screening is to reduce subsequent breast cancer mortality through early detection. However, FFDM's overall sensitivity of 75% to 85% [141] may be as low as 30% to 60% for dense-breasted women [142]. Additionally, many women receive false-positive diagnoses because the positive-predictive value of FFDM is 18% to 31% [143]. These women undergo unnecessary supplemental imaging and, occasionally, invasive procedures to evaluate the suspicious findings. Patient stress, inconvenience, cost, and potential for harm because of these additional procedures are concerning. Overall, supplemental imaging and biopsy affect 3.2 million American women at an estimated cost of $2.8 billion [144].

Dual-energy contrast-enhanced spectral mammography (DE-CESM or CESM) (also called dual-energy contrast-enhanced digital mammography [DE-CEDM]) improves the accuracy of breast cancer diagnosis [145]. CESM generates a low-energy mammographic image along with a recombined contrast-enhanced image, reflecting contrast accumulation within a breast (Figure 5.1). Breast regions with increased or leaky vasculature, two common characteristics of neoplasms, can be identified using intravenously administered, iodinated-contrast material, thus improving lesion detection and characterization [146]. Although studies have shown that CESM is superior to FFDM, specificity is estimated to be 58% to 70% [147], leaving room for further improvement in diagnostic accuracy, for example, by applying quantitative image analyses.

Quantitative image analysis is a topic of active research that includes well-established applications like computer-aided diagnosis (CAD) algorithms. CAD algorithms, when coupled with traditional mammography, have shown promise for identifying suspicious breast lesions

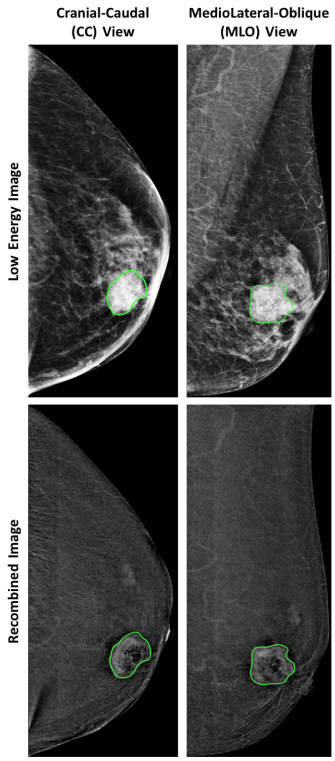


Figure 5.1: CESM images. A 64 year-old woman with a malignant appearing lesion on the left breast. This was later shown via biopsy to be invasive ductal carcinoma (IDC). Each lesion was captured in 2 views with 2 images per view, low-energy (LE) and recombined image. The green contours are the lesion contours marked by the radiologist.

[148]. These algorithms have evolved and are improved by leveraging large datasets generated from high-throughput sequencing experiments [149]. However, no studies thus far have evaluated the use of CAD algorithms in CESM.

Recently, there has been some interest in the use of texture features to distinguish benign and malignant lesions on magnetic resonance [150] and mammographic imaging [151]. We hypothesized that lesion texture and shape features can often capture often-missed information regarding the characteristics of a tumor and can provide details that have prognostic or diagnostic value [13]. These texture features describe intensity distributions within the lesion and capture spatial and spectral frequency patterns, as well as characterize the relationships between different intensity levels within the lesion. Some of these features

might not be visually apparent to the radiologists and therefore have the potential to complement the diagnostic skillset of radiologists. The purpose of our current study was to evaluate a prototype CAD-CESM tool, using texture and morphologic analysis to differentiate benign and malignant breast lesions.

5.4 Material and Methods

5.4.1 Study Population

The study was considered exempt by the Institutional Review Board (IRB). We retrospectively reviewed CESM examinations on a Selenia mammography system (Hologic, Bedford, MA). The examinations were performed between August 1, 2014, and December 31, 2015. Informed consent was obtained from all patients having a BI-RADS (Breast Imaging Reporting and Data Systems) [152] classification of 4 and 5 in a preexisting IRB-approved study to determine if CESM could lower the false-positive biopsy rates in mammography. As part of the aforementioned study, the radiologist counseled patients on the risks and benefits of biopsy and CESM. CESM was offered as an adjunct, not as an alternative to the recommended breast biopsy. The patient was informed that CESM might increase or decrease the level of suspicion of a lesion and might show additional suspicious areas. CESM was performed before the biopsy, typically on the diagnostic examination day or on the day of the biopsy.

The cohort for the current study included examinations that met the following criteria: (1) a diagnostic mammogram that received a BI-RADS rating of 4 or 5 and (2) studies that corresponded with available pathologic results from a surgical or image-guided biopsy. We limited the cohort to BI-RADS 4 and 5 lesions because the analysis required the criterion standard of lesion pathology. We identified 50 studies that met the above inclusion criteria, comprising 24 benign and 26 malignant, biopsy-proven lesions (Table 5.1). We analyzed 1 lesion per patient. If a patient had multiple enhancing lesions, the annotating radiologist

selected the largest lesion to ensure the best feature section. Statistical analysis was performed in Python 2.7.0 environment [99] using the open-source SciPy library [101].

5.4.2 CESM Image Acquisition Protocol

The patients were seated in the mammography suite to minimize vasovagal episodes. They received contrast via a single-lumen power injector: 1.5 mL/kg of Omnipaque 350 (GE Healthcare, Inc, Princeton, NJ) at a rate of 3 mL/second. We waited 2 minutes following the injection, compressing the breast and obtaining images exactly 2 minutes after contrast administration. The examination began with the mediolateral oblique (MLO) view of the affected breast because this view encompasses the most breast tissue. Next, images on the craniocaudal (CC) view were acquired. Image acquisition was completed within 7 minutes. The low-energy and recombined images were immediately available to an interpreting radiologist after the study.

5.4.3 Segmentation of Lesions and the Regions of Interest

All DICOM (Digital Imaging and Communications in Medicine) images were transferred to a database and loaded into the open source image processing tool OsiriX (OsiriX Foundation, Geneva, Switzerland) [153]. DICOM images were anonymized and prepared for blinded reading by a radiologist. A fellowship-trained breast radiologist with over 8 years of imaging experience interpreted the mammogram blindly and used the OsiriX tool to outline lesion contours. Contours were drawn on the CC and MLO recombined views for each patient. These contours were then cloned onto low-energy images. All lesions were visible in both views. For texture analysis, we required a fixed-sized rectangular area entirely contained within each lesion contour. From each enclosed region, we selected a 32x32 pixel area at the centroid of each rectangle as the region of interest (ROI) for feature extraction (Figure

Table 5.1: Summary of the study population

Biopsied Lesion	N (%)
All lesions	50
Benign cases	24 / 50 (48.0%)
- Fibroadenoma	6 / 24 (25.0%)
- Fibrosis	5 / 24 (20.8%)
- Intraductal papilloma	2 / 24 (8.3%)
- PASH	1 / 24 (4.2%)
- Fat necrosis	1 / 24 (4.2%)
- Sclerotic intraductal papilloma	1 / 24 (4.2%)
- Cyst	1 / 24 (4.2%)
Malignant cases	26 / 50 (52.0%)
- Ductal carcinoma in situ	6 / 26 (23.1%)
∗ Grade 1	1 / 26 (3.8%)
∗ Grade 2	1 / 26 (3.8%)
∗ Grade 3	4 / 26 (15.4%)
- Invasive ductal carcinoma	17 / 26 (65.4%)
∗ Grade 1	5 / 26 (19.2%)
∗ Grade 2	6 / 26 (23.1%)
∗ Grade 3	6 / 26 (23.1%)
- Invasive lobular carcinoma	3 / 26 (11.5%)
∗ Grade 1	1 / 26 (3.8%)
∗ Grade 2	2 / 26 (7.7%)

Abbreviation: PASH, PseudoAngiomatous Stromal Hyperplasia.

5.2). The size of the ROI was selected based on the size of the smallest lesion in the cohort, excluding areas with post biopsy image change.

5.4.4 Feature Computation

A set of four shape features were extracted per lesion contour [100]: shape bending energy, shape compactness, shape entropy, and shape radial-length ratio. Thus, 8 shape features were extracted per lesion (4 features/contour, 2 contours per lesion). Next, we used statistical and spectral methods [1, 17, 56] to measure texture at various spatial and spectral scales of the lesions. Feature calculation was performed using an in-house pipeline written in Python 2.7.0 [99]. A total of 236 texture features were computed from each subregion. Texture features included 3 from raw-image intensity, 26 from Gray Level Co-occurrence

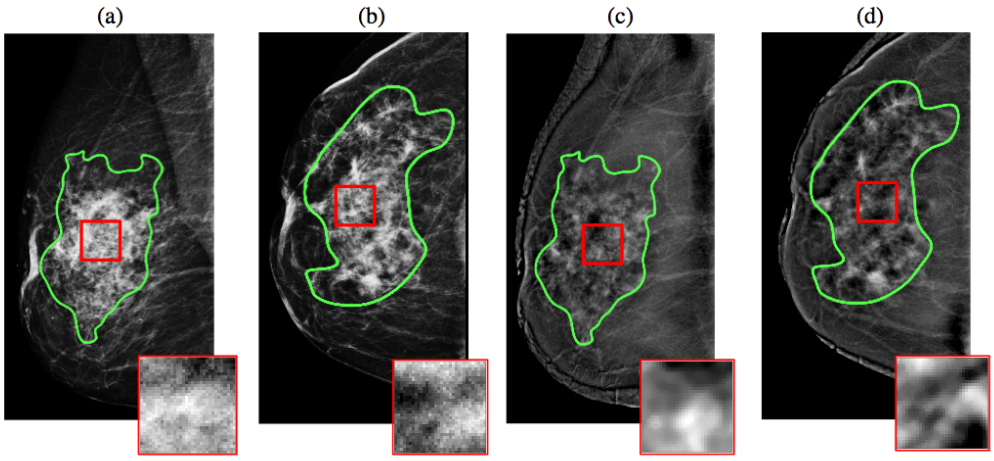


Figure 5.2: Regions of interest for shape and texture analysis. 61 year-old woman with a malignant appearing lesion on the right breast. (a) view: Mediolateral-Oblique or MLO, low-energy image (LE) - (b) view: Cranial-Caudal or CC, Low Energy image (LE), (c) view: MLO, recombined image and (d) view: CC, recombined image. ROIs are identical for the low-energy and contrast-enhanced images on each view. Shape features are extracted from contours (green) and texture features are extracted from a rectangular area inside each contour (red).

Matrices (GLCM) [126], 18 from Laplacian-of-Gaussian Histograms (LoGHist) [94], 126 from a Gabor Filter Banks (GFB) [93], 48 from Local Binary Patterns (LBP) [127], and 15 from the Discrete Orthonormal Stockwell Transform (DOST) [91]. Detailed descriptions of the texture features are provided in Table 5.2.

5.4.5 Feature Selection and Statistical Analysis

Feature selection, analysis, and model generation were performed using a Python-based platform developed in-house utilizing the Scikit-learn package [102]. We used a sequential forward feature selection (SFFS) approach in conjunction with quadratic discriminant analysis [104] to identify the most discriminating features. SFFS starts with an empty set of features and gradually adds features that offer minimal redundancy but maximal relevance. The setting used for quadratic discriminant analysis was no priors on classes and threshold of 0.0001 for rank estimation.This approach was repeated 100 times with a random-sampling

scheme to reduce training bias and improve model generalizability. In each trial, we randomly selected one-third of the benign and malignant samples and fed the combined samples to SFFS. The 10 most frequently appearing features among all iterations were selected for further analysis. Inclusion of more features did not improve classification accuracy.

To simplify results of the analysis, features resulting from a shared-texture method (eg, GLCM, GFB) were combined using principal component analysis (PCA) [104]. PCA creates linear combinations of features called principal components (PCs). One to three PCs were generated per feature group, sufficient to describe 90% of the feature variability. The final set of features was tested for statistical significance by using a 2-sample t test, with a P-value of less or equal to 0.05 considered significant. The significant features were used for classification.

5.4.6 Machine Learning Classification

The statistically significant features, described above, were provided to a Support Vector Machine (SVM) classifier 20 for construction of a lesion malignancy model. SVM model was build using a radial basis function kernel with coefficient value of 1 and gamma parameter equal to $1/numberoffeatures$. We utilized a leave-one-out cross-validation (LOOCV) technique [104] for model construction and estimation of classification performance. LOOCV iteratively sets aside 1 sample from the dataset as the test sample and uses the remaining samples as the training set for model development. The model is validated on the test sample in each trial. This process is repeated until every sample serves as the test sample. We used this approach to predict the malignancy status of the lesions. The predicted diagnosis across all test samples was used to estimate performance measures, such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve.

Table 5.2: Summary of the texture features used for breast lesion characterization

Method	Type of features	Details and Settings
Raw intensity	Statistical	Mean, range, standard deviation of the ROI scaled intensities before scaling the dynamic range to [0,255]
Gray Level Co-occurrence Matrices (GLCM) [126]	Statistical	Mean and range of 13 GLCM features at 4 angles (radius of 1).
Discrete Orthonormal Stockwell Transform (DOST) [91]	Spectral	Rotationally invariant dost features for all possible dyadic frequency bands insidethe ROI. With the ROI with the size of 32x32, 15 Dost features were extracted.
Laplacian of Gaussian Histogram (LoGHist) [94]	Spectral	Kurtosis, skewness, standard deviation, average intensity, entropy and uniformity of the histogram resulting from convolutions of band pass filtersand the ROI. Band pass filter sigma range (2.0, 4.0, 6.0).
Local Binary Patterns (LBP) [94]	Statistical	Values of histogram bins generated from intensity patterns found within the ROI. The intensity patterns are categorized into 'uniform' and 'non-uniform' based on the number of bit-wise transitions from 0 to 1 or vice versa within patches within the ROI. Radius of patches: (3.0, 5.0, 8.0, 12.0).
Gabor Filter Banks(GFB) [93]	Spectral	Gabor filters (sigma: [1.0 - 5.0], increments of 0.5, frequency: [0.2 - 0.8] increments of 0.1) were used to focus on different frequency resolution in the ROI. Mean and standard deviation of intensities of each filtered image was used as a GFB feature.

5.4.7 Comparative Human Reader study

Two breast-fellowship-trained radiologists with 5 years of total radiology experience and 6 months experience interpreting CESM images, not involved in other aspects of the study, reviewed the lesions to provide a non-computer-aided prediction of lesion pathology. Each lesion location was made apparent to the radiologist with the quadrant location written next to the case number. The images were reviewed according to standard diagnostic mammography workflow at our institution. Both breasts in the standard CC and MLO

views were read first as a diagnostic mammogram on dedicated Hologic workstations (and, if needed, associated ultrasound images) and decided on as a malignant or benign lesion. Next, the CESM images were reviewed independently by the readers and designated as malignant or benign based on low-energy and recombined images. The readers were blinded to the final pathology of the findings. No additional clinical information was provided to the readers. Next, performance measures (accuracy, sensitivity, and specificity) of each reader were calculated. The performance measures of the readers were compared to the performance measures of our automated model.

5.5 Results

The average age of patients was 57.1 years (52.2 years, patients with benign lesions; 61.9 years, patients with malignant lesions). The age difference between the 2 groups (9.7 years) was significant ($P<.05$). Twenty-six (52%) of the 50 biopsied lesions were diagnosed with breast cancer at histopathology. The remaining 24 (48%) lesions were diagnosed with benign pathology after core needle biopsy (Table 5.1).

Table 5.3 compares the performance measures of human readers and the computational model. Reader 1 provided a sensitivity of 92.3% (2 false-negatives) and a specificity of 62.5% (9 false-positives). Overall accuracy for Reader 1 was 78.0%. Reader 2 provided a sensitivity of 100.0% (0 false-negatives) and a specificity of 70.8% (7 false-positives). Overall accuracy for Reader 2 was 86.0%.

Differences between CAD-CESM and the readers are shown in Table 5.3. CAD-CESM provided a sensitivity of 88% (3 false-negatives from 26 malignant cases). Two false-negative results were invasive ductal carcinoma, and 1 was ductal carcinoma in situ. The false-negative lesions ranged from 7 to 12 mm in diameter. CAD-CESM provided a specificity of 92% (2 false-positives from 24 benign cases). Overall, CAD-CESM accuracy was 90%. The differences

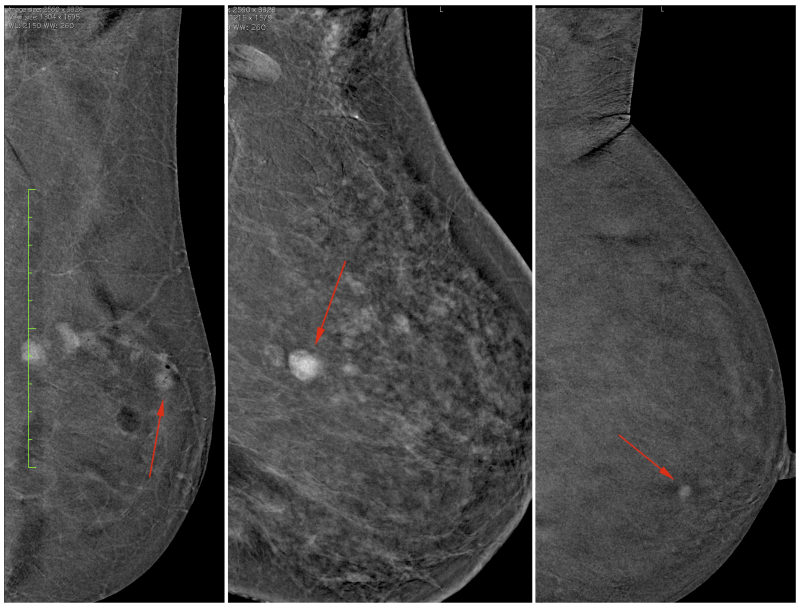


Figure 5.3: Benign lesions misclassified by the radiologists. Three benign lesions (red arrows) for which the label 'malignant' was assigned by the blinded human readers but not by the machine-learning developed quantitative model. Lesions are presented on the digital subtraction (DES) images Mediolateral-Oblique (MLO) views. The Reader's descriptions for these cases were as follows: (a) multiple avidly enhancing lesions, (b) avidly enhancing mass, satellite lesions, (c) obscured margins, not well defined.

Table 5.3: Exact McNemar's test between CESM and Reader 1/Reader 2

	Sensitivity	Specificity	Accuracy
Denominator	Malignant (N = 26)	Benign (N = 24)	Overall (N = 50)
CESM	88%	92%	90%
Reader 1	92%	63%	78%
Reader 2	100%	71%	86%
Pvalue1 [a]	>.99	0.016	0.109
Pvalue2 [b]	NA	0.063	0.726

Abbreviation: CESM, Contrast-Enhanced Spectral Mammography.

[a] Comparing CESM and Reader 1.

[b] Comparing CESM and Reader 2. Pvalue is NA because of 0 cell.

between CAD-CESM and the radiologists were statistically significant for reader 1 specificity (P=.016). The readers were marginally superior to the model in detecting malignant cases (0 and 2 misdiagnosed malignant cases for radiologists compared with 3 for the model). In comparison, benign cases were detected more successfully using the quantitative model. 7 and 9 benign cases were diagnosed as malignant by radiologists compared with only 3 for the model. Among the misclassified benign lesions by the readers, 3 cases were identical (Figure 5.3). The readers described these lesions using BI-RADS lexicon terms such as "obscured margin", "spiculated margin", "distortion", and "avidly enhancing", which typically portends a malignant diagnosis. The 3 lesions misclassified by the radiologists were, in fact, benign cellular fibroadenoma, pseudoangiomatous stromal hyperplasia (PASH), and fibrosis.

Table 5.4 shows the list of features selected by the SFFS process. Among the selected 10 features, 7 were extracted from the recombined image, 1 from a low-energy image; 2 were selected from the lesion shape features. After feature selection, PCA was used to combine the GFB, GLCM, DOST, and LoGHist texture features. Figure 5.4, A shows the result of univariate analysis (P<.05) on the texture PCs and morphology features. Four texture features survived the threshold of statistical significance, including 2 Gabor PCs (combination of the standard deviations of features at high (f=0.8) and medium (f=0.4) frequencies), 1 GLCM PC (combination of GLCM angular second moment, a measure of homogeneity of the ROI, and GLCM difference variance), and 1 medium-frequency content DOST feature. Among the morphology features, shape radial-length ratio corresponding to irregularity of lesion shape was also identified as statistically significant.

The CESM-CAD malignancy model was constructed using 3 of the 5 statistically significant features described above. The 3 features were Gabor PC2, Gabor PC3 – both texture features extracted from the recombined images – and the morphology feature for lesion radial-length ratio. The area under the ROC curve was 0.95. Figure 5.4, (B, C) shows the quantitative model ROC curve, as well as the contributions of shape and texture features to the model accuracy.

Table 5.4: Frequently selected features in repeated sequential forward feature selection algorithm on random samples

CESM Image and View	Feature Name	Description
MLO, DES	Standard deviation of Gabor filter sigma =1.0, frequency = 0.7	Spectral texture feature corresponding to smaller kernels that are sensitive to higher frequencies.
MLO, DES	Standard deviation of Gabor filter sigma = 3.0 frequency = 0.4	Spectral texture feature corresponds to kernels that are sensitive to medium frequencies.
CC, DES	Standard deviation of Gabor filter sigma = 3.5 frequency = 0.8	Spectral texture feature corresponds to smaller kernels that are sensitive to higher frequencies.
CC, LE	Dost5	Spectral texture feature corresponding to medium level frequencies.
CC, DES	Dost3	Spectral texture feature corresponding with low to medium frequencies.
MLO, DES	GLCM Angular Second Moment	Statistical texture feature, measure of orderliness in intensity values of neighboring pixels.
MLO, DES	GLCM Difference Variance	Statistical texture feature, measure of uniformity in the intensity of neighboring pixels.
MLO, DES	LoGHist Entropy with sigma = 4.0	Spectral Texture feature, measure of entropy of medium frequency band-pass filters.
MLO	Shape entropy	Morphology feature, Measure of predictability of the lesion's radial length [a].
CC	Shape radial length ratio	Morphology feature, Ratio of minimum to maximum radial length [a].

Key: CC, CranioCaudal; CESM, Contrast-Enhanced Spectral Mammography; DOST, Discrete Orthonormal Stockwell Transform; GFB, Gabor Filter Banks; GLCM, Gray Level Co-occurrence Matrices; LE, Low Energy; LoGHist, Laplacian-of-Gaussian Histogram; MLO, MedioLateral Oblique.
[a]Radial length = distance of boundary points from the center.

5.6 Discussion

The moderate specificity of mammography might result in additional imaging exams, costs, and patient anxiety, and may cause avoidable breast biopsies. CESM was introduced in 2011 to improve the diagnostic accuracy of breast imaging through mammography [154]. Although CESM is superior to FFDM, its specificity reaches a maximum of approximately

70% [155]. Therefore, CAD tools might be used as an adjunct to radiologists' assessments. In this study, we aimed to evaluate the performance of a CAD-CESM tool based on quantitative morphology and texture features extracted from lesions on the recombined CESM images. We showed that by using this tool, the accuracy of CESM could be further improved when compared with that of 2 experienced radiologists, especially regarding specificity of this diagnostic test.

CESM is a promising imaging technique, which provides information from standard digital mammography combined with enhancement characteristics related to underlying neoangiogenesis. CESM has been reported to offer improved sensitivity, specificity, and accuracy compared with conventional mammography [156]. Cheung et al. [144] investigated the performance of CESM vs mammography in dense breasts. Their results suggested that using CESM improves diagnosis by 21.2% in sensitivity, 16.1% in specificity, and 12.8% in overall accuracy. Tagliafico et al. [147] recently summarized the diagnostic performance of CESM in a systematic review of 8 eligible studies. The pooled sensitivity of CESM was 98% (95% CI, 96%-100%); with a pooled specificity being moderate at 58% (95% CI, 38%-77%). This moderate specificity is mainly explained by a preponderance of data from a Polish study group, which showed very poor diagnostic accuracies, even for FFDM. When the results of the review are recalculated with the results from the Polish group removed, the specificity increases to 78% (95% CI, 56%-90%). CAD tools might aid in further improving the specificity of CESM, preferably while not decreasing the already high sensitivity of CESM.

In our study, the model of the CAD-CESM tool was constructed by using lesion-shape irregularity and spectral texture features from an ROI. Lesion shape is one of the primary diagnostic factors used by radiologists. In fact, radiologists describe the morphology of lesions on imaging reports based on criteria suggested by BI-RADS shape vs margin [152]. The shape of a mass in radiologic reports is generally described as round, oval, or irregular, and the

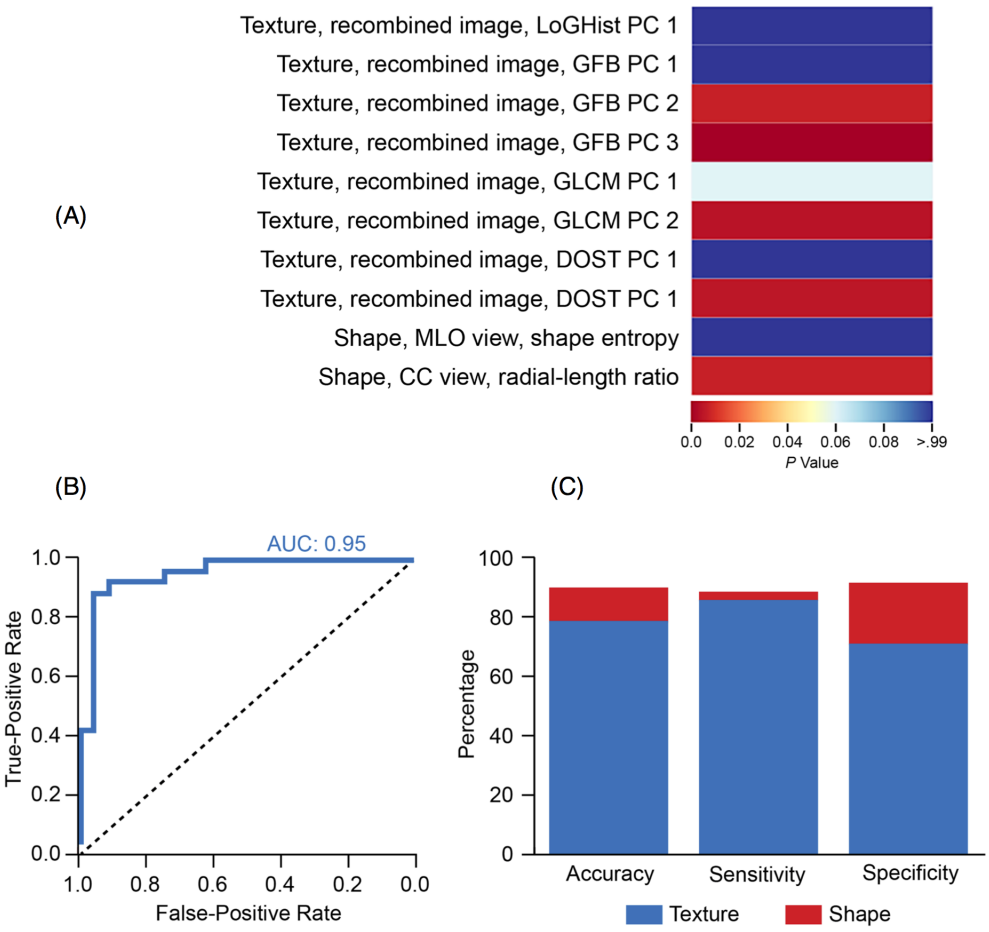


Figure 5.4: A) Univariate analysis of the selected imaging features compared with the pathologic finding (benign vs. malignant). We used a 2-tailed t test to evaluate the final set of features used for classification purposes. Red to white cells show significant features at P < 0.05 (corrected for false-discovery rate). Principal component analysis was only performed when multiple texture features resulted from the common texture analysis method. B) The receiver operating characteristic curve (solid line) demonstrates the performance of the automated model using machine-learning techniques coupled with contrast-enhanced spectral mammographic imaging features (area under the curve or AUC $= 0.95$). C) Across all features, the overall model accuracy (90%), sensitivity (88%), and specificity (92%) were predominantly contributed from texture features, with minor contributions from shape features. [Key: CC, CranioCaudal; DOST, Discrete Orthonormal Stockwell Transform; GFB, Gabor Filter Banks; GLCM, Gray Level Co-occurrence Matrices; LoGHist, Laplacian-of-Gaussian Histogram; MLO, MedioLateral Oblique; PC, Principal Component].

margins are categorized as circumscribed, obscured, microlobulated, indistinct, or spiculated. The positive predictive value for carcinoma of masses visible in digital mammograms with spiculated margins is 81% and with irregular shape is 73% [157].

Lesion shape has been used previously to predict malignancy [158]. We found benign lesions to be associated with more round shapes and malignant lesions to have more irregular shapes. Texture features that contributed to the model included GFB features. GFB uses filters that are capable of detecting patterns at a certain scale and orientation. Using an exhaustive approach, we generated a high number of GFBs to maximize the possibility of finding patterns present at any scale or orientation. We found different enhancing patterns for the benign and malignant groups on the recombined CESM images. These patterns were detectable via medium and high-frequency GFB features. Combining shape irregularity and texture feature, the model reached an area under the ROC curve of 0.95, a sensitivity of 88%, and a specificity of 92% for 50 lesions.

A strength of this study is the use of a multi-parametric texture analysis approach for feature extraction. This approach has been designed to simplify inclusion of additional computer vision algorithms and to maximize the chance of detecting informative features. It has been shown to be beneficial, especially when applied to multiple images with varying contrast characteristics [17, 56]. The texture analysis methods in our package have been previously cited in the literature in various applications, such as for classifying on magnetic resonance imaging [159] and computed tomography of the breast [160], assessing response to therapy [161], and predicting tumor invasion [162].

The modest sensitivity of CAD-CESM will be addressed in future models. Contrary to the reader study, the entirety of the lesions was not provided to the texture analysis pipeline. Regions used for texture analysis were small (ROI size: 32x32 pixels), with well circumscribed areas within the contours. The region size therefore was constrained by the largest available cross-sectional area across the cohort, as well as the presence of marks from

previous biopsies on some cases. An automated approach, using the geometric center of the lesion, placed the ROI at the center of the lesion for consistency purposes. Although this approach to selection of the region of interest provided us with a framework for automation, we suspect that the 2 factors (size and location of the ROIs) might have caused the false negative cases. In future work, we aim to address these limitations by pursuing a global lesion texture analysis approach.

Our study has some limitations. We acknowledge a high prevalence of malignancy in our study (52%). While our result shows great promise, future plans will include a validation study in which we will test CAD in a larger population with other disease prevalence and preferably in images from multiple contrast enhanced mammography vendors to improve generalizability. We also plan to include and report on how this tool compares between variable BI-RADS breast densities. We developed a preliminary computer-aided diagnosis tool for contrast-enhanced spectral mammography (CAD-CESM) to test whether there is potential to reduce false positives in breast cancer screening. CESM-based texture features complemented the lesion boundary features in predicting lesion malignancy status. The computational model showed higher specificity and accuracy than the human reader using BI-RADS descriptors suggesting the potential of CESM-CAD (computer aided detection) systems to improve breast cancer detection specificity. Although the sensitivity of the model was lower than human readers, future work will include margin assessments which may improve sensitivity of the CAD-CESM.

While our result shows great promise, it will need to be evaluated with external validation datasets, including more patients. We did not assess the impact of contouring variability on the model accuracy. In future work, we intend to implement automatic lesion segmentation to eliminate any reader bias in drawing the ROI. Also, we did not include lesion margin descriptors in this preliminary study. Earlier reader studies have shown a negative predictive value of 80% to 84% for circumscribed masses and a positive predictive value of 90% to 93%

for noncircumscribed masses [163]. In the future, we intend to extend our image analytics package to include margin descriptors, which we hope will improve sensitivity of the model.

5.7 Conclusion

We developed a preliminary computer-aided diagnosis tool for contrast-enhanced spectral mammography (CAD-CESM) to test whether there was potential to reduce false-positives in breast cancer screening. CESM-based texture features complemented the lesion boundary features in predicting malignancy status of the lesion. The computational model showed higher specificity and accuracy than the human readers using BI-RADS descriptors, suggesting the potential of CAD-CESM (computer aided detection) systems to improve breast cancer detection specificity. Although the sensitivity of the model was lower than that of the human readers, future work will include margin assessments, which may improve the sensitivity of the CAD-CESM.

5.8 Achnowledgement

We would like to acknowledge the Mayo Clinic-ASU Seed grant for funds for this project. We would also like to thank Yania Lopez-Alvarez, MD, for participating as a reader in this study and Jennifer Palmieri, RT, and Nan Zhang, PhD, for assistance with the data analysis portions of this manuscript.

Chapter 6

VALIDATION STUDY III: COMPUTED TOMOGRAPHY BASED TEXTURE ANALYSIS TO DETERMINE HUMAN PAPILLOMAVIRUS STATUS OF OROPHARYNGEAL SQUAMOUS CELL CARCINOMA

6.1 Summary

This chapter reports on a retrospective study to determine whether machine learning can accurately classify human papillomavirus (HPV) status of oropharyngeal squamous cell carcinoma (OPSCC) using computed tomography (CT)-based image features. This work has been accepted by Journal of Computed Assisted Tomography (JCAT). My contributions to this manuscript are 1) devising technical design of the study, 2) implementing the methods section, 3) implementing the machine learning process, 4) producing the results, and 5) writing and editing the manuscript.

Authors Sara Ranjbar[1], Shuluo Ning[2], Christine M. Zwart[3], Christopher P. Wood[4], Steven M. Weindling[5], Teresa Wu[2], J Ross Mitchell[6], Jing Li[2], and Joseph M. Hoxworth[3].

Affiliations [1]Department of Biomedical Informatics, Arizona State University, Tempe, AZ, [2] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, [3]Department of Radiology, Mayo Clinic, Phoenix, AZ, [4]Department of Radiology, Mayo Clinic, Rochester, MN, [5]Department of Radiology, Mayo Clinic, Jacksonville, FL, [6]Department of Research, Mayo Clinic, Phoenix, AZ.

6.2 Abstract

Objective: To determine whether machine learning can accurately classify human papillomavirus (HPV) status of oropharyngeal squamous cell carcinoma (OPSCC) using computed tomography (CT)-based texture analysis.

Methods: Texture analyses were retrospectively applied to regions of interest from OPSCC primary tumors on contrast-enhanced neck CT, and machine learning was used to create a model that classified HPV status with the highest accuracy. Results were compared against the blinded review of 2 neuroradiologists.

Results: The HPV-positive (n = 92) and -negative (n=15) cohorts were well-matched clinically. Neuroradiologist classification accuracies for HPV status (44.9%, 55.1%) were not significantly different (P = 0.13), and there was a lack of agreement between the two neuroradiologists (kappa = -0.145). The best machine learning model had an accuracy of 75.7%, which was greater than either neuroradiologist (P<0.001, P = 0.002).

Conclusions: Useful diagnostic information regarding HPV infection can be extracted from the CT appearance of OPSCC beyond what is apparent to the trained human eye.

Keywords: Squamous cell carcinoma; human papilloma virus; oropharynx; oropharyngeal cancer; texture analysis; machine learning; radiomics

6.3 Introduction

Infection with high risk human papillomavirus (HPV), particularly HPV16, has become increasingly recognized as a risk factor for oropharyngeal squamous cell carcinoma (OPSCC), and the prevalence of HPV-positive OPSCC has dramatically risen in recent decades [164, 165]. This represents a clinically distinct group comprised of predominantly younger male patients, many of whom lack the typical risk factors of alcohol consumption and cigarette smoking. Because of improved prognosis and treatment response over HPV-negative OPSCC,

establishing HPV-positivity is clinically relevant but requires tissue sampling [166, 167]. Given the complex molecular mechanisms underlying HPV-induced carcinogenesis, HPV infection can be determined pathologically with a variety of techniques that most commonly focus on p16 overexpression or viral DNA detection [168, 169].

Oropharyngeal squamous cell carcinoma demonstrates significant intratumoral genomic heterogeneity [170]. This is particularly relevant for patients undergoing primary chemoradiation whose diagnosis may be based on limited biopsy samples, as incomplete tumor characterization could lead to treatment resistance. The field of radiomics offers the potential to help in this regard by identifying the spatial distribution of tumor phenotypes and genotypes on conventional radiology images [171]. This process involves extracting mineable data from the complex visual pattern within a biomedical image by using histogram, texture, and higher order statistical analyses in an attempt to mathematically characterize subpatterns and similarities (or dissimilarities). Because of the high dimensional nature of this feature data, machine learning methodology is well suited for data analysis and classification.

Computed tomography CT-based texture analysis was previously used in OPSCC to identify individual image feature differences based on HPV status [172]. To advance this field, the primary aims of the current study were to determine if CT-based texture analysis and machine learning can accurately classify the HPV status of OPSCC and to compare this classification result to that of the neuroradiologist.

6.4 Materials and Methods

This retrospective study is compliant with the Health Insurance Portability and Accountability Act and was approved by the Institutional Review Board at our institution with waiver of informed consent.

6.4.1 Study Cohort

The pathology database was queried from January 1, 2010 through December 31, 2014 to find adult patients (age $\geqslant$ 18 years) who underwent surgical resection of OPSCC originating in the tonsil or base of the tongue, and patients were excluded if the OPSCC originated in any other oropharyngeal subsite. Subject demographic and clinical data were obtained from the electronic medical record and included age at diagnosis, gender, subsite of tumor origin within the oropharynx, pathologic tumor T stage [173], tumor histologic type, and HPV-status. The HPV testing was performed as part of standard clinical care utilizing DNA in situ hybridization, and "HPV-positive" indicates positivity for one of more of the following high risk HPV types: 16, 18, 31, 33, or 51. The OPSCC had to be of keratinizing or non-keratinizing histologic type, and less common squamous cell carcinoma variants such as papillary, small cell, undifferentiated, adenosquamous, and spindle cell were excluded [174]. Based upon review of the medical record, subjects who had any history of prior head and neck tumor, surgery, or irradiation were also excluded. Potential subjects meeting the above criteria were cross-referenced against the radiology information management system to identify those subjects who had a pretreatment contrast-enhanced neck CT electronically archived and available for review. This initial study cohort consisted of 131 HPV-positive and 23 HPV-negative subjects, which is concordant with the high rate of HPV-related OPSCC found in our practice.

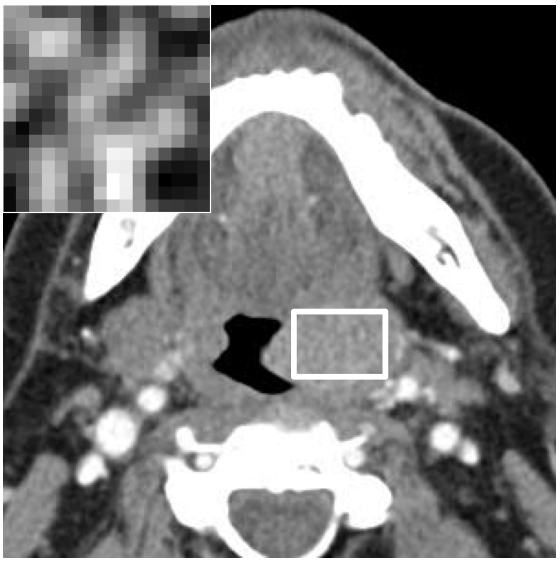
To maximize generalizability of the study results, there was no requirement for the use of a single CT scanner type, common imaging protocol, or standardized bolus of intravenous contrast administration. Neck CT scans were eligible for inclusion even if performed at an outside facility and, in fact, 60 of 107 neck CT's (56.1%) were performed before patient referral to our institution. For inclusion, the contrast-enhanced neck CT had to include axial images with a slice thickness ranging from 1.5 mm to 3 mm, and the images had to qualitatively be rendered with a smooth reconstruction kernel. The pretreatment contrast-enhanced neck CT was reviewed for each potential subject by a board-certified neuroradiologist with 11 years of

experience. To permit reliable region of interest (ROI) assessment, it was required that the primary oropharyngeal tumor had to be clearly discernible on CT with minimum orthogonal measurements of 1x1 cm^2 on at least one axial image, leading to the further elimination of 22 cases (19 HPV-positive and 3 HPV-negative). Subjects for whom the primary tumor assessment on neck CT was degraded by patient motion or beam-hardening streak artifact were also excluded (17 HPV-positive, 4 HPV-negative). Lastly, tumors containing any air due to ulceration were removed (3 HPV-positive, 1 HPV-negative). The final subject cohort therefore consisted of 107 cases, 92 HPV-positive and 15 HPV-negative cases, whose neck CT scans were then anonymized to undergo texture analysis in a blinded fashion.

6.4.2 Texture Analysis

The DICOM imaging data was imported into OsiriX (version 6.5; Pixmeo SARL, Bernex, Switzerland) and, to account for differences in imaging protocols, the images were resampled to have uniform 0.7mm x 0.7 mm pixel spacing and 512x512 matrix size. Because ROI assessment was performed by the same neuroradiologist who screened the neck CT's for study eligibility, a period of two months was allowed to elapse between CT anonymization and further imaging evaluation to preserve blinding and eliminate potential bias. For descriptive purposes, tumor margins were manually traced on all axial images, and tumor volume was estimated by the "ROI Volume" function in OsiriX. In order to follow a standard approach, the axial image with the largest tumor cross-sectional area was selected for each subject as the basis for texture analysis. The neuroradiologist then inserted the largest possible rectangular ROI that would fit within the margins of each tumor (Figure 6.1, b). To avoid the potentially confounding effect of oversampling larger tumors, an automated script determined the largest common square ROI that would fit within the borders of the manually placed rectangular ROIs for all 107 cases. This least-common square ROI, which was found to be 16x16 pixels, was automatically centered within each respective rectangular ROI and served as the input for two-dimensional texture analysis.

(a) A 55 year-old man with HPV-negative squamous cell carcinoma of the left palatine tonsil (T2N1M0).



(b) A 70 year-old man with HPV-positive squamous cell carcinoma of the left palatine tonsil (T2N0M0).

Figure 6.1: Regions of interest for texture analysis. White boxes represent the largest manually drawn rectangular ROI that would fit within the confines of the tumor. The inset images in the upper left corner of each image are the 16x16 matrices obtained from the isocenter of the manually drawn ROI and served as the input for 2-dimensional texture analysis.

The texture operators used in this study have been previously described [17, 56] and include first-order textural features, second-order Ggray Level Co-occurrence Matrices (GLCM)[126], Laplacian of Gaussian Histogram (LoGHist)[94], Discrete Orthonormal Stockwell Transform (DOST)[91], Gabor Filter Banks (GFB)[93], and Local Binary Patterns (LBP)[127], summing up to a total of 77 features per patient that were exported for lesion classification. The texture analyses were implemented in Python programming language using custom written code based on these previous publications and open source libraries.

First, mean, standard deviation, and range of the gray level values across the ROI in the original image (i.e. 'raw' features) were reported as the first order statistics of texture. Next, the gray value dynamic range of the ROI was reduced to 8 bits per pixel intensity levels for computational feasibility of analysis. GLCM is a well-established statistical method

of extracting second-order statistics of texture [126]. GLCM features such as entropy, energy, angular second moment, homogeneity, dissimilarity, and correlation are descriptors of randomness/orderliness, homogeneity, or dependency of gray-tones of the ROI to each other. GLCM features corresponding to a distance of one pixel and four angles (0°, 45°, 90°, and 135°) were calculated. Mean and range of each feature over the four directions were calculated and exported for a total of 26 GLCM features. The LBP method finds the most frequent intensity patterns within a certain radius of each pixel in the ROI [127], and a radius of three pixels was selected to focus on medium textures for a total of 12 features per sub image.

Transform-based texture analyses such as DOST, LoGHist, and GFB capture the spectral composition of the lesion in intermediate scales not accessible to first and second order statistical methods. DOST is a rotationally-invariant multi-resolution spatial-frequency representation of an image based on dyadic sampling of Fourier representation of the image [91], while LoGHist uses different sizes of band-pass filters to highlight and enhance different scales of texture ranging from fine to coarse details [94]. Gaussians with three different values of sigma (2.0, 4.0 and 6.0) were used to cover the range of fine to medium scale textures. The GFB method uses localized filters to capture details in various frequency resolutions [93]. Using two levels of sigma (1.0 and 3.0) and two levels of frequency (0.6 and 1.0), four different Gabor filters were applied with mean and standard deviation of the filtering result determined. In total, 10 DOST, 18 LoGHist, and 8 GFB features were calculated, yielding a total of 36 spectral texture features per sub image.

6.4.3 Neuroradiologist Qualitative Classification

In order for texture analysis to become a useful tool to assist with clinical image interpretation, it needs to detect tumor characteristics not well appreciated by radiologists. Although differences have been described between HPV-positive and HPV-negative OPSCC, these were related to features, such as muscle invasion, exophytic morphology, and cystic

nodal metastases [175, 176]. To date, no qualitative features (i.e. necrosis or enhancement pattern) limited to the internal tumor appearance have been reported as a predictor of HPV status in OPSCC [175]. To determine if texture analysis could provide useful information beyond what was obvious to the trained human eye, two board-certified neuroradiologists, with 25 and 28 years of experience, respectively, reviewed the tumors for all 107 subjects in a blinded fashion. Using OsiriX, they independently assessed the primary tumors with any window, level, and magnification adjustments that they deemed necessary and subsequently rendered an opinion of HPV-positive or -negative for each case. Of note, the CT images were zero-filled outside of the ROIs that had been manually traced around the tumor margins thereby preventing ancillary features such as tumor location, invasive behavior, and cystic metastatic lymph nodes from being taken into account. This was performed to allow an isolated comparison based exclusively on the internal appearance of the tumor, because this was the only aspect being interrogated with texture analysis. Again, because no differentiating features have been published in this regard, the neuroradiologists were not given any explicit guidelines on image review, other than to use their best judgment based on decades of experience in head and neck radiology.

6.4.4 Machine Learning and Statistical Analysis

Open-source statistical R software (`http://www.R-project.org`) was used to perform the analyses. Classification of OPSCC HPV status was achieved using widely accepted machine learning methodology, which was performed utilizing the open-source python-based package scikit-learn [102] in conjunction with custom written code. The texture features were first standardized by removing the mean and standard deviation and scaling them to unit variance. Principal component analysis (PCA) was used to improve model generalizability to unseen cases and to reduce data dimension [177–179]. Principal component analysis creates principal components (PCs), which are linear combinations of lesion features that best explain the variability in the dataset. To aid with clinical interpretation, PC sets were

constructed from each texture analysis algorithm separately resulting in 5 sets of PCs. One to 3 PCs per subset, sufficient to describe 90% of the subset variability, were selected for further processing. Several comparative datasets were generated with PCA to find the optimal level of variance in the PCs for gaining the highest accuracy while avoiding over-fitting. The resulting PCs were then merged with raw first order features as the final set predictors in the analysis. HPV status was used as the class label for supervised machine learning.

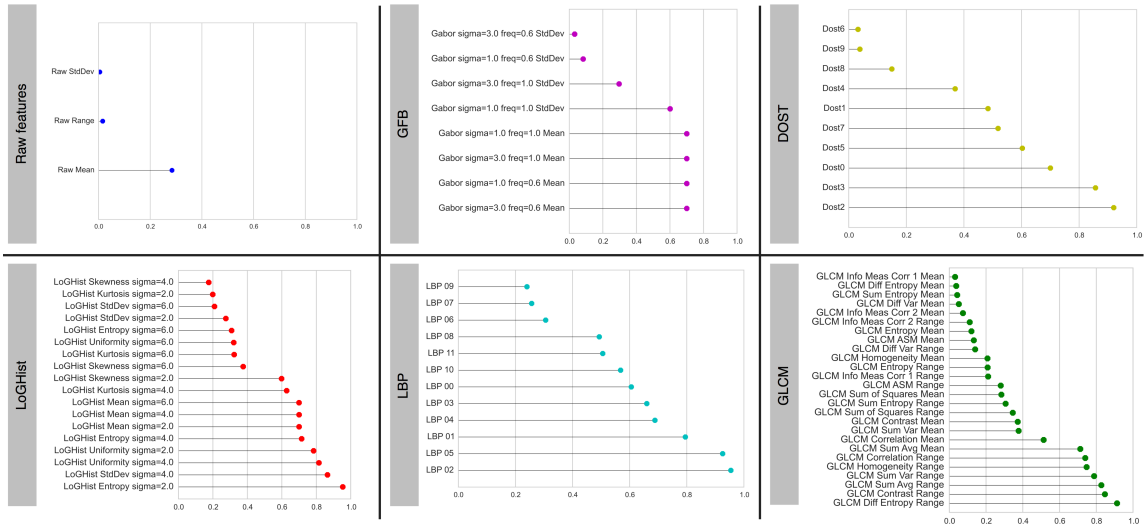


Figure 6.2: Univariate comparison of texture features, which are adjusted for multiple comparisons using the False Discovery Rate, are plotted for comparison of texture features between HPV-positive and -negative OPSCC tumors using the two sample t test.

Based on prevalence within our practice, there was a domain-specific reason for dataset skewness towards the HPV-positive class. With 86% HPV-positive samples, imbalanced class distributions rendered standard machine learning algorithms unsuitable. Instead, a sampling approach was used to ensure that the training model had the opportunity to learn the inherent properties of each class without the dominance of the majority class [180, 181]. An algorithm was designed that utilized a leave one out cross validation technique combined with ensemble down-sampling. In each round, one sample was selected and set aside as the test sample, while the remaining samples were used for training. The majority class was split into 6 randomly selected non-overlapping subsets with sizes roughly matching that of

the minority class. Each of these subsets together with the minority class was used to train a classifier, rounding up to a total of 6 classifiers per test sample. The class label for the training task was HPV status. Diagonal quadratic discriminant analysis, which was used to perform supervised classification for each batch of samples [133, 182], is a parsimonious model that avoids overfitting and has elicited good performance for high dimensional data with small sample sizes [183–185]. The setting used for quadratic discriminant analysis was no priors on classes and threshold of 0.0001 for rank estimation.

Table 6.1: Subject Demographics and Tumor Features

	HPV-Positive (n = 92)	HPV-Negative (n = 15)	P Value
Gender			0.66[a]
Male	86 (93.5%)	14 (93.3%)	
Female	6 (6.5%)	1 (6.7%)	
Age	54.6 ± 9.1	65.7± 13.4	0.007[b]
Squamous Cell Carcinoma Histology			< 0.001[a]
Keratinizing	11 (12.0%)	12 (80.0%)	
Non-keratinizing	81 (88.0%)	3 (20%)	
Oropharyngeal Subsite of Tumor Origin			0.48[c]
Base of the Tongue	40 (43.5%)	8 (53.3%)	
Tonsil	52 (56.5%)	7 (46.7%)	
Tumor Volume	19.5 ± 13.4 cm^3	23.0 ± 17.9 cm^3	0.38[b]
Tumor T Stage			0.91[d]
T1	10 (10.9%)	1 (6.7%)	
T2	36 (39.1%)	7 (46.7%)	
T3	15 (16.3%)	3 (20.0%)	
T4	31 (33.7%)	4 (26.7%)	

Age and tumor volume are presented as mean ± standard deviation. All other values are numbers of subjects in each category with percentages included as parentheticals. The reported P values represent comparisons using the Fisher exact test[a], two sample t test[b], Pearson chi-square test[c] , and the Freeman-Halton extension of the Fisher exact test[d].

Tumor T stage reflects the American Joint Committee on Cancer staging system (7th edition) [173], which was in effect during the study period.

Selecting features for classifier construction is an important issue in the machine learning process. The sequential forward feature selection algorithm was utilized, as this selects features that have minimal redundancy yet are maximally relevant for prediction. PCs were

selected if they improved the classifier's performance by at least 5%, as estimated using leave one out cross validation [45]. After training, the 6 models were used to predict HPV status in the test sample using the average posterior probability of the models as the final prediction. Receiver operating characteristic curves were used to evaluate classification performance metrics with a higher percentage of the area under the curve indicative of better model performance.

Additional statistical analyses were performed to assess for potential differences between the OPSCC HPV-positive and -negative groups. Differences in mean age and tumor volume were assessed using the 2-sample t test. Frequencies for gender and tumor histologic type were compared using the Fisher exact test, while the Pearson chi-square test was employed for tumor subsite (i.e. base of the tongue versus tonsil). Tumor T stage frequencies were evaluated using the Freeman-Halton extension of the Fisher exact test. The 2-sample t test was used for a univariate comparison of texture features between HPV-positive and -negative tumors prior to model development, and P values were corrected for multiple comparisons using the False Discovery Rate [186]. Cohen's κ coefficient was calculated to measure inter-rater agreement for the prediction of HPV status by the 2 blinded neuroradiologists, while potential differences in accuracy were assessed with a 2-tailed test of population proportion. All hypothesis tests were 2-sided, and statistical significance was defined as P less than 0.05.

6.5 Results

6.5.1 Study Cohort

Demographic and tumor characteristics are summarized in Table 6.1. The HPV-positive and -negative cohorts were similarly male predominant, while the subjects with HPV-positive tumors were significantly younger on average. No statistically significant differences were identified for tumor volume, T stage, or oropharyngeal subsite of tumor origin based on HPV status. Concordant with previous publications, conventional keratinizing squamous

Table 6.2: Diagnostic Performance of Texture Analysis With Machine Learning Compared With Neuroradiologist Review for the Differentiation of HPV-Positive from HPV-Negative OPSCC

	In-Situ Hybridization HPV-positive (n = 92)	HPV-negative (n = 15)		
Texture Analysis (Diagonal Quadratic Discriminant Analysis)				
			Accuracy	[ab]75.7%
HPV-positive	69	3	Sensitivity	75.0%
			Specificity	80.0%
HPV-negative	23	12	PPV	95.8%
			NPV	34.3%
Neuroradiologist #1				
			Accuracy	[ac]44.9%[a]
HPV-positive	44	11	Sensitivity	47.8%
			Specificity	26.7%
HPV-negative	48	4	PPV	80.0%
			NPV	7.7%
Neuroradiologist #2				
			Accuracy	[bc]55.1%
HPV-positive	53	9	Sensitivity	57.6%
			Specificity	40.0%
HPV-negative	39	6	PPV	85.5%
			NPV	13.3%

Comparison of accuracies was performed using a 2-tailed test of population proportion:
[a]$P < 0.001$, [b]$P = 0.002$, [c]$P = 0.13$.
PPV indicates positive predictive value; NPV, negative predictive value.

cell carcinoma predominated in HPV-negative OPSCC, while the majority of HPV-positive tumors were non-keratinizing [174].

6.5.2 Neuroradiologist Qualitative Classification

Neuroradiologist classification results for tumor HPV status are documented in Table 6.2. The respective accuracies of 44.9% and 55.1% were not significantly different ($P = 0.13$). The percent agreement between the 2 raters was 43.0%, while the κ statistic was -0.145 indicating a lack of agreement between the 2 neuroradiologists.

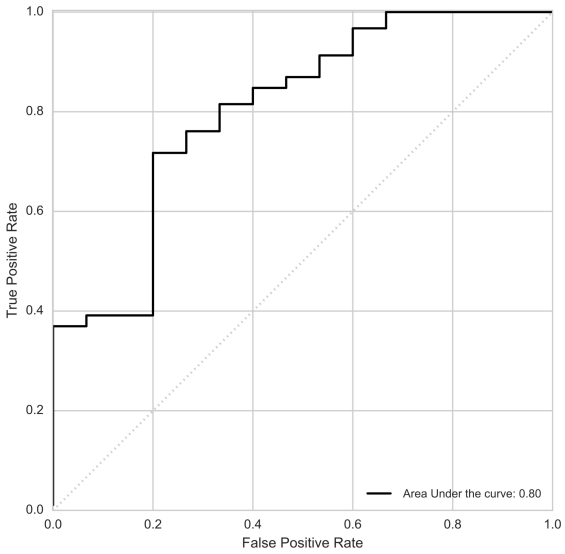


Figure 6.3: Model performance depicted by receiver operating characteristic curve. Classifiers were trained to predict tumor HPV status, and average posterior probability of models was used as the prediction for each classifier. Area under the curve for the best diagonal quadratic discriminant analysis model was 0.80.

6.5.3 Texture Analysis and Machine Learning

Based on a univariate comparison, all forms of texture analysis had at least some features that were significantly different between HPV-positive and -negative OPSCC ($P < 0.05$), except for LoGHist and LBP (Figure 6.2). The performance of the best diagonal quadratic discriminant analysis model is depicted by the receiver operating characteristic curve in Figure 6.3, which had an area under the curve of 0.80. The model was comprised of texture features with the highest predictive value for classification of HPV status. Figure 6.4 shows how different texture analysis methods compare in terms of their relative contribution to accuracy across all models, with the raw features contributing the least. The best-performing model from diagonal quadratic discriminant analysis was significantly more accurate as compared to the performance of either neuroradiologist (Table 6.2).

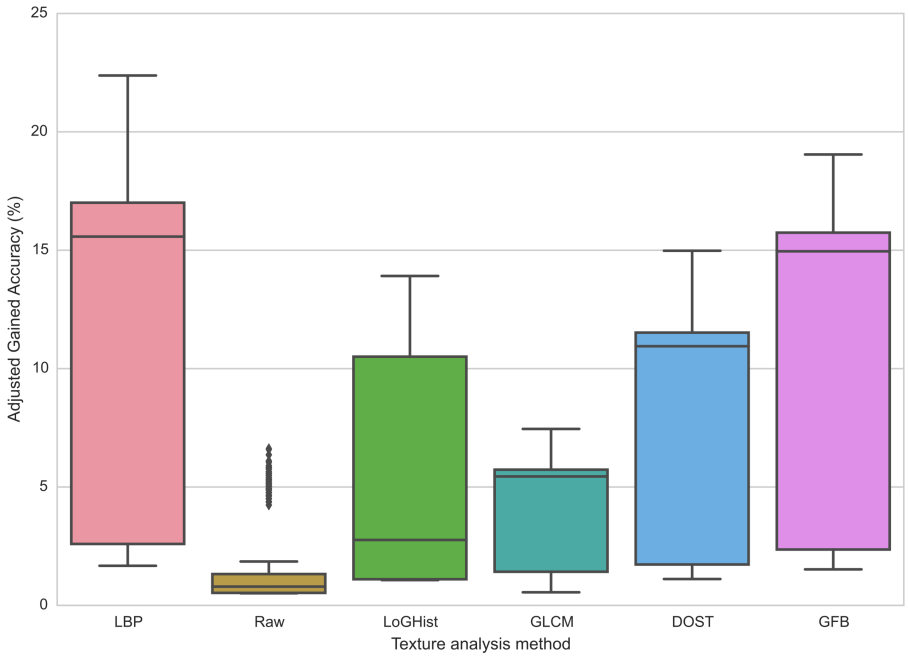


Figure 6.4: Comparison of predictive value among texture analysis methods. As detailed in the materials and methods, numerous models were created in an iterative fashion using leave one out cross validation. Boxplots were generated using the statistics of all models to summarize relative contributions to accuracy of the various texture analyses. Individual outliers are plotted with the accuracy gained by the raw features.

6.6 Discussion

The potential for CT-based texture analysis to determine HPV status of squamous cell carcinoma both within and outside of the oropharynx has been recently demonstrated [172, 187]. In applying a Student t test for independent samples, the authors assessed mean differences in histogram, GLCM, gray level run length, gray level gradient matrix, and Law features. For OPSCC, a limited number of statistically significant differences were identified in histogram parameters (median and entropy) and GLCM entropy [172]. To show potential future clinical applicability, the current study applied a more sophisticated battery of texture analyses to OPSCC and synthesized these results using a machine-learning model in lieu of simply attempting to show statistical differences for individual features. The best classification model demonstrated a promising result of 80% area under the receiver

operating characteristic curve. Based on the inaccurate results from the neuroradiologist review and the lack of interobserver agreement, the difference between HPV-positive and HPV-negative tumors is not obvious to the trained eye of an experienced neuroradiologist when solely evaluating the intrinsic appearance of the primary tumor. Hence, although the machine-learning classification accuracy was not perfect, it underscores that texture analysis has the potential to complement the interpretative skills of a neuroradiologist for determining HPV status of OPSCC. In addition, these results show promise for the generalizability of CT-based texture analysis for OPSCC since CT scans were not acquired with a standard imaging protocol at a single institution, in contrast to previous studies [172, 187]. This is reassuring in light of results showing the impact that CT parameter variation can have on texture features [188, 189].

It is well recognized that HPV-related OPSCC is a molecularly and morphologically distinct tumor type, with the majority demonstrating a poorly differentiated, basaloid, and non-keratinizing histologic appearance [174]. The presence of cystic lymph node metastases from OPSCC is strongly correlated with HPV positivity [175, 176] However, the imaging appearance of the primary tumor has historically been considered a less reliable predictor. Cantrell and colleagues [175] found that HPV-negative tumors were significantly more likely to invade adjacent muscle, while increased enhancement, exophytic morphology, and well defined borders were found more frequently in HPV-positive cases, but these observed differences did not all reach statistical significance at the P less than 0.05 level. The low accuracy from the neuroradiologist review in the current study further supports that it is challenging to predict HPV status of OPSCC based solely on the appearance of the primary tumor (i.e. without taking tumor margins and cystic nodal metastases into account). However, there is evidence that HPV-positive and -negative tumors have differences that can be objectively quantified with imaging.

In using immunohistochemistry to identify p16 overexpression as a surrogate marker for HPV infection, mean and minimum apparent diffusion coefficients were significantly

lower in p16-positive OPSCC [190]. By applying cut-off values of 1.027 x 10-3 mm_2/s and 0.641 x 10-3 mm_2/s for mean and minimum apparent diffusion coefficients respectively, p16 status was predicted with 80.77% accuracy in a small retrospective cohort. These results are presumably related to differences in tumor cellularity and water content and likely reflect underlying histology. The p16-negative cohort was comprised of 64.3% keratinizing tumors, while the p16-positive group was 66.7% non-keratinizing. However, no subgroup analysis was reported to determine if diffusion characteristics correlated with histologic type, independent of p16 expression.

Previous studies have shown the potential for texture analysis to extract other clinically relevant features from imaging data in patients with OPSCC. Zhang and colleagues [191] performed CT histogram and texture measurements in a cohort of 72 patients with locally advanced head and neck squamous cell carcinoma, of which 40% were OPSCC. Primary mass entropy (pixel complexity or irregularity in space) and skewness of the pixel intensity distribution were independently associated with overall survival in patients treated with induction chemotherapy. Unfortunately, HPV status was not available for most of their patients, histologic data was not presented, and their Cox proportional hazards models were not adjusted for primary mass location. As a result, it is not possible to delineate whether the impact of these texture and histogram features on improved survival could be partially or completely attributed to the better prognosis of HPV-related OPSCC, which is known to be more responsive to chemoradiation, have lower recurrence rates, and demonstrate improved disease-specific survival [166, 167]. Likewise, an additional study used texture analysis to predict p53 expression with 81.3% accuracy in a small cohort of patients with squamous cell carcinoma of the oropharynx and hypopharynx, but without specific mention of p16 or HPV status [192] Because p53 expression is not tightly coupled to HPV expression, the results are of limited value in comparing with the current study [193].

This study is limited by its retrospective nature and, based upon patient demographics during the study period, the HPV-negative cases were underrepresented as the minority class.

Over-sampling and under-sampling are 2 commonly used approaches in machine learning for dealing with class imbalance [194]. We avoided over-sampling since it can lead to excessive copies of the minority samples and result in over-fitting. Under-sampling, on the other hand, can discard potentially useful majority samples. To avoid this drawback, we adopted an ensemble under-sampling strategy that has shown success in fully utilizing samples of both classes [195–197].

Future prospective validation is necessary, and this would benefit from a much larger sample size to determine whether texture analysis is merely detecting keratinizing versus non-keratinizing histologic differences or some other unique feature related to HPV infection. Lastly, this study was not intended to compare texture analysis to a neuroradiologist's clinical interpretive skills, as the latter synthesizes numerous additional imaging features that were deliberately hidden from view. Instead, the study was structured solely to determine whether the CT appearance of an OPSCC primary tumor contains imaging data that has not yet been recognized to be of qualitative value in predicting HPV status.

In conclusion, CT texture analysis and machine learning classified the HPV status of OPSCC with a higher accuracy than neuroradiologist review when limited to the internal appearance of the primary tumor. This suggests that useful diagnostic information can be extracted from the CT appearance of OPSCC above and beyond what is apparent to the trained human eye. Consequently, this technique may have potential to complement image interpretation by using a CT-based radiomic analysis to predict HPV-status in patients with OPSCC, though future prospective validation is necessary.

Chapter 7

VALIDATION STUDY IV: BRAIN MR RADIOMICS TO DIFFERENTIATE COGNITIVE DISORDERS

7.1 Summary

This chapter reports the result of a retrospective study to investigate the utility of MRI imaging features in capturing the subtle effects of cognitive decline on the brain. Initial results of the work was presented at Scientific Computing with Python (SciPy) conference, July 2015, Austin, TX. The manuscript is pending submission. My contributions to this manuscript are 1) devising technical design of the study, 2) data selection, data gathering, and data cleaning, 3) implementing the methods section including the image analysis process as well as the machine learning and statistical analysis processes, 4) writing and editing the manuscript.

Authors Sara Ranjbar[1*], Stefanie N. Velgos[2*], Amylou C. Dueck [3], Yonas E. Geda[4,5], J Ross Mitchell[6], and [†]the Alzheimer's Disease Neuroimaging Initiative (*authors contributed equally).

Affiliations [1]Department of Biomedical Informatics, Arizona State University, Tempe, AZ, [2]Center for Clinical and Translational Science, Mayo Clinic Graduate School of Biomedical Sciences, Mayo Clinic, Scottsdale, Arizona, [3]Department of Biostatistics, Mayo Clinic Arizona, Scottsdale, AZ,[4]Department of Psychiatry and Psychology, Mayo Clinic Arizona, Scottsdale, AZ, [5]Department of Neurology, Mayo Clinic Arizona, Scottsdale, AZ, [6]Department of Research, Mayo Clinic, Phoenix, AZ.

[†] Data used in preparation of this article were obtained from the Alzheimer's Disease

Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found here [1].

7.2 Abstract

Subtle and gradual changes in brain anatomy may occur years prior to cognitive impairment. Radiomics can help identify these underlying pathophysiological changes via quantitative image analyses. This study explores the utility of hippocampal texture analysis and volumetric features to differentiate between cognitive states (cognitively normal (CN), Mild Cognitive Impairment (MCI), Alzheimer's disease (AD), Clinical Dementia Rating (CDR) scores. 175 cases from the Alzheimer's Disease Neuroimaging Initiative (ADNI) with 3T T1-weighted structural magnetic resonance (MR) images were selected. Texture and volume features from the bilateral hippocampal area were extracted and machine learning techniques were applied to perform binary classification using Diagonal Quadratic Discriminant Analysis (DQDA) in a leave one out cross validation approach. Area under the receiver operating characteristic (ROC) curve (AUC), sensitivity and specificity were used to assess model performance. Texture analysis was better able to differentiate between no impairment and mild impairment (CN-MCI: AUC of 0.86 and CDRs 0-1: AUC of 0.95). Volumetric features achieved a better classification between no impairment and advanced stages of impairment (CN-AD: AUC of 0.86 and CDRs 0-2: AUC of 0.98). Brain MR radiomics may be a promising tool for assessing early cognitive impairment, as many features are sensitive to early AD pathology.

Keywords: Texture analysis; Alzheimer's disease; Mild cognitive impairment; Machine learning; Radiomics

[1] `https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

7.3 Introduction

The identification of where an individual lies on the cognitive continuum can lead to an earlier diagnosis of cognitive impairment. This is important for rapid communication for management decisions and selective pharmacologic treatment options, which may be more responsive to treatments given at this early stage. It is possible that the current therapeutics for Alzheimer's disease (AD) dementia fail because the intervention is occurring too late in the disease progression. Thus, identifying quantifiable biomarkers of disease progression beyond a collection of cognitive symptoms is paramount.

The existence of a grey zone of cognitive impairment has been reinforced by the emerging clinical need of something beyond the binary diagnosis of the presence or absence of AD [198]. In Alzheimer's disease dementia, the hallmark signatures of extracellular deposits of beta amyloid (Aβ) and intraneuronal neurofibrillary tangles (NFTs), along with neuronal injury and synaptic loss may precede clinical manifestations by several decades, placing it on the far end of the cognitive continuum [199, 200]. Mild cognitive impairment (MCI) falls in the intermediate stage and is often, but not always, a transitional phase from cognitive changes of normal aging to those typically found in AD [198, 201]. Patients with MCI constitute a high-risk group because they develop dementia at 10-15% per year compared with the general population at 1-2% [202]. There is no gold standard for which neuropsychological test battery is used for diagnosis of MCI, but it is important that all the main cognitive areas are examined; these typically include executive function, attention, language, memory and visuospatial skills [198, 203].

Neuropsychological tests in the context of AD help stage dementia severity [204]. A popular brief cognitive screening test to detect cognitive impairment and classify patients as having normal cognition, MCI, or dementia is the Mini Mental State Examination (MMSE) [205]. Staging is needed to generate homogenous or comparable patient samples, define endpoints in clinical drug trials, and to guide pharmacologic treatment options that have

been approved for different levels of disease severity [204, 206]. However, the use of MMSE performance in AD clinical trials and diagnostic studies for inclusion and endpoints may have limitations [206]. Several studies have addressed the MMSE's large ceiling and floor effects, sensitivity to practice effects, and general psychometric limitations [207, 208]. In determining the correspondence between the MMSE and Clinical Dementia Rating (CDR) scores, Perneczky et al. [204] found that MMSE lacks both accuracy in the identification of MCI or mild AD dementia, and validity in detecting and discriminating across the various disease stages. Clinical Dementia Rating is considered the gold standard for staging dementia severity, has good inter-rater reliability and criterion validity, and has also been shown to predict neuropathology [209–212]. Thus, Clinical Dementia Rating scores could be a closer approximation to brain pathology and magnetic resonance (MR) imaging and may be sensitive in early cognitive impairment.

Multiple pattern recognition based methods applied to structural MR images have been developed to predict Alzheimer's disease [9, 213, 214]. Several studies found that local hippocampal and total brain volume are significantly reduced in AD and MCI as compared to healthy elderly [215–219]. Hippocampal volume is the most studied structural biomarker of AD and is used in the criteria for AD diagnosis [220]. In addition, prediction of MCI to AD conversion has been correlated with the rate and amount of hippocampal, medial temporal lobe, and total brain atrophy [221–223].

A current challenge is to discriminate patients who are at a very early stage of cognitive impairment. While volumetric macro-structural correlations provide useful information, another approach that incorporates the well-established microscopic pathological changes in AD is texture analysis. Texture analysis focuses on deriving quantitative and reproducible mathematical metrics from images that describe the interrelations of pixel intensities across multiple spatial scales. Texture analysis has been used previously in the context of AD [215, 221, 224, 225]. For example, Sorensen et al. describes hippocampal texture measures that may serve as a prognostic neuroimaging biomarker of early cognitive impairment [215, 225].

Our group has developed a radiomics texture analysis platform that was previously used characterize gene expression and tumor content heterogeneity of brain cancers[17, 56], and pathology of head and neck cancer [1]. Radiomics involves the high-throughput extraction of quantitative features that converts images into minable data. These features can be used to build descriptive and predictive models that may reveal quantitative predictive or prognostic associations between images and clinical outcomes [26, 171, 226]. The overall goal of this study is to classify neuropsychological performances as they relate to subtle brain changes. In general, we hypothesize that changes in neuropsychological functions has a morphological counterpart, detectable via structural MRI. We hypothesize that the application of brain MR radiomics can achieve comparable classification accuracies with volume and texture features. Our objectives are 2-fold: 1) to differentiate between cognitive states (cognitively normal (CN), Mild Cognitive Impairment (MCI), Alzheimer's disease (AD), and 2) to estimate neuropsychological performance using Clinical Dementia Rating (CDR) scores.

7.4 Methods

7.4.1 Image Dataset

We used T1 MR images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database for this study [227]. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [228]. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). ADNI-1 started in October 2004 as a 5-year study, with a cohort of 200 CN, 200 MCI and 400 AD cases. We selected cases from the shared image collection ADNI1: The patients were divided into the CN, MCI,

and AD groups and subjected to 3T imaging at the following time points: baseline, 6, 12, 18 (MCI only), and 24 months.

7.4.2 Cognitive measures

Clinical Dementia Rating (CDR) scores is one of our classification targets as it focuses on decline in cognitively-driven everyday function, rather than objective cognitive deficits [209]. The CDR score is obtained through semi-structured interviews of patients and informants with the goal of evaluating the patient in six domains: memory, orientation, judgment and problem solving, community affairs, home & hobbies, and personal care [209]. An algorithm creates an overall rating of impairment severity: 0 (no impairment), 0.5 (questionable impairment), 1.0 (mild dementia), 2.0 (moderate dementia), or 3.0 (severe dementia). Typically, a score of 0.5 is given to individuals with a diagnosis of MCI [209, 229].

We also used the categorical division of patients into three cognitive group classifications: cognitively normal (CN), MCI and AD. Group specific inclusion criteria are available on ADNI's website under the General Procedures Manual [230].

7.4.3 Patient Cohort

The initial patient selection criteria were as follows: 1) available CDR score associated with the time of image acquisition, and 2) available 3T scanning protocol to ensure maximum resolution for the image analysis.

We found 204 unique patients in ADNI-1 with available 3T MR images. Image data was available for all patients at different time points ranging from baseline to month 24. Since we were interested in predicting static cognition levels (CDR scores, cognitive groups), time point was irrelevant and we focused on the CDR scores. To ensure unique patients across groups, we selected one time point per patient. To maximize group sizes, we started with finding the minority group cases first: patients who were assigned the CDR score of 2 at

any given time. These patients were excluded from all the other groups. The process was repeated for CDR scores 1 and 0.5. All the remaining patients not assigned to the above groups were placed in the CDR 0 group. The group of CDR score 3 was excluded due to small sample size. Next, we proceeded to find the 3T MR scan time points associated with the assigned group labels for patients. The image data acquired at the selected time points were used for analysis. 29 patients were excluded because the time point of the 3T MR image and CDR score acquisition date did not match, which was a requisite in our selection process. 8 patients were excluded due to image unavailability. The final sample size was 175: 67 non-impaired (CDR 0), 50 questionable (CDR 0.5), 39 mild (CDR 1) and 19 moderate (CDR 2) cognitively impaired individuals.

Table 7.1 shows the demographic information and clinical parameters of the final cohort. Note that to receive a diagnosis of MCI or AD, in addition to clinician judgement, intra-individual decline must be obtained with serial cognitive measurements (multiple CDR scores over time), or by a history of change from previously attained levels [198]. Thus, the numbers of patients between cognitive grouping and CDR scores may differ.

Table 7.1: Demographic and Clinical Parameters

	CN (N = 62)	MCI (N = 71)	AD (N = 42)	
Male N (%)	26 (41.9)	44 (61.9)§	17 (40.5)	
Age Mean (SD)	75.2 (4.7)	75.8 (8.5)	76.2 (8.5)	
	CDR 0 (N = 67)	**CDR 0.5 (N = 50)**	**CDR 1 (N = 39)**	**CDR 2 (N = 19)**
Male N (%)	29 (43.3)	27 (54)	22 (56.4)	9 (47.4)
Age Mean (SD)	74.9 (5.3)	75.6 (7.3)	74.3 (8.9)	81.3 (7.6)*

[Key: CDR, clinical dementia rating; AD, Alzheimer's disease; MCI, mild cognitive impairment; CN, cognitively normal].
§indicates significantly higher proportion of males in the MCI group (Pearson chi-square = 4.0704, $p < 0.001$), and *indicates a significantly higher age in the CDR 2 group (student's t test, $p < 0.05$).

7.4.4 Image preprocessing

Each MPRAGE image in ADNI database is linked with related image files which have undergone specific image preprocessing correction steps which are described on the ADNI website [231]. To ascertain the correction of intensity values within a single scan, we normalized the images based on cerebrospinal fluid (CSF) intensity variance (mean and standard deviation) [232]. Texture and volume analysis were performed using the normalized images.

7.4.5 Regions of Interest

The bilateral hippocampal area was extracted from images in two different methods for texture and volume analysis. For texture analysis, we manually performed 2.5D hippocampal segmentation using the MIPAV (Medical Image Processing, Analysis, and Visualization) application version 7.2.0 [233]. 3 sequential coronal image slices with the largest cross-sectional views of the hippocampus tissue were selected. Unlike other views, coronal view had a common pixel spacing across the cohort as all images were 256x256 pixels with 1.02x1.02 mm_2 of pixel spacings. To minimize the inclusion of areas outside the hippocampus in the analysis, we manually marked the left and right hippocampal area in each slice using square 16x16 pixel ROIs (Figure 7.1a). The ROIs were cropped out of each image and set aside for further analysis.

Segmentation of hippocampi volume was performed automatically as part of volumetric feature extraction using the online volBrain framework [234]. volBrain is a framework that automatically segments parenchyma, brain tissues, macrostructure and subcortical structures of provided MR images in a 3D manner. Figure 7.1b shows volBrain report for one image in the cohort.

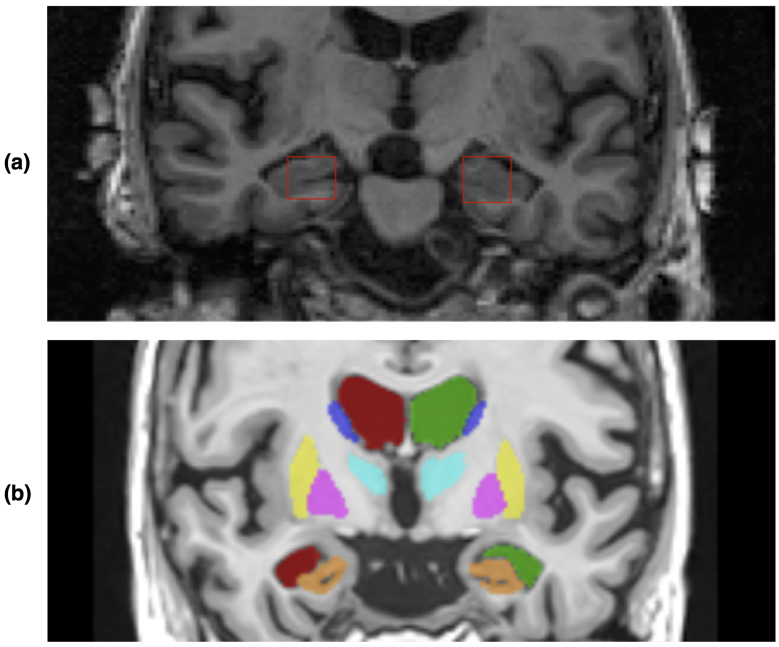


Figure 7.1: Segmentation of the hippocampus in texture and volume analysis (a) Texture analysis regions of interest: The left and right hippocampal areas are manually marked using 16x16 pixel squares (contour in red). This process is repeated on 3 coronal slices with the largest cross-sectional view of the hippocampus area. (b) Volume analysis region of interest: volBrain framework automatically segments subcortical brain tissues. The image shows overlay of segmentation result in volBrain report for one of the patients in our cohort. Hippocampal areas, the region of interest in our analysis, are shown in orange.

7.4.6 Feature Extraction

For each patient, 6 16x16 ROIs (3 image slices x 2 hippocampi) were used as feature extraction inputs. First, to ensure comparability of spectral content across the cohort, the dynamic range of intensities of ROIs were normalized to 0-255. Next, each ROI was analyzed using a texture analysis pipeline with prior success in the analysis of MR images [1, 17, 56]. Developed in Python version 2.7 [99], this in-house-developed pipeline utilizes several available libraries [235, 236] to describe texture of a region of interest statistically and spectrally. Texture descriptors used in this study and their settings are listed in Table 7.2. Overall, we extracted 119 texture features for each ROI. To account for sampling variability,

we averaged the features across each hippocampus, which resulted in a total of 238 texture features (119 per hippocampus) per patient.

Hippocampal volumetric features were generated using VolBrain pipeline [234]. volBrain reported 3 volumetric features for the hippocampus area including absolute volume (cm_3), relative volume (%), and asymmetry index (%). Relative volume represents the sum of hippocampi volumes in relation to the volume of intracranial cavity. The asymmetry index shows the difference between right and left volumes divided by their mean. We used %volume and %asymmetry for our analysis.

Table 7.2: Texture Analysis Methods, Settings, and Number of Features for a Region of Interest

Method	Feature Type	Details	# Features
Raw intensity	Statistical	Mean, range, and standard deviation of the ROI before dynamic range normalization.	3
Gray Level Co-occurrence Matrices (GLCM) [126]	Statistical	Mean and range of 13 GLCM features including entropy, heterogeneity, uniformity, etc. Refer to [126] for complete list of features.	26
Discrete Orthonormal Stockwell Transform (DOST) [91]	Spectral	Rotationally invariant dost features for dyadic frequency bands inside the ROI.	10
Laplacian of Gaussian Histogram (LoGHist) [94]	Spectral	Mean, standard deviation, kurtosis, skewness, entropy and uniformity of the histogram constructed from convolutions of the ROI with band pass filters with sigmas: [0.7, 0.8, 1.2, 1.7, 4.0, 6.0].	36
Local Binary Patterns (LBP) [127]	Statistical	Normalized counts of a histogram generated from aggregated intensity patterns found within patches with 3px radius inside the ROI.	36
Gabor Filter Banks (GFB) [93]	Spectral	Mean and standard deviation of Filtered ROI. Filters were Gabor filters with sigmas: [0.7, 0.8, 1.0, 1.2, 1.7, 2.0, 2.4, 3.0], frequencies: [1.0, 3.0].	32

7.4.7 Statistical analysis and machine learning

Open-source statistical R software (`http://www.R-project.org`) and Scipy 0.18.1 package [101] in Python version 2.7 were used to perform the analyses. The area under the receiver operating characteristic curve (AUC-ROC) was used to assess classification performance. Estimates of P-value (0.05) and 95% confidence interval using DeLong et al. method [237] was used to estimate ROC curve significance. Open-source statistical R software and pRoc package [72] were used for this analysis. Statistical differences between groups were tested using Student's t test for age and Pearson's chi-square test for sex. Statistical significance level was defined as $p<0.05$. We performed univariate analysis to compare the difference in texture and volume feature values for both CDR groups and cognitive groups. The p-values of the two-sample t tests were adjusted for multiple comparisons using the Benjamini and Hochberg False Discovery Rate (FDR) correction method [186].

We used Principle Component Analysis (PCA) [238] to reduce data dimensionality of texture features. To ensure interpretability of the principle components, PCA was applied to features stemming from a common texture analysis method. Several comparative datasets were generated with PCA to find the optimal level of variance. Final set of PCs represented 90% of the variance in the original features, the optimal level of variance in the PCs that resulted in the highest accuracy while avoiding over-fitting. Texture PCs combined with volume features were used in supervised classification experiments using 2 label variables: (1) cognitive groups (CN, MCI, AD); (2) CDR scores. Machine learning was conducted utilizing the open-source python-based package scikit-learn library [102] and custom-written script. Features were selected using Sequential Forward Feature Selection (SFFS) [238] method. Starting from an empty set, SFFS sequentially adds features to the set if they lead to improved accuracy beyond a pre-defined threshold. We used 5% accuracy improvement as the threshold for adding features to the model. We used Diagonal Quadratic Discriminant Analysis (DQDA) [238] to perform supervised classification using a leave one out cross validation (LOOCV) approach. LOOCV uses all samples except for one as the training set

while the left-out sample serves as the test case. This process is repeated so that every sample in the dataset serves as the test sample. The setting used for quadratic discriminant analysis was no priors on classes and threshold of 0.0001 for rank estimation.

We encountered class-imbalance in classification experiments that involved CDR 2. Imbalance in class size is a well-known issue in predictive modeling in which dominance of the majority class hinders the classifier's ability to learn the inherent properties of each class. To ensure generalizability of the result, we used an ensemble down-sampling approach coupled with leave one out cross validation in these experiments. In each cross-validation round, train sample was divided into majority and minority groups. The majority group was then randomly divided into subsets roughly the same size as the minority group. Each of the subsets were merged with the minority group and used to build a classification model. Average of the probabilities of the models was used as the result of the cross validation round.

7.5 Results

The MCI group had a higher proportion of males than the CN and AD groups (Pearson chi-square = 4.0704, $p < 0.001$). No significant difference was observed in gender ratio of other groups. As expected, the age of patients in the CDR 2 group was significantly higher than that of patients at other CDR levels. Figure 7.2 compares volume features across groups and CDR scores. Figure 7.3 shows the univariate comparison of features across feature groups. Features extracted from left and right hippocampi showed similar significance levels. Increasing the level of variance included in the principle components of texture feature did not improve the results.

7.5.1 Classification of Cognitive Groups

Figure 7.4A compares Area under the ROC curves (AUCs) of this section. Classification of Cognitive Groups reached AUC levels of 0.89 (CI: 0.82-0.94) for CN-AD, 0.86 (CI: 0.79-0.91)

for CN-MCI, and 0.70 (CI: 0.61-0.77) for MCI-AD classification experiments, respectively. Table 7.3 shows the performance measures, selected features, and ROC curve analysis of this section. All models were significant at $p<0.05$. Texture feature used in MCI-AD and CN-MCI was the first principle component of Laplacian of Gaussian Histogram (LoGHist) extracted from the left hippocampal region. %Volume was predictive of cognition only in CN-AD models.

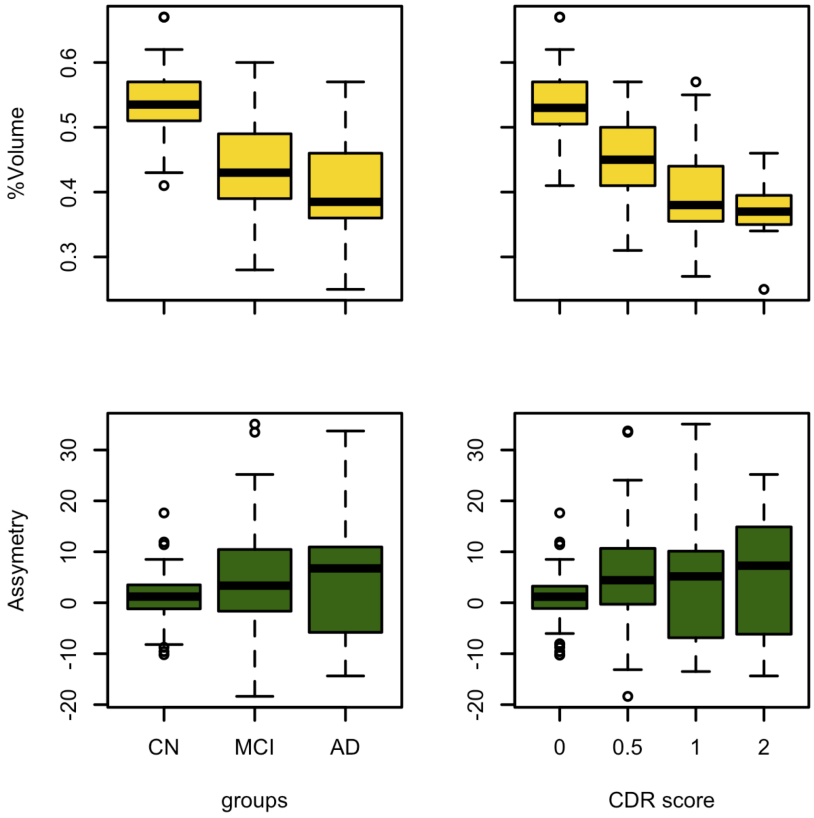


Figure 7.2: Comparison of boxplots for the volumetric features across cognitive groups and CDR scores. The plot shows the distribution of the two volume features (y-axis) across different grouping of patients: cognitive states and CDR scores (x-axis). [%Volume shows the sum of hippocampal volumes in relation to the volume of intracranial cavity. The asymmetry index shows the difference between right and left hippocampal volumes divided by their mean.]

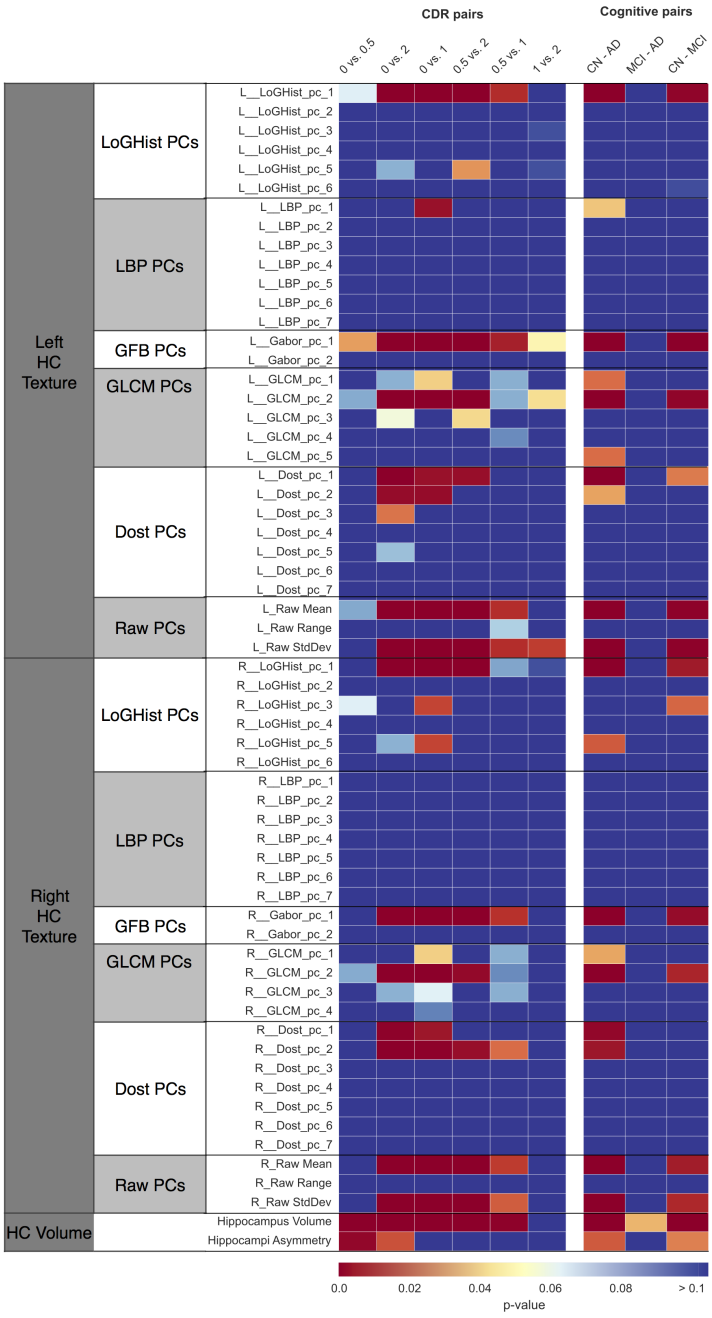


Figure 7.3: Heatmap of univariate analysis of features across groups. Dependent variables are listed above columns (CDR score and cognitive group). Color maps show the False Discovery Rate (FDR) corrected p-values of a two-sample t test. Red to white colors indicates significance (p). [HC, Hippocampus; Dost, Discrete orthonormal Stockwell transform; GFB, Gabor Filter Banks; GLCM, Gray level Co-occurrence Matrices; LBP, Local Binary Patterns; LoGHist, Laplacian of Gaussian Histograms].

Table 7.3: Classification Results for Prediction of Cognitive Groups

	AUC	CI[b]	SE[c]	Z-stats	P-value	Model Features
CN-MCI	0.86	(0.79, 91)	0.03	11.58	<0.0001	LoGHist and LBP PCs of left HC
MCI-AD	0.70	(0.61,0.77)	0.05	4.16	<0.0001	LBP PC of left HC
CN-AD	0.89	(0.82, 0.94)	0.03	12.31	<0.0001	%Volume

[Key: AUC, Area Under the ROC Curve; PC, Principal Component; HC, Hippocampus; LoGHist, Laplacian of Gaussian Histogram; LBP, Local Binary Patterns].
[b]Binomial exact 95% Confidence Interval for the AUC.
[c]Standard Error of the AUC (Delong et al. [237]).

7.5.2 Classification of Clinical Dementia Rating Scores

Figure 7.4B compares Area under the ROC curves (AUCs) of this section. Overall, models were more successful in classification when the target groups were further apart on the CDR spectrum. The AUC levels of the models built in this section respectively were: 0.98 (CI: 0.93-0.99) for CDRs 0-2, 0.95(CI: 0.9-0.98) for CDRs 0-1, 0.84 (CI: 0.76-0.89) for CDRs 0-0.5, 0.73 (CI:0.61-0.83) for CDRs 0.5-2, 0.71 (CI: 0.61-0.8) for CDRs 0.5-1, and 0.56 (CI: 0.42-0.69) for CDRs 1-2. Table 7.4 presents details of models performance and statistical analysis for this section. All models were significant except for classification model of CDR 1-2. The most predictive texture features in this classification experiment were principal components from Discrete Orthonormal Stockwell Transform (DOST) and Gabor Filter Banks (GFB). Similar to the previous experiment, %volume was a predictive volumetric feature. We conducted further analysis to assess whether age accounted for relative volume significantly differentiating between CDR 0 versus 2 in Table 7.4. When age was included in the model, relative hippocampi volume(%) remained highly statistically significant (p=0.003) while age was not significant (p=0.35). The AUC only slightly increased from 0.98 (model with relative volume alone) to 0.9910 (model with relative volume and age). Note, a model containing age by itself resulted in an AUC of only 0.785, and the addition of relative volume

significantly improved the model fit ($p<0.0001$). Thus, relative volume is meaningful in differentiating between CDR 0 and 2 above and beyond age.

Table 7.4: Classification Result for Prediction of Clinical Dementia Rating (CDR) Score

	AUC	CI[b]	SE[c]	Z-stats	P-value	Model Features
CDRs 0,0.5	0.84	(0.76, 89)	0.04	9.67	<0.0001	%Volume
CDRs 0.5,1	0.71	(0.61,0.8)	0.05	4.03	<0.0001	DOST PC of right HC
CDRs 1,2	0.56	(0.42, 0.69)	0.08	0.74	0.46	GFB PC of left HC
CDRs 0,1	0.95	(0.9,0.98)	0.02	22.88	<0.0001	PCs from DOST, LoGHist, and GFB of left HC, GLCM of right HC
CDRs 0.5,2	0.73	(0.61,0.83)	0.08	2.89	0.0038	PCs from DOST and GFB of left HC
CDRs 0,2	0.98	(0.93,0.99)	0.01	46.5	<0.0001	%Volume

[Key: AUC, area under the ROC curve; HC, hippocampus; PC, principle component;DOST, Discrete orthonormal Stockwell transform; GFB, Gabor Filter Banks; GLCM, Gray-Level Co-occurence Matrices].

[b]Binomial exact 95% Confidence Interval for the AUC.

[c]Standard Error of the AUC (Delong et al. [237]).

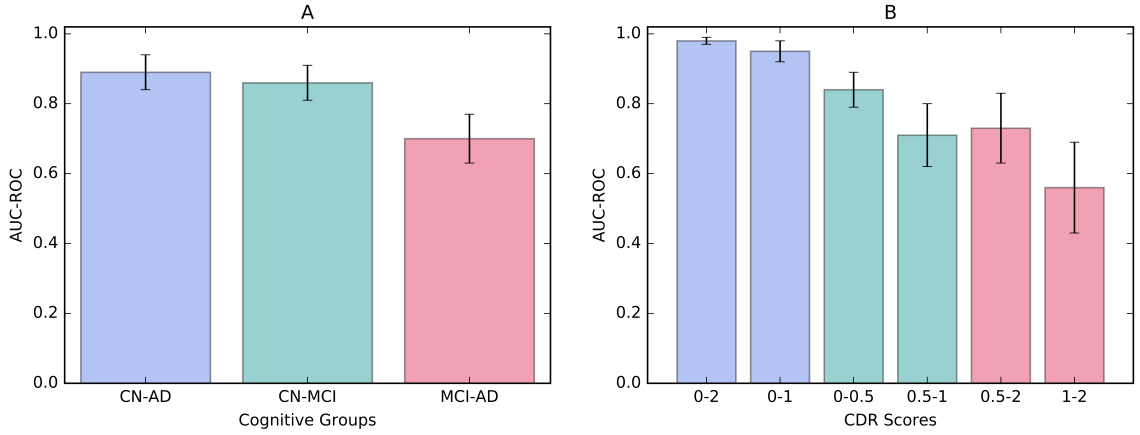


Figure 7.4: Comparison of Area Under ROC Curves. A) Area under the ROC curves of cognitive group classification models, B) Area under the receiver operator curve of CDR score classification models. Error bars show the confidence interval of AUCs. [Key: CN, control; MCI, mild cognitive impairment; AD, Alzheimer's; CDR, clinical dementia rating; ROC, receiver operating characteristic curve; AUC, area under curve].

7.6 Discussion

Despite innate difficulties in identifying MCI due to its heterogeneity, we found that the well-established volume features and texture features had comparable classification accuracies. In distinguishing between CDR scores, we found that texture features had a high classification accuracy (AUC: 0.95) between no cognitive impairment (CDR 0) and mild impairment (CDR 1). Volume had a high classification accuracy (AUC: 0.98) when distinguishing between no impairment (CDR 0) and moderate impairment (CDR 2). It is noteworthy that texture and volume had close accuracies. Volume features performed well in extremes (CN-AD; CDR 0-2), and texture features generally had high accuracies earlier in cognitive impairment (CN-MCI; CDR 0-1).

The experiment to distinguish between CDR 1 and 2 was inconclusive as AUC was poor, not achieving statistical significance (P=0.46). The transition from mild to moderate impairment appears to be a subtle shift without pronounced, discernable steps. While texture features suggest that CDR scores and neuropathology may have a relationship early in cognitive impairment, i.e. early deposition of amyloid or tau, the lack of discrimination accuracy between CDR 1 and 2 suggests the pathological depositions may not help in improving classification accuracy. Aisen et al. [239] posit that the terminology behind mild and moderate AD is inaccurate, because the individual has had the disease present for many years. The clinical staging nomenclature infers a clear distinction between various stages, but in reality, the process progresses in a more continuous manner [239].

There are debates in the literature about what exactly texture captures. In one study, Sorensen et al. suggested that texture captures patterns related to function of hippocampus since it seems to have significant correlation with FDG-PET uptake [215]. The same group also found that hippocampal texture was the most important feature in the algorithm followed by hippocampal volume in discriminating cognitive groups [225]. In our study, texture features contributed to both CN-MCI and MCI-AD models for classification of

sequential stages of cognitive decline. The most comparable study to ours is by Sorensen et al [215]. Our findings are in par with their result. However One technical difference between our methods and Sorensen's is in preprocessing of MR images. They resampled MR images to have consistency in voxel size across their cohort, a necessary preprocessing step when images are obtained from different institutes. However, resampling involves interpolation, which changes the high spatial frequency content (fine level texture) of the image. To establish a reliable baseline for utility of texture features, we focused our study to images with a common voxel size. In the future, we aim to validate this result on external datasets.

Due to technical limitations of our pipeline, we did not perform 3D texture analysis on our data. Instead, we used manual placement of fixed-sized 2-dimensional ROIs (16x16 pixels) on 3 slices with the largest cross sectional view of the hippocampus, which may include immediate anatomical structures including the entorhinal cortex. Additionally, the 3T imaging criteria within ADNI-1 dataset was a limitation which imposed a constraint on the sample size of our study (N = 175). However, we felt it was justified given that 3T images tend to have higher spatial resolution and contrast-to-noise ratios than 1.5T images. As a result, we did not partition the dataset into independent train and test sets but instead, used leave one out cross validation for evaluation of model performance. Therefore, we cannot claim the clinical utility of textural biomarkers introduced in this study since the models were not tested prospectively.

7.7 Conclusion

We utilized existing resources (ADNI-1) to introduce a new application of brain MR radiomics using texture analysis and volumetric features in the field of aging, neuropsychiatry, and dementia. Our study findings support the use of brain MR radiomics for assessing early cognitive impairment, as many features are sensitive to early AD pathology. With time and finesse, brain MR radiomics may have the potential for individualized prediction of cognitive impairment or conversion to AD, and may identify individuals at the earliest

stage of cognitive impairment. Early identification of cases has implication in identifying a target population for clinical intervention to delay cognitive deterioration. However, future studies need to replicate our findings and should examine the clinical utility of image texture features as AD biomarkers.

7.8 Acknowledgments

This publication was made possible by CTSA Grant Number UL1 TR000135 from the National Center for Advancing Translational Sciences (NCATS), a component of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; MesoScale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University

of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Chapter 8

DISCUSSION

This thesis proposed a tool for deriving quantitative imaging features from diagnostic imaging. For AIM I, I presented an analytical informatics-driven pipeline that can contribute to the advancement of biomedical imaging informatics in a variety of ways. Data provided with this pipeline can be used to enhance context-driven image interpretations to improve diagnostic accuracies. They can also be utilized in the development of advanced applications for computer-aided diagnosis to improve performance compared with the current clinical standard of care.

The processes of biomarker detection and measurement need to be accurate, reproducible and feasible over time [240–242]. Imaging biomarkers are no exception in this regard. In order to produce reproducible and reliable imaging biomarkers standardized feature extraction processes need to be developed. Among the criticisms to traditional quantitative image analysis studies are that they are fragmented studies that are difficult to repeat and are rarely validated. These studies are sometimes referred to "feature engineering" experiments, in which models are built on too specific data using overly selective feature extraction and analysis method. Radiomics approach addresses at least one of these criticisms by reducing the effect of feature extraction method. Using a variety of feature extraction methods, radiomics remains open to all possible options including widening the range of parameter settings, leaving the decision on the value of each feature to the machine learning. By facilitating close-to-optimal extraction of information from diagnostic imaging, radiomics serves as a bench mark for faster standardization of feature extraction and analysis processes. This approach also expedites standardization and technical validation of potential imaging biomarkers. Several recent standardization studies support this argument in which the radiomics approach was used to assess robustness and reproducibility of imaging features

[243], to determine the significance of inter-scanner variability on imaging features [244], to find the optimal image intensity quantization method [245], and to compare across feature selection and classification methods [105]. This pipeline currently provides a number of highly-cited texture analysis methods. Inclusion of additional texture analysis algorithms and other clinically relevant quantifiers such as descriptors of lesion boundary/margin can be potential future directions for the this pipeline.

In this pipeline, we focused on the analysis of small regions (ROIs) within larger suspicious regions identified by experts. Focusing the analysis to a smaller-sized region limits spatial support encroached by the texture operators and avoids mixing properties from distinct texture classes [246]. However, in general having common-sized ROI for all patients is not necessarily a requirement for texture analysis. This requirement in our pipeline is a restriction imposed by the discrete orthonormal Stockwell transform, in which different sizes of ROI leads to variance in the number and content of features. The default settings of parameters for the texture analysis operators included here are also the result of our limited experiments with small biomedical texture. Optimal settings of parameters for feature extraction and optimal size of the ROIs should be addressed with rigorous examination in future studies. If available, test-retest data are also extremely helpful, as they can help prioritize features on the basis of their reproducibility and reliability.

In my humble opinion, parameter tuning is also a limitation imposed on classic quantitative imaging approaches. Traditionally quantitative imaging relies on classic machine learning approaches that require previously-crafted features for learning. In recent years deep learning has shown excelled performance in segmentation [247–250], detection [251], diagnosis [252], and classification [253] of tissue structures. The appeal of deep learning lies its abstract data representations and elimination of the need for hand-crafted features. Although deep learning requires large amount of training data with attached ground truth. Moreover, the collective data produced overtime for rare diseases probably will never suffice for the effective use of deep learning. Scarcity of large amount of labeled data is an obstacle that is expected

to lead to slower adoption of deep learning approach in clinical practice. Transfer learning [254–256], in which deep trained networks trained on very large datasets (eg. ImageNet [257]) are trained with biomedical images for a different task are often suggested as an intermediate solution to expedite the process.

We used classic classification approaches such as support vector machines in the experiments presented here. Overly specific feature selection and complicated classification methods were avoided to reduce the potential for overfitted models, specially since we had small datasets and no separate validation or test sets. In general the success of traditional machine learning approach in finding meaningful features heavily depends on the data available and the parameter setting used for machine learning. I acknowledge that the models presented here may have potential for improvement as a result of our conservative classification approaches. These models should be considered as "proof of concept" or "pilot" models and should be examined for the effect of parameter setting as well as the choice of feature selection and classification method in the future.

The current pipeline has several limitations. Firstly, we performed image analysis on 2-dimensional (2D) images in Chapters 4-6. Two-dimensional image analysis is straightforward to implement, yet does not allow for optimal or comprehensive image analysis, particularly for volumetric data acquired using magnetic resonance imaging (MRI) machines or computed tomography (CT) scanners. In the Alzheimer's study (Chapter 7) I attempted to address this shortcoming by performing image analysis on three slices instead of one to allow for 2.5D analysis of the regions of interest and improved sampling on image data. However, 2.5D approach has also not been without disadvantages since it results in the lose of spatial information related to distinct local habitats. To become fully comprehensive, this pipeline needs to allow for 3D image analysis. Hence, future directions for this pipeline should include its refinement into a volumetric analysis tool.

The second limitation of this pipeline lies in its segmentation process. Success of a

radiomics workflow highly relies on the result of intermediate steps such as the segmentation process. Segmentation is of utmost importance particularly for extraction of morphological features which heavily depend on the result of segmentation. This pipeline uses manual segmentation (outline drawn by an expert) as the basis for the selection of a region of interest. In breast cancer study (Chapter 5) outlines drawn by radiologists were used for computation of quantitative morphological features. However, as previously discussed in Chapter 2 (and also in the discussion section of Chapter 5), manual segmentation is time-consuming and can suffer from inter- and intra-reader variability, factors that can limit the prospect of this pipeline's use in fully elaborated workflows and the validity and reliability of the features. In the future, this pipeline should incorporate an automatic or a semi-automatic segmentation system to expedite the implementation process and clinical efficiency, and to result in objective morphology features.

Studies conducted for completion of AIM II (Chapters 4-7) assessed the utility of imaging-derived features in diagnostic applications related to head and neck cancers, breast cancer, and Alzheimer's disease. These retrospective studies demonstrated that identifying imaging traits can lead to improved diagnosis accuracy and potentially result in advancement of diagnostic biomarker discovery. Given that the image data used in these studies was acquired from multiple sites (with the exception of the breast cancer study), I am hopeful that repeating these experiments on datasets acquired from other institutes would render good results in the future. In this thesis we focused on evaluation of the utility of imaging-derived features in diagnostic applications In cancer-related studies the performance of predictive models was compared with that of radiologists to investigate the benefits of quantitative image analysis to routine care. In these studies we often compared the performance of models with human readers. It should be noted that the motivation of these comparisons was to assess the potential utility of the pipeline to supplement radiologists' interpretive skills in diagnostic tasks and not to compete with them.

Chapter 5 discussed the potential utility of this technique in breast cancer screening.

The model built on quantitative image features provided reduced number of false positives compared to trained radiologists. The study suggested that a computer aided diagnosis system based on quantitative image features can provide benefit in improving the accuracy of diagnosis by providing a recommendation. This can result in reduced number of unnecessary supplemental imaging and invasive biopsy procedures for the patients leading to reduced cost on the health care system as well as enhancement in patient outcome. This study is under final stages of review by European Journal of radiology.

Chapter 6 and 7 discuss the utility of this pipeline in biomedical discovery. The study reported in chapter 6 (accepted for publication by the Journal of Computed Assisted Tomography) showed that quantitative image features could be used to detect differences in lesion representation of HPV-positive and HPV-negative OPSCC lesions. Provided with the same region of interest, experienced neuroradiologists did not observe any detectable visual difference between the two cohorts. HPV is an important factor in patient prognosis and can only be detected via biopsies. Thus a system built on this approach can potentially enhance the diagnosis process by providing critical information at an early stage. This study has been accepted by the Journal of Computed Tomography in 2017 (publication details to be defined).

Chapter 7 showed texture features can provide information when proved pathological diagnosis does not exists. Definite diagnosis of Alzheimer's is delayed until the disease is fully advanced and can only be pathologically proven post-mortem. By detecting earlier signs of cognitive decline, model based on image analysis can lead to earlier interventions in the treatment process and improved patient management.

Imaging features can also be used along with other types of biomedical data such as genomics and clinical data to generate multi-scale models to improve patient outcome and survival. Bridging imaging features with other sources of clinical data can provide a path for advancement of individualized medicine [258]. It should be emphasized that these types of

analyses identify correlations, not causes [31]. However, these correlations can inform the decision about whether to test for certain gene alterations in biopsy samples as well as the choice of biopsy sites [31].

Apart from the studies presented in Chapters 4-7, we also have conducted two studies using this pipeline that support the radiogenomics argument. In 2015 our group conducted a study in which imaging features extracted from MRI images were used to predict the tumor content of biopsied locations of Glioblastoma Multiforme (GBM) [17]. The model was used to generate maps of predicted tumor content across the entire lesions. These prediction maps showed promising result when tested against a previously unseen validation set, suggesting the utility of a radiomics models in guiding biopsy procedures. In a followup study in 2016 image features were identified as non-invasive imaging signatures for the genomics pathways of GBM lesions. We generated prediction maps of genomics pathways of GBM across the entirety of lesions using models built on imaging features. The models showed promising result on unseen validation data, suggesting a potential for characterization of regional genetic heterogeneity in glioblastoma.

Although promising, the results presented here should not be interpreted as the performance of an established diagnostic imaging test. Radiomics is a young discipline that will need to undergo a slow process of standardization and characterization before it reaches the level of reliability and reproducibility required for clinical implemented diagnostic and therapeutic systems [240–242]. The obstacles along the way can be associated with a number of causes, including technical complexities of identifying reliable and reproducible image features. These features need to be portraying disease manifestation across the population and not be the result of unrecognized confounding variable such as image acquisition. To address this concern, radiomics investigations should be validated against external completely unseen datasets preferably from other institutes. These studies should be conducted prospectively and directed to assess imaging-pathology, imaging-molecular or imaging-genomics

correlations [259]. Patient risk factor data (such as subject age, biology) also need to be incorporated into radiomics as these factors can influence the image features [260].

Large standardized datasets representing the full spectrum of patients are required to result in unbiased inferences of radiomics analyses. Initiatives such as Quantitative Imaging Biomarkers Alliance (QIBA), sponsored by Radiological Society of North America (RSNA) and the National Institute for Biomedical Imaging and Bioengineering [261], are among promising steps that encourage collaboration among the stakeholders and advance the field through provision of a consensus on imaging protocols, and the measurement accuracy of imaging biomarkers. The Cancer Imaging Archive (TCIA) [262], Quantitative Imaging Data Warehouse (QIDW) [263], Alzheimer's Disease Neuroimaging Initiative (ADNI) [264] are examples of targeted data collections that aim to generate a consensus among the stakeholders on how to optimize, harmonize, and standardize data collection and analysis processes.

Chapter 9

CONCLUSION

Despite major advances in computational biology, gaps still exist in our understanding of cancer dynamics. Radiomics offer the potential for a non-invasive tool to complement existing methods of assessing disease characteristics and behavior to enhance diagnosis, monitoring, and prediction of response to treatment for various cancers. The field of radiomics is still young and requires rigorous technical and clinical validation studies. As the field grows, we can expect radiomics to bridge the gaps of knowledge and help shift medicine closer to a fundamental genetic-based and truly personalized practice. With expanding radiomics cohorts and feature dimensions higher prediction performance in future radiomics studies is expected.

This thesis aimed to develop a radiomics pipeline for biomedical image analysis, and to examine its utility in diagnostic applications. The data provided with this pipeline showed potential in the following tasks: a) to provide incremental benefit to the neuroradiologists in differentiating small head and neck lesions (chapter 4), b) to reduce the number of false positives in breast cancer screening (chapter 5), c) to provide insight about the presence of HPV in a malignant type of head and neck cancer (chapter 6) and finally, d) to detect early signs of cognitive decline in patients with Alzheimer's disease. These studies suggest that converting images to quantitative data can enable improved accuracy in the diagnosis of head and neck cancers, breast cancer, and Alzheimer's disease. Enhancing diagnostic accuracy results in improved patient care and patient management as well as earlier and better-informed decisions in treatment planning. Overall, this thesis is a step forward towards the enhancements of radiomics based clinical predictions.

The Future path for this pipeline can include conducting integrative studies that contribute

to further advancement of imaging biomarker research. Examples are standardization studies that assess reproducibility and reliability of imaging features over image acquisition setting, feature extraction, and feature analysis method or validation studies that assess imaging-pathology, imaging-molecular or imaging-genomics correlations.

NOTES

REFERENCES

[1] S. Ramkumar et al. "MRI-Based Texture Analysis to Differentiate Sinonasal Squamous Cell Carcinoma from Inverted Papilloma". In: *American Journal of Neuroradiology* 38.5 (2017), pp. 1019–1025. ISSN: 0195-6108. DOI: 10.3174/ajnr.A5106. URL: http://www.ajnr.org/content/38/5/1019.

[2] Sara Ranjbar and J. Ross Mitchell. "Chapter 8 - An Introduction to Radiomics: An Evolving Cornerstone of Precision Medicine". In: *Biomedical Texture Analysis*. Ed. by Adrien Depeursinge, Omar S. Al-Kadi, and J.Ross Mitchell. Academic Press, 2017, pp. 223–245. ISBN: 978-0-12-812133-7. URL: http://www.sciencedirect.com/science/article/pii/B9780128121337000089.

[3] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. "Cancer statistics, 2017". In: *CA: A Cancer Journal for Clinicians* 67.1 (Feb. 2017), pp. 7–30.

[4] Angela B Mariotto, K Robin Yabroff, Yongwu Shao, Eric J Feuer, and Martin L Brown. "Projections of the cost of cancer care in the United States: 2010–2020". In: *Journal of the National Cancer Institute* 103.2 (2011), pp. 117–128.

[5] Mia A Levy and Daniel L Rubin. "Current and future trends in imaging informatics for oncology". In: *Cancer journal (Sudbury, Mass.)* 17.4 (2011), p. 203.

[6] Andrew J Buckler et al. "The use of volumetric CT as an imaging biomarker in lung cancer". In: *Academic radiology* 17.1 (2010), pp. 100–106.

[7] AR Padhani and L Ollivier. "The RECIST criteria: implications for diagnostic radiologists". In: *The British journal of radiology* 74.887 (2001), pp. 983–986.

[8] Charles L Sawyers. "The cancer biomarker problem". In: *Nature* 452.7187 (2008), pp. 548–552.

[9] Catharine M West and Gillian C Barnett. "Genetics and genomics of radiotherapy toxicity: towards prediction". In: *Genome medicine* 3.8 (2011), p. 52.

[10] David E Gerber. "Targeted therapies: a new generation of cancer treatments". In: *Am Fam Physician* 77.3 (2008), pp. 311–319.

[11] Michele E Gaguski and Phyllis Begyn. "Advances in cancer therapy: Targeted agents". In: *Nursing management* 39.5 (2008), pp. 45–46.

[12] Adnan Mujahid Khan and Yinyin Yuan. "Biopsy variability of lymphocytic infiltration in breast cancer subtypes and the ImmunoSkew score". In: *Scientific reports* 6 (2016), p. 36231.

[13] Philippe Lambin et al. "Radiomics: extracting more information from medical images using advanced feature analysis". In: *European journal of cancer* 48.4 (2012), pp. 441–446.

[14] Shinichi Yachida et al. “Distant metastasis occurs late during the genetic evolution of pancreatic cancer”. In: *Nature* 467.7319 (2010), pp. 1114–1117.

[15] Marco Gerlinger et al. “Intratumor heterogeneity and branched evolution revealed by multiregion sequencing”. In: *N Engl j Med* 2012.366 (2012), pp. 883–892.

[16] Andrea Sottoriva et al. “Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics”. In: *Proceedings of the National Academy of Sciences* 110.10 (2013), pp. 4009–4014.

[17] Leland S Hu et al. “Multi-parametric MRI and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma”. In: *PloS one* 10.11 (2015), e0141506.

[18] Mary Claire Skrzynski, J Sybil Biermann, Anthony Montag, and Michael A Simon. “Diagnostic accuracy and charge-savings of outpatient core needle biopsy compared with open biopsy of musculoskeletal tumors”. In: *J Bone Joint Surg Am* 78.5 (1996), pp. 644–9.

[19] Paul L Nguyen et al. “The impact of pathology review on treatment recommendations for patients with adenocarcinoma of the prostate”. In: *Urologic Oncology: Seminars and Original Investigations*. Vol. 22. 4. Elsevier. 2004, pp. 295–299.

[20] Valerie L Staradub, Kathleen A Messenger, Nanjiang Hao, and Monica Morrow. “Changes in breast cancer therapy because of pathology second opinions”. In: *Annals of surgical oncology* 9.10 (2002), pp. 982–987.

[21] JK Aronson. “Biomarkers and surrogate endpoints”. In: *British journal of clinical pharmacology* 59.5 (2005), pp. 491–494.

[22] JA Wagner. “Biomarkers: principles, policies, and practice”. In: *Clinical Pharmacology & Therapeutics* 86.1 (2009), pp. 3–7.

[23] Darren R Hodgson, Robin D Whittaker, Athula Herath, Dereck Amakye, and Glen Clack. “Biomarkers in oncology drug development”. In: *Molecular oncology* 3.1 (2009), pp. 24–32.

[24] Mitch Dowsett. “Preoperative models to evaluate endocrine strategies for breast cancer”. In: *Clinical cancer research* 9.1 (2003), 502s–510s.

[25] Jeffrey William Prescott. “Quantitative imaging biomarkers: the application of advanced image processing and analysis to clinical and preclinical decision making”. In: *Journal of digital imaging* 26.1 (2013), pp. 97–108.

[26] Virendra Kumar et al. “Radiomics: the process and the challenges”. In: *Magnetic resonance imaging* 30.9 (2012), pp. 1234–1248.

[27] Hugo JWL Aerts et al. “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach”. In: *Nature communications* 5 (2014).

[28] Robert Ford et al. “Lessons learned from independent central review”. In: *European Journal of Cancer* 45.2 (2009), pp. 268–274.

[29] Philippe Thiesse et al. “Response rate accuracy in oncology trials: reasons for interobserver variability. Groupe Français d’Immunothérapie of the Fédération Nationale des Centres de Lutte Contre le Cancer.” In: *Journal of Clinical Oncology* 15.12 (1997), pp. 3507–3514.

[30] Aaron M Rutman and Michael D Kuo. “Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging”. In: *European journal of radiology* 70.2 (2009), pp. 232–241.

[31] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. “Radiomics: images are more than pictures, they are data”. In: *Radiology* 278.2 (2015), pp. 563–577.

[32] S Herlidou-Meme et al. “MRI texture analysis on texture test objects, normal brain and intracranial tumors”. In: *Magnetic resonance imaging* 21.9 (2003), pp. 989–993.

[33] Doaa Mahmoud-Ghoneim, Grégoire Toussaint, Jean-Marc Constans, and D Jacques. “Three dimensional texture analysis in MRI: a preliminary evaluation in gliomas”. In: *Magnetic resonance imaging* 21.9 (2003), pp. 983–987.

[34] RA Lerski et al. “VIII. MR image texture analysis – an approach to tissue characterization”. In: *Magnetic resonance imaging* 11.6 (1993), pp. 873–887.

[35] Evangelia I Zacharaki et al. “Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme”. In: *Magnetic resonance in medicine* 62.6 (2009), pp. 1609–1618.

[36] Naga R Mudigonda, R Rangayyan, and JE Leo Desautels. “Gradient and texture analysis for the classification of mammographic masses”. In: *IEEE transactions on medical imaging* 19.10 (2000), pp. 1032–1043.

[37] Peter Gibbs and Lindsay W Turnbull. “Textural analysis of contrast-enhanced MR images of the breast”. In: *Magnetic resonance in medicine* 50.1 (2003), pp. 92–98.

[38] Yuanjie Zheng et al. “STEP: Spatiotemporal enhancement pattern for MR-based breast tumor diagnosis”. In: *Medical physics* 36.7 (2009), pp. 3192–3204.

[39] Dustin Newell et al. “Selection of diagnostic features on breast MRI to differentiate between malignant and benign lesions using computer-aided diagnosis: differences in lesions presenting as mass and non-mass-like enhancement”. In: *European radiology* 20.4 (2010), pp. 771–781.

[40] Shannon C Agner et al. “Textural kinetics: a novel dynamic contrast-enhanced (DCE)-MRI feature for breast lesion classification”. In: *Journal of Digital Imaging* 24.3 (2011), pp. 446–463.

[41] SA Waugh et al. “Magnetic resonance imaging texture analysis classification of primary breast cancer”. In: *European radiology* 26.2 (2016), pp. 322–330.

[42] Adrien Depeursinge et al. “Near-affine-invariant texture learning for lung tissue analysis using isotropic wavelet frames”. In: *IEEE Transactions on Information Technology in Biomedicine* 16.4 (2012), pp. 665–675.

[43] Omar S Al-Kadi and D Watson. "Texture analysis of aggressive and nonaggressive lung tumor CE CT images". In: *IEEE transactions on biomedical engineering* 55.7 (2008), pp. 1822–1830.

[44] Andreas Wibmer et al. "Haralick texture analysis of prostate MRI: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores". In: *European radiology* 25.10 (2015), pp. 2840–2850.

[45] Sylvia Drabycz et al. "An analysis of image texture, tumor location, and MGMT promoter methylation in glioblastoma using magnetic resonance imaging". In: *Neuroimage* 49.2 (2010), pp. 1398–1405.

[46] Balaji Ganeshan, Sandra Abaleke, RC Young, Christopher R Chatwin, and Kenneth A Miles. "Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage". In: *Cancer imaging* 10.1 (2010), pp. 137–143.

[47] Francesca Ng, Balaji Ganeshan, Robert Kozarski, Kenneth A Miles, and Vicky Goh. "Assessment of primary colorectal cancer heterogeneity by using whole-tumor texture analysis: contrast-enhanced CT texture as a biomarker of 5-year survival". In: *Radiology* 266.1 (2013), pp. 177–184.

[48] Sheng-Xiang Rao et al. "Whole-liver CT texture analysis in colorectal cancer: Does the presence of liver metastases affect the texture of the remaining liver?" In: *United European gastroenterology journal* 2.6 (2014), pp. 530–538.

[49] Kenneth A Miles, Balaji Ganeshan, Matthew R Griffiths, Rupert CD Young, and Christopher R Chatwin. "Colorectal cancer: texture analysis of portal phase hepatic CT images as a potential marker of survival 1". In: *Radiology* 250.2 (2009), pp. 444–452.

[50] Vicky Goh et al. "Assessment of response to tyrosine kinase inhibitors in metastatic renal cell cancer: CT texture as a predictive biomarker". In: *Radiology* 261.1 (2011), pp. 165–171.

[51] Ash A Alizadeh et al. "Toward understanding and exploiting tumor heterogeneity". In: *Nature medicine* 21.8 (2015), pp. 846–853.

[52] Michael D Kuo and Neema Jamshidi. "Behind the numbers: decoding molecular phenotypes with radiogenomics – guiding principles and technical considerations". In: *Radiology* 270.2 (2014), pp. 320–325.

[53] Zhen Guo, Yan Shu, Hui Wang, Honghao Zhou, and Wei Zhang. "Radiogenomics helps to achieve personalized therapy by evaluating patient responses to radiation treatment". In: *Carcinogenesis* (2015), bgv007.

[54] Proud MSN Carol. "Radiogenomics: the promise of personalized treatment in radiation oncology?" In: *Clinical journal of oncology nursing* 18.2 (2014), p. 185.

[55] Hui Yang et al. "Healthcare intelligence: turning data into knowledge". In: *IEEE Intelligent Systems* 29.3 (2014), pp. 54–68.

[56] Leland S Hu et al. “Radiogenomics to characterize regional genetic heterogeneity in glioblastoma”. In: *Neuro-oncology* 19.1 (2016), pp. 128–137.

[57] Chintan Parmar et al. “Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer”. In: *Scientific reports* 5 (2015), p. 11044.

[58] Thibaud P Coroller et al. “CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma”. In: *Radiotherapy and Oncology* 114.3 (2015), pp. 345–350.

[59] Pascal O Zinn et al. “Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme”. In: *PloS one* 6.10 (2011), e25451.

[60] Olivier Gevaert et al. “Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features”. In: *Radiology* 273.1 (2014), pp. 168–174.

[61] Maximilian Diehn et al. “Identification of noninvasive imaging surrogates for brain tumor gene-expression modules”. In: *Proceedings of the National Academy of Sciences* 105.13 (2008), pp. 5213–5218.

[62] Christoph A Karlo et al. “Radiogenomics of clear cell renal cell carcinoma: associations between CT imaging features and mutations”. In: *Radiology* 270.2 (2014), pp. 464–471.

[63] Eran Segal et al. “Decoding global gene expression programs in liver cancer by noninvasive imaging”. In: *Nature biotechnology* 25.6 (2007), pp. 675–680.

[64] Shota Yamamoto, Daniel D Maki, Ronald L Korn, and Michael D Kuo. “Radiogenomic analysis of breast cancer using MRI: a preliminary study to define the landscape”. In: *American Journal of Roentgenology* 199.3 (2012), pp. 654–663.

[65] Rob J van Klaveren et al. “Management of lung nodules detected by volume CT scanning”. In: *New England Journal of Medicine* 361.23 (2009), pp. 2221–2229.

[66] Yi Wang et al. “Diffusion-weighted magnetic resonance imaging of pancreatic adenocarcinomas: Association with histopathology and tumor grade”. In: *Journal of Magnetic Resonance Imaging* 33.1 (2011), pp. 136–142.

[67] J-L Van Laethem et al. “New strategies and designs in pancreatic cancer research: consensus guidelines report from a European expert panel”. In: *Annals of oncology* (2011), mdr351.

[68] Ona Wu et al. “Tracer arrival timing-insensitive technique for estimating flow in MR perfusion-weighted imaging using singular value decomposition with a block-circulant deconvolution matrix”. In: *Magnetic resonance in medicine* 50.1 (2003), pp. 164–174.

[69] Harriet C Thoeny and Brian D Ross. “Predicting and monitoring cancer treatment response with diffusion-weighted MRI”. In: *Journal of Magnetic Resonance Imaging* 32.1 (2010), pp. 2–16.

[70] Thomas Beyer, David W Townsend, Tony Brun, and Paul E Kinahan. “A combined PET/CT scanner for clinical oncology”. In: *The Journal of nuclear medicine* 41.8 (2000), p. 1369.

[71] Wolf-Dieter Heiss. "The potential of PET/MR for brain imaging". In: *European journal of nuclear medicine and molecular imaging* 36.1 (2009), pp. 105–112.

[72] Andrés Larroza et al. "Support vector machine classification of brain metastasis and radiation necrosis based on texture analysis in MRI". In: *Journal of Magnetic Resonance Imaging* 42.5 (2015), pp. 1362–1368.

[73] Abigail L Stockham et al. "Conventional MRI does not reliably distinguish radiation necrosis from tumor recurrence after stereotactic radiosurgery". In: *Journal of neuro-oncology* 109.1 (2012), pp. 149–158.

[74] Samuel T Chao et al. "Challenges with the diagnosis and treatment of cerebral radiation necrosis". In: *International Journal of Radiation Oncology* Biology* Physics* 87.3 (2013), pp. 449–457.

[75] Christine Tanner, Michael Khazen, Preminda Kessar, Martin O Leach, and David J Hawkes. "Classification improvement by segmentation refinement: Application to contrast-enhanced MR-mammography". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2004, pp. 184–191.

[76] Chintan Parmar et al. "Robust radiomics feature quantification using semiautomatic volumetric segmentation". In: *PloS one* 9.7 (2014), e102107.

[77] Woo Kyung Moon et al. "Quantitative ultrasound analysis for classification of BI-RADS category 3 breast masses". In: *Journal of digital imaging* 26.6 (2013), pp. 1091–1098.

[78] Robert M Haralick. "Statistical and structural approaches to texture". In: *Proceedings of the IEEE* 67.5 (1979), pp. 786–804.

[79] Mary M Galloway. "Texture analysis using gray level run lengths". In: *Computer graphics and image processing* 4.2 (1975), pp. 172–179.

[80] Joan S Weszka, Charles R Dyer, and Azriel Rosenfeld. "A comparative study of texture measures for terrain classification". In: *IEEE Transactions on Systems, Man, and Cybernetics* 4 (1976), pp. 269–285.

[81] Richard W Conners and Charles A Harlow. "A theoretical comparison of texture algorithms". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 3 (1980), pp. 204–222.

[82] Aswini Kumar Mohanty, Swapnasikta Beberta, and Saroj Kumar Lenka. "Classifying benign and malignant mass using GLCM and GLRLM based texture features from mammogram". In: *International Journal of Engineering Research and Applications* 1.3 (2011), pp. 687–693.

[83] K Raghesh Krishnan and R Sudhakar. "Automatic classification of liver diseases from ultrasound images using GLRLM texture features". In: *Soft Computing Applications*. Springer, 2013, pp. 611–624.

[84] Hari Wibawanto, Adhi Susanto, Thomas Sri Widodo, and S Maesadji Tjokronegoro. "Discriminating cystic and non cystic mass using GLCM and GLRLM-based texture features". In: *Int J Electron Eng Res* 2 (2010), pp. 569–580.

[85] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. "Gray scale and rotation invariant texture classification with local binary patterns". In: *European Conference on Computer Vision*. Springer. 2000, pp. 404–420.

[86] Bidyut Baran Chaudhuri and Nirupam Sarkar. "Texture segmentation using fractal dimension". In: *IEEE Transactions on pattern analysis and machine intelligence* 17.1 (1995), pp. 72–77.

[87] Mika Seppä and Matti Hämäläinen. "Visualizing human brain surface from T1-weighted MR images using texture-mapped triangle meshes". In: *Neuroimage* 26.1 (2005), pp. 1–12.

[88] Stephane G Mallat. "A theory for multiresolution signal decomposition: the wavelet representation". In: *IEEE transactions on pattern analysis and machine intelligence* 11.7 (1989), pp. 674–693.

[89] Adrien Depeursinge, Antonio Foncubierta-Rodriguez, Dimitri Van de Ville, and Henning Müller. "Lung texture classification using locally–oriented Riesz components". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2011, pp. 231–238.

[90] Robert Glenn Stockwell, Lalu Mansinha, and RP Lowe. "Localization of the complex spectrum: the S transform". In: *IEEE transactions on signal processing* 44.4 (1996), pp. 998–1001.

[91] Sylvia Drabycz, Robert G Stockwell, and J Ross Mitchell. "Image texture characterization using the discrete orthonormal S-transform". In: *Journal of digital imaging* 22.6 (2009), p. 696.

[92] Hongmei Zhu et al. "MR multi-spectral texture analysis using space-frequency information". In: *Proceedings of the International Conference on Mathematics and Engineering Techniques in medicine and Biological Sciences, METMBS'04*. 2004.

[93] Anil K Jain and Farshid Farrokhnia. "Unsupervised texture segmentation using Gabor filters". In: *Pattern recognition* 24.12 (1991), pp. 1167–1186.

[94] Fergus Davnall et al. "Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice ?" In: *Insights into imaging* 3.6 (2012), pp. 573–589.

[95] Piotr M Szczypiński, MichaeI Strzelecki, Andrzej Materka, and Artur Klepaczko. "MaZda – a software package for image texture analysis". In: *Computer methods and programs in biomedicine* 94.1 (2009), pp. 66–76.

[96] Richard Arnold Johnson and Dean W Wichern. *Applied multivariate statistical analysis*. Vol. 5. 8. Prentice hall Upper Saddle River, NJ, 2002.

[97] Pang-Ning Tan. *Introduction to data mining*. Pearson Education India, 2006.

[98] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[99] *Python programming language*. https://www.python.org/.

[100] Shantanu Sinha et al. "Multifeature analysis of Gd-enhanced MR images of breast lesions". In: *Journal of Magnetic Resonance Imaging* 7.6 (1997), pp. 1016–1026.

[101] Eric Jones, Travis Oliphant, and Pearu Peterson. "{SciPy}: open source scientific tools for {Python}". In: (2014).

[102] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830.

[103] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society. Series B (Methodological)* (1995), pp. 289–300.

[104] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*. Vol. 5. 8. Prentice hall Upper Saddle River, NJ, 2002.

[105] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts. "Machine learning methods for quantitative radiomic biomarkers". In: *Scientific reports* 5 (2015), p. 13087.

[106] S Anari and S Carrie. "Sinonasal inverted papilloma: narrative review". In: *The Journal of Laryngology & Otology* 124.07 (2010), pp. 705–715.

[107] Alexander Sauter, Rubina Matharu, Karl Hormann, and Ramin Naim. "Current advances in the basic research and clinical management of sinonasal inverted papilloma (review)". In: *Oncology reports* 17.3 (2007), pp. 495–504.

[108] Leon Barnes, M Brandwein, PM Som, et al. "Diseases of the nasal cavity, paranasal sinuses, and nasopharynx". In: *Surgical pathology of the head and neck. New York: Informa Healthcare* (2009), pp. 343–422.

[109] S Mirza, PJ Bradley, A Acharya, M Stacey, and NS Jones. "Sinonasal inverted papillomas: recurrence, and synchronous and metachronous malignancy". In: *The Journal of Laryngology & Otology* 121.09 (2007), pp. 857–864.

[110] Myung Woul Han, Bong-Jae Lee, Yong Ju Jang, and Yoo-Sam Chung. "Clinical value of office-based endoscopic incisional biopsy in diagnosis of nasal cavity masses". In: *Otolaryngology–Head and Neck Surgery* 143.3 (2010), pp. 341–347.

[111] Roberto Maroldi et al. "Magnetic resonance imaging findings of inverted papilloma: differential diagnosis with malignant sinonasal tumors". In: *American journal of rhinology* 18.5 (2004), pp. 305–310.

[112] Peter M Som, William Lawson, and Mika W Lidov. "Simulated aggressive skull base erosion in response to benign sinonasal disease." In: *Radiology* 180.3 (1991), pp. 755–759.

[113] Tae Yeon Jeon et al. "18F-FDG PET/CT findings of sinonasal inverted papilloma with or without coexistent malignancy: comparison with MR imaging findings in eight patients". In: *Neuroradiology* 51.4 (2009), pp. 265–271.

[114] TY Jeon et al. "Sinonasal inverted papilloma: value of convoluted cerebriform pattern on MR imaging". In: *American Journal of Neuroradiology* 29.8 (2008), pp. 1556–1560.

[115] Hiroya Ojiri, Masuo Ujita, Shimpei Tada, and Kunihiko Fukuda. "Potentially distinctive features of sinonasal inverted papilloma on MR imaging". In: *American Journal of Roentgenology* 175.2 (2000), pp. 465–468.

[116] David M Yousem et al. "Inverted papilloma: evaluation with MR imaging." In: *Radiology* 185.2 (1992), pp. 501–505.

[117] Anna M Brown et al. "Multi-institutional validation of a novel textural analysis tool for preoperative stratification of suspected thyroid tumors on diffusion-weighted MRI". In: *Magnetic resonance in medicine* (2015).

[118] K Buch et al. "Using texture analysis to determine human papillomavirus status of oropharyngeal squamous cell carcinomas on CT". In: *American Journal of Neuroradiology* 36.7 (2015), pp. 1343–1348.

[119] M Dang et al. "MRI texture analysis predicts p53 status in head and neck squamous cell carcinoma". In: *American Journal of Neuroradiology* 36.1 (2015), pp. 166–170.

[120] Julia Fruehwald-Pallamar et al. "Texture-based and diffusion-weighted discrimination of parotid gland lesions on MR images at 3.0 Tesla". In: *NMR in biomedicine* 26.11 (2013), pp. 1372–1379.

[121] Akifumi Fujita et al. "Difference between HPV-positive and HPV-negative non-oropharyngeal head and neck cancer: texture analysis features on CT". In: *Journal of computer assisted tomography* 40.1 (2016), pp. 43–47.

[122] Jia Liu et al. "Use of texture analysis based on contrast-enhanced MRI to predict treatment response to chemoradiotherapy in nasopharyngeal carcinoma". In: *Journal of Magnetic Resonance Imaging* (2016).

[123] Elisa Scalco, Claudio Fiorino, Giovanni Mauro Cattaneo, Giuseppe Sanguineti, and Giovanna Rizzo. "Texture analysis for the assessment of structural changes in parotid glands induced by radiotherapy". In: *Radiotherapy and Oncology* 109.3 (2013), pp. 384–387.

[124] BT Yang, SP Li, YZ Wang, JY Dong, and ZC Wang. "Routine and dynamic MR imaging study of lobular capillary hemangioma of the nasal cavity with comparison to inverting papilloma". In: *American Journal of Neuroradiology* 34.11 (2013), pp. 2202–2207.

[125] Haowei Zhang et al. "Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy". In: *Radiology* 269.3 (2013), pp. 801–809.

[126] Robert M Haralick, Karthikeyan Shanmugam, et al. "Textural features for image classification". In: *IEEE Transactions on systems, man, and cybernetics* 3.6 (1973), pp. 610–621.

[127] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.7 (2002), pp. 971–987.

[128] John H Krouse. "Development of a staging system for inverted papilloma". In: *The Laryngoscope* 110.6 (2000), pp. 965–968.

[129] Stephen B Edge and Carolyn C Compton. "The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM". In: *Annals of surgical oncology* 17.6 (2010), pp. 1471–1474.

[130] Erez Eyal et al. "Principal component analysis of dynamic contrast enhanced MRI in human prostate cancer". In: *Investigative radiology* 45.4 (2010), pp. 174–181.

[131] IT Jolliffe. *Principal component analysis*. 2003.

[132] Timothy Pok Chi Yeung et al. "Improving quantitative CT perfusion parameter measurements using principal component analysis". In: *Academic radiology* 21.5 (2014), pp. 624–632.

[133] Sarah J Dixon and Richard G Brereton. "Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure". In: *Chemometrics and Intelligent Laboratory Systems* 95.1 (2009), pp. 1–17.

[134] Geng-Cheng Lin, Wen-June Wang, Chuin-Mu Wang, and Sheng-Yih Sun. "Automated classification of multi-spectral MR images using linear discriminant analysis". In: *Computerized Medical Imaging and Graphics* 34.4 (2010), pp. 251–268.

[135] N Fujima et al. "Differentiation of squamous cell carcinoma and inverted papilloma using non-invasive MR perfusion imaging". In: *Dentomaxillofacial Radiology* 44.9 (2015), p. 20150074.

[136] Wang Xinyan, Zhang Zhengyu, Chen Xiaoli, Li Jing, and Xian Junfang. "Value of magnetic resonance imaging including dynamic contrast-enhanced magnetic resonance imaging in differentiation between inverted papilloma and malignant tumors in the nasal cavity". In: *Chin Med J (Engl)* 127 (2014), pp. 1696–1701.

[137] Yoshitaka Miyazaki et al. "Non-necrotic invasive squamous cell carcinoma associated with an inverted papilloma: MRI features". In: *Radiation medicine* 24.2 (2006), pp. 143–146.

[138] RA Lerski et al. "Multicentre magnetic resonance texture analysis trial using reticulated foam test objects". In: *Magnetic resonance imaging* 17.7 (1999), pp. 1025–1031.

[139] Marius E Mayerhoefer, Pavol Szomolanyi, Daniel Jirak, Andrzej Materka, and Siegfried Trattnig. "Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study". In: *Medical physics* 36.4 (2009), pp. 1236–1243.

[140] J Fruehwald-Pallamar et al. "Texture-Based Analysis of 100 MR Examinations of Head and Neck Tumors–Is It Possible to Discriminate Between Benign and Malignant Masses in a Multicenter Trial?" In: *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Vol. 188. 02. 2016, pp. 195–202.

[141] Karla Kerlikowske et al. "Performance of screening mammography among women with and without a first-degree relative with breast cancer". In: *Annals of Internal Medicine* 133.11 (2000), pp. 855–863.

[142] Margaret T Mandelson et al. "Breast density as a predictor of mammographic detection: comparison of interval-and screen-detected cancers". In: *Journal of the National Cancer Institute* 92.13 (2000), pp. 1081–1087.

[143] Jean Weigert and Sarah Steenbergen. "The Connecticut experiment: the role of ultrasound in the screening of women with dense breasts". In: *The breast journal* 18.6 (2012), pp. 517–522.

[144] Yun-Chung Cheung et al. "Diagnostic performance of dual-energy contrast-enhanced subtracted mammography in dense breasts compared to mammography alone: interobserver blind-reading analysis". In: *European radiology* 24.10 (2014), pp. 2394–2403.

[145] Elzbieta Luczyńska et al. "Contrast-enhanced spectral mammography: comparison with conventional mammography and histopathology in 152 women". In: *Korean journal of radiology* 15.6 (2014), pp. 689–696.

[146] S Badr et al. "Dual-energy contrast-enhanced digital mammography in routine clinical practice in 2013". In: *Diagnostic and interventional imaging* 95.3 (2014), pp. 245–258.

[147] Alberto Stefano Tagliafico et al. "Diagnostic performance of contrast-enhanced spectral mammography: systematic review and meta-analysis". In: *The Breast* 28 (2016), pp. 13–19.

[148] Robyn L Birdwell. "The Preponderance of Evidence Supports Computer-aided Detection for Screening Mammography 1". In: *Radiology* 253.1 (2009), pp. 9–16.

[149] James PB O'Connor, Alan Jackson, Geoff JM Parker, and Gordon C Jayson. "DCE-MRI biomarkers in the clinical evaluation of antiangiogenic and vascular disrupting agents". In: *British journal of cancer* 96.2 (2007), pp. 189–195.

[150] Yuanjie Zheng et al. "STEP: Spatiotemporal enhancement pattern for MR-based breast tumor diagnosis". In: *Medical physics* 36.7 (2009), pp. 3192–3204.

[151] Marcelo Zanchetta Do Nascimento et al. "Classification of masses in mammographic image using wavelet domain features and polynomial classifier". In: *Expert Systems with Applications* 40.15 (2013), pp. 6213–6221.

[152] Carl J D'Orsi. *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. 2013.

[153] Antoine Rosset, Luca Spadola, and Osman Ratib. "OsiriX: an open-source software for navigating in multidimensional DICOM images". In: *Journal of digital imaging* 17.3 (2004), pp. 205–216.

[154] EM Fallenberg et al. "Contrast-enhanced spectral mammography versus MRI: initial results in the detection of breast cancer and assessment of tumour size". In: *European radiology* 24.1 (2014), pp. 256–264.

[155] UC Lalji et al. "Contrast-enhanced spectral mammography in recalls from the Dutch breast cancer screening program: validation of results in a large multireader, multicase study". In: *European radiology* 26.12 (2016), pp. 4371–4379.

[156] Marc BI Lobbes et al. "Contrast-enhanced spectral mammography in patients referred from the breast cancer screening programme". In: *European radiology* 24.7 (2014), pp. 1668–1676.

[157] Laura Liberman et al. "The breast imaging reporting and data system: positive predictive value of mammographic features and final assessment categories." In: *AJR. American journal of roentgenology* 171.1 (1998), pp. 35–40.

[158] Woo Kyung Moon et al. "Quantitative ultrasound analysis for classification of BI-RADS category 3 breast masses". In: *Journal of digital imaging* 26.6 (2013), pp. 1091–1098.

[159] Peter Gibbs and Lindsay W Turnbull. "Textural analysis of contrast-enhanced MR images of the breast". In: *Magnetic resonance in medicine* 50.1 (2003), pp. 92–98.

[160] Shonket Ray, Nicolas D Prionas, Karen K Lindfors, and John M Boone. "Analysis of breast CT lesions using computer-aided diagnosis: an application of neural networks on extracted morphologic and texture features". In: *SPIE Medical Imaging*. International Society for Optics and Photonics. 2012, 83152E–83152E.

[161] Jyoti Parikh et al. "Changes in primary breast cancer heterogeneity may augment midtreatment MR imaging assessment of response to neoadjuvant chemotherapy". In: *Radiology* 272.1 (2014), pp. 100–112.

[162] Balaji Ganeshan et al. "Heterogeneity of focal breast lesions and surrounding tissue assessed by mammographic texture analysis: preliminary evidence of an association with tumor invasion and estrogen receptor status". In: *Frontiers in oncology* 1 (2011), p. 33.

[163] José Hermes Ribas do Nascimento, Vinicius Duval da Silva, and Antonio Carlos Maciel. "Accuracy of sonographic findings in breast cancer: correlation between BI-RADS categories and histological findings". In: *Radiologia Brasileira* 42.4 (2009), pp. 235–240.

[164] Gypsyamber Dsouza et al. "Case–control study of human papillomavirus and oropharyngeal cancer". In: *New England Journal of Medicine* 356.19 (2007), pp. 1944–1956.

[165] Hisham Mehanna et al. “Prevalence of human papillomavirus in oropharyngeal and nonoropharyngeal head and neck cancer systematic review and meta-analysis of trends by time and region”. In: *Head & neck* 35.5 (2013), pp. 747–755.

[166] MA Ororke et al. “Human papillomavirus related head and neck cancer survival: a systematic review and meta-analysis”. In: *Oral oncology* 48.12 (2012), pp. 1191–1201.

[167] Fausto Petrelli, Enrico Sarti, and Sandro Barni. “Predictive value of human papillomavirus in oropharyngeal carcinoma treated with radiotherapy: An updated systematic review and meta-analysis of 30 trials”. In: *Head & neck* 36.5 (2014), pp. 750–759.

[168] Haitham Mirghani et al. “Human papilloma virus testing in oropharyngeal squamous cell carcinoma: what the clinician should know”. In: *Oral oncology* 50.1 (2014), pp. 1–9.

[169] Theodoros Rampias, Clarence Sasaki, and Amanda Psyrri. “Molecular mechanisms of HPV induced carcinogenesis in head and neck”. In: *Oral oncology* 50.5 (2014), pp. 356–363.

[170] Xinyi Cindy Zhang et al. “Tumor evolution and intratumor heterogeneity of an oropharyngeal squamous cell carcinoma revealed by whole-genome sequencing”. In: *Neoplasia* 15.12 (2013), 1371IN6–1378IN7.

[171] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. “Radiomics: images are more than pictures, they are data”. In: *Radiology* 278.2 (2015), pp. 563–577.

[172] K Buch et al. “Using texture analysis to determine human papillomavirus status of oropharyngeal squamous cell carcinomas on CT”. In: *American Journal of Neuroradiology* 36.7 (2015), pp. 1343–1348.

[173] American Joint Committee on Cancer. “Pharynx (Including Base of Tongue, Soft Palate, and Uvula)”. In: *AJCC Cancer Staging Manual, Seventh Edition. New York, NY*. Springer, 2010, pp. 41–56.

[174] Samir K. El-Mofty. “Human papillomavirus-related head and neck squamous cell carcinoma variants”. In: *Seminars in Diagnostic Pathology* 32.1 (2015), pp. 23–31.

[175] Sarah C Cantrell et al. “Differences in imaging characteristics of HPV-positive and HPV-negative oropharyngeal cancers: a blinded matched-pair analysis”. In: *American Journal of Neuroradiology* 34.10 (2013), pp. 2005–2009.

[176] David Goldenberg et al. “Cystic lymph node metastasis in patients with head and neck cancer: An HPV-associated phenomenon”. In: *Head & neck* 30.7 (2008), pp. 898–903.

[177] Erez Eyal et al. “Principal component analysis of dynamic contrast enhanced MRI in human prostate cancer”. In: *Investigative radiology* 45.4 (2010), pp. 174–181.

[178] Ian Jolliffe. *Principal component analysis*. Wiley StatsRef: Statistical Reference Online, 2014.

[179] Timothy Pok Chi Yeung et al. “Improving quantitative CT perfusion parameter measurements using principal component analysis”. In: *Academic radiology* 21.5 (2014), pp. 624–632.

[180] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[181] Yang Liu, Nitesh V Chawla, Mary P Harper, Elizabeth Shriberg, and Andreas Stolcke. "A study in machine learning from imbalanced data for sentence boundary detection in speech". In: *Computer Speech & Language* 20.4 (2006), pp. 468–494.

[182] RC Conceicao, M Ohalloran, M Glavin, and E Jones. "Evaluation of features and classifiers for classification of early-stage breast cancer". In: *Journal of Electromagnetic Waves and Applications* 25.1 (2011), pp. 1–14.

[183] Jae Won Lee, Jung Bok Lee, Mira Park, and Seuck Heun Song. "An extensive comparison of recent classification tools applied to microarray data". In: *Computational Statistics & Data Analysis* 48.4 (2005), pp. 869–885.

[184] Grace S Shieh, YC Jiang, and Yu-shan Shih. "Comparison of support vector machines to other classifiers using gene expression data". In: *Communications in Statistics, Simulation and Computation* 35.1 (2006), pp. 241–256.

[185] Jieping Ye, Tao Li, Tao Xiong, and Ravi Janardan. "Using uncorrelated discriminant analysis for tissue classification with gene expression data". In: *IEEE Transactions on Computational Biology and Bioinformatics* 1.4 (2004), pp. 181–190.

[186] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society. Series B (Methodological)* (1995), pp. 289–300.

[187] Akifumi Fujita et al. "Difference between HPV-positive and HPV-negative non-oropharyngeal head and neck cancer: texture analysis features on CT". In: *Journal of computer assisted tomography* 40.1 (2016), pp. 43–47.

[188] K Buch et al. "Quantitative Assessment of Variation in CT Parameters on Texture Features: Pilot Study Using a Nonanatomic Phantom". In: *American Journal of Neuroradiology* 38.5 (2017), pp. 981–985.

[189] Kenneth A Miles, Balaji Ganeshan, Matthew R Griffiths, Rupert CD Young, and Christopher R Chatwin. "Colorectal cancer: texture analysis of portal phase hepatic CT images as a potential marker of survival 1". In: *Radiology* 250.2 (2009), pp. 444–452.

[190] Mitsuhiko Nakahira, Naoko Saito, Hiroshi Yamaguchi, Kiyomi Kuba, and Masashi Sugasawa. "Use of quantitative diffusion-weighted magnetic resonance imaging to predict human papilloma virus status in patients with oropharyngeal squamous cell carcinoma". In: *European Archives of Oto-Rhino-Laryngology* 271.5 (2014), pp. 1219–1225.

[191] Haowei Zhang et al. "Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall

survival in patients treated with induction chemotherapy". In: *Radiology* 269.3 (2013), pp. 801–809.

[192] M Dang et al. "MRI texture analysis predicts p53 status in head and neck squamous cell carcinoma". In: *American Journal of Neuroradiology* 36.1 (2015), pp. 166–170.

[193] Elaine M Smith et al. "Association between p53 and human papillomavirus in head and neck cancer survival". In: *Cancer Epidemiology and Prevention Biomarkers* 17.2 (2008), pp. 421–427.

[194] Haibo He and Edwardo A Garcia. "Learning from imbalanced data". In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[195] Shohei Hido, Hisashi Kashima, and Yutaka Takahashi. "Roughly balanced bagging for imbalanced data". In: *Statistical Analysis and Data Mining* 2.5-6 (2009), pp. 412–426.

[196] Pilsung Kang and Sungzoon Cho. "EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems". In: *Neural Information Processing*. Springer. 2006, pp. 837–846.

[197] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. "Exploratory undersampling for class-imbalance learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.2 (2009), pp. 539–550.

[198] Ronald C Petersen et al. "Mild cognitive impairment: a concept in evolution". In: *Journal of internal medicine* 275.3 (2014), pp. 214–228.

[199] Dennis J Selkoe and John Hardy. "The amyloid hypothesis of Alzheimer's disease at 25 years". In: *EMBO molecular medicine* 8.6 (2016), pp. 595–608.

[200] Heiko Braak and Eva Braak. "Demonstration of amyloid deposits and neurofibrillary changes in whole brain sections". In: *Brain Pathology* 1.3 (1991), pp. 213–216.

[201] Ronald C Petersen et al. "Mild cognitive impairment: ten years later". In: *Archives of neurology* 66.12 (2009), pp. 1447–1455.

[202] Michael Grundman et al. "Mild cognitive impairment can be distinguished from Alzheimer disease and normal aging for clinical trials". In: *Archives of neurology* 61.1 (2004), pp. 59–66.

[203] Ronald C Petersen. "Mild cognitive impairment". In: *CONTINUUM: Lifelong Learning in Neurology* 22.2 Dementia (2016), p. 404.

[204] Robert Perneczky et al. "Mapping scores onto stages: mini-mental state examination and clinical dementia rating". In: *The American journal of geriatric psychiatry* 14.2 (2006), pp. 139–144.

[205] M Folstein Folstein. "A practical method for grading the cognitive state of patients for the clinician". In: *J Psychiatr res* 12 (1975), pp. 189–198.

[206] Kimberly R Chapman et al. “Mini Mental State Examination and Logical Memory scores for entry into Alzheimer’s disease trials”. In: *Alzheimer’s research & therapy* 8.1 (2016), p. 9.

[207] Francisco Franco-Marina et al. “The Mini-mental State Examination revisited: ceiling and floor effects after score adjustment for educational level in an aging Mexican population”. In: *International psychogeriatrics* 22.1 (2010), pp. 72–81.

[208] Robert J Spencer et al. “Psychometric limitations of the mini-mental state examination among nondemented older adults: an evaluation of neurocognitive and magnetic resonance imaging correlates”. In: *Experimental aging research* 39.4 (2013), pp. 382–397.

[209] Charles P Hughes, Leonard Berg, Warren L Danziger, Lawrence A Coben, and RL Martin. “A new clinical scale for the staging of dementia.” In: *The British journal of psychiatry* 140.6 (1982), pp. 566–572.

[210] John Carl Morris et al. “Clinical Dementia Rating training and reliability in multicenter studies The Alzheimer’s Disease Cooperative Study experience”. In: *Neurology* 48.6 (1997), pp. 1508–1510.

[211] John C Morris. “The Clinical Dementia Rating (CDR): current version and scoring rules.” In: *Neurology* (1993).

[212] Ma Shwe Zin Nyunt et al. “Reliability and validity of the Clinical Dementia Rating for community-living elderly subjects without an informant”. In: *Dementia and geriatric cognitive disorders extra* 3.1 (2013), pp. 407–416.

[213] Rémi Cuingnet et al. “Automatic classification of patients with Alzheimer’s disease from structural MRI: a comparison of ten methods using the ADNI database”. In: *neuroimage* 56.2 (2011), pp. 766–781.

[214] Farshad Falahati, Eric Westman, and Andrew Simmons. “Multivariate data analysis and machine learning in Alzheimer’s disease with a focus on structural magnetic resonance imaging”. In: *Journal of Alzheimer’s Disease* 41.3 (2014), pp. 685–708.

[215] Lauge Sørensen et al. “Early detection of Alzheimer’s disease using MRI hippocampal texture”. In: *Human brain mapping* 37.3 (2016), pp. 1148–1161.

[216] Emilie Gerardin et al. “Multidimensional classification of hippocampal shape features discriminates Alzheimer’s disease and mild cognitive impairment from normal aging”. In: *Neuroimage* 47.4 (2009), pp. 1476–1486.

[217] Hakim C Achterberg et al. “Hippocampal shape is predictive for the development of dementia in a normal, elderly population”. In: *Human brain mapping* 35.5 (2014), pp. 2359–2371.

[218] Sergi G Costafreda et al. “Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment”. In: *Neuroimage* 56.1 (2011), pp. 212–219.

[219] WJP Henneman et al. “Hippocampal atrophy rates in Alzheimer disease Added value over whole brain volume measures”. In: *Neurology* 72.11 (2009), pp. 999–1007.

[220] Clifford R Jack et al. “Steps to standardization and validation of hippocampal volumetry as a biomarker in clinical trials and diagnostic criterion for Alzheimer’s disease”. In: *Alzheimer’s & Dementia* 7.4 (2011), pp. 474–485.

[221] Andrea Chincarini et al. “Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer’s disease”. In: *Neuroimage* 58.2 (2011), pp. 469–480.

[222] Roman Filipovych, Christos Davatzikos, Alzheimer’s Disease Neuroimaging Initiative, et al. “Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI)”. In: *NeuroImage* 55.3 (2011), pp. 1109–1119.

[223] Shannon L Risacher et al. “Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort”. In: *Current Alzheimer Research* 6.4 (2009), pp. 347–361.

[224] Jing Zhang, Chunshui Yu, Guilian Jiang, Weifang Liu, and Longzheng Tong. “3D texture analysis on MRI images of Alzheimer’s disease”. In: *Brain imaging and behavior* 6.1 (2012), pp. 61–69.

[225] Lauge Sørensen et al. “Differential diagnosis of mild cognitive impairment and Alzheimer’s disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry”. In: *NeuroImage: Clinical* 13 (2017), pp. 470–482.

[226] Philippe Lambin et al. “Radiomics: extracting more information from medical images using advanced feature analysis”. In: *European journal of cancer* 48.4 (2012), pp. 441–446.

[227] MS De Oliveira et al. “MR imaging texture analysis of the corpus callosum and thalamus in amnestic mild cognitive impairment and mild Alzheimer’s disease”. In: *American Journal of Neuroradiology* 32.1 (2011), pp. 60–66.

[228] *Alzheimer’s Disease Neuroimaging Initiative*. http://adni.loni.usc.edu/. 2016.

[229] Sid E O’Bryant et al. “Staging dementia using Clinical Dementia Rating Scale Sum of Boxes scores: a Texas Alzheimer’s research consortium study”. In: *Archives of neurology* 65.8 (2008), pp. 1091–1095.

[230] Leonard Berg, Daniel W McKeel, J Philip Miller, Jack Baty, and John C Morris. “Neuropathological indexes of Alzheimer’s disease in demented and nondemented persons aged 80 years and older”. In: *Archives of Neurology* 50.4 (1993), pp. 349–358.

[231] *ADNI General Procedures Manual.* http://adni.loni.usc.edu/wpcontent/uploads/2010/09/ADNI_GeneralProceduresManual.pdf.. 2016.

[232] *MRI Pre-processing: Image Corrections Provided by ADNI.* http://adni.loni.usc.edu/methods/mri-analysis/mri-pre-processing/.. 2016.

[233] *MIPAV (Medical Image Processing, Analysis, and Visualization) application*. https://mipav.cit.nih.gov/. 2016.

[234] *volBrain: An online MRI brain volumetry system.* http://volbrain.upv.es/.. 2016.

[235] Stefan Van der Walt et al. "scikit-image: image processing in Python". In: *PeerJ* 2 (2014), e453.

[236] Luis Pedro Coelho. "Mahotas: Open source software for scriptable computer vision". In: *arXiv preprint arXiv:1211.4907* (2012).

[237] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach". In: *Biometrics* (1988), pp. 837–845.

[238] Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis.* Vol. 4. Prentice-Hall New Jersey, 2014.

[239] Paul S Aisen et al. "On the path to 2025: understanding the Alzheimer's disease continuum". In: *Alzheimer's research & therapy* 9.1 (2017), p. 60.

[240] John J Smith, A Gregory Sorensen, and James H Thrall. "Biomarkers in Imaging: Realizing Radiology's Future 1". In: *Radiology* 227.3 (2003), pp. 633–638.

[241] JA Wagner, SA Williams, and CJ Webster. "Biomarkers and Surrogate End Points for Fit-for-Purpose Development and Regulatory Evaluation of New Drugs". In: *Clinical Pharmacology & Therapeutics* 81.1 (2007), pp. 104–107.

[242] John C Waterton and Liisa Pylkkanen. "Qualification of imaging biomarkers for oncology drug development". In: *European journal of cancer* 48.4 (2012), pp. 409–415.

[243] Ralph TH Leijenaar et al. "Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability". In: *Acta Oncologica* 52.7 (2013), pp. 1391–1397.

[244] Dennis Mackin et al. "Measuring Computed Tomography Scanner Variability of Radiomics Features." In: *Investigative radiology* 50.11 (2015), pp. 757–765.

[245] Ralph TH Leijenaar et al. "The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis". In: *Scientific reports* 5 (2015), p. 11075.

[246] Adrien Depeursinge. "Chapter 2 ? Multiscale and Multidirectional Biomedical Texture Analysis: Finding the Needle in the Haystack". In: *Biomedical Texture Analysis.* Ed. by Adrien Depeursinge, Omar S. Al-Kadi, and J.Ross Mitchell. Academic Press, 2017, pp. 29–53.

[247] Tom Brosch et al. "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1229–1239.

[248] Konstantinos Kamnitsas et al. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation". In: *Medical image analysis* 36 (2017), pp. 61–78.

[249] Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. "Deep learning and structured prediction for the segmentation of mass in mammograms". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 605–612.

[250] Youngjin Yoo, Tom Brosch, Anthony Traboulsee, David KB Li, and Roger Tam. "Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2014, pp. 117–124.

[251] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2013, pp. 403–410.

[252] Jie-Zhi Cheng et al. "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans". In: *Scientific reports* 6 (2016), p. 24454.

[253] John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. "Representation learning for mammography mass lesion classification with convolutional neural networks". In: *Computer methods and programs in biomedicine* 127 (2016), pp. 248–257.

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

[255] Yaniv Bar et al. "Chest pathology detection using deep learning with non-medical training". In: *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE. 2015, pp. 294–297.

[256] Bram van Ginneken, Arnaud AA Setio, Colin Jacobs, and Francesco Ciompi. "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans". In: *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE. 2015, pp. 286–289.

[257] *ImageNet*. http://www.image-net.org/. Accessed: 2017-09-01.

[258] William Hsu, Mia K Markey, and May D Wang. "Biomedical imaging informatics in the era of precision medicine: progress, challenges, and opportunities". In: *Journal of the American Medical Informatics Association* 20.6 (2013), pp. 1010–1013. DOI: 10.1136/amiajnl-2013-002315.

[259] Michael D Kuo and Shota Yamamoto. "Next generation radiologic-pathologic correlation in oncology: Rad-Path 2.0". In: *American Journal of Roentgenology* 197.4 (2011), pp. 990–997.

[260] Adrien Depeursinge et al. “Fusing visual and clinical information for lung tissue classification in high-resolution computed tomography”. In: *Artificial intelligence in medicine* 50.1 (2010), pp. 13–21.

[261] Andrew J Buckler, Linda Bresolin, N Reed Dunnick, and Daniel C Sullivan. “A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging”. In: *Radiology* 258.3 (2011), pp. 906–914.

[262] *The Cancer Imaging Archive*. http://www.cancerimagingarchive.net.

[263] *Quantitative Imaging Data Warehouse*. http://qidw.rsna.org/.

[264] *Alzheimer's Disease Neuroimaging Initiative*. http://www.adni-info.org/.

APPENDIX A

ABBREVIATIONS

DOST:	Discrete Orthonormal Stockwell Transform
GFB:	Gabor Filter Banks
GLCM:	Gray Level Co-occurrence Matrices
LBP:	Local Binary Patterns
LoGHist:	Laplacian of Gaussian Histogram
ROI:	Region Of Interest
PC:	Principal Component
PCA:	Principal Component Analysis
SVM:	Support Vector Machine
QDA:	Quadratic Discriminant Analysis
DQDA:	DIagonal Quadratic Discriminant Analysis
LDA:	Linear Discriminant Analysis
SFFS:	Sequential Forward Feature Selection
AD:	Alzheimer's Disease
ADNI:	Alzheimer's Disease Neuroimaging Initiative
CDR:	Clinical Dementia Rating
MCI:	Mild Cognitive Impairment
CN:	Cognitively Normal
CSF:	Cerebro-Spinal Fluid
IP:	Inverted Papilloma
SCC:	Squamous Cell Carcinoma
HPV:	Human PapillomaVirus
OPSCC:	OroPharyngeal Squamous Cell Carcinoma
MRI:	Magnetic Resonance Imaging
CT:	Computed Tomography Imaging
PET:	Positron Emission Tomography
T1:	axial T1-weighted MRI pulse sequence
T2:	axial T2-weighted MRI pulse sequence with frequency selective fat-suppression
T1C:	axial T1-weighted post-contrast MRI pulse sequence with frequency selective fat-suppression
CESM:	Contrast-Enhanced Spectral Mammography
MLO:	Mediolateral Oblique
CC:	Cranial-Caudal
FN:	False Negative
FP:	False Positive
PPV:	Positive Predictive Value
NPV:	Negative Predictive Value
ROC:	Receiver Operating Characteristic curve
AUC-ROC:	Area Under the ROC curve

APPENDIX B

COPYRIGHT PERMISSIONS

Table 1.1 is reproduced from Molecular Oncology, 3/1, Hodgson, D.R., Whittaker, R.D., Herath, A., Amakye, D., Clack, G., "Biomarkers in oncology drug development", 2009, pp.24-32, Copyright (2017), with permission from John Wiley and Sons Inc, License Number: 4211460054660.

Figure 2.1 is reproduced from PloS one 10/11, Hu, L.S., Ning, S., Eschbacher, J.M., Gaw, N., Dueck, A.C., Smith, K.A., Nakaji, P., Plasencia, J., Ranjbar, S., Price, S.J. and Tran, N., Loftus, J., Jenkins, R., O'Neill, B.P., Elmquist, W., Baxter, L.C., Gao, F., Frakes., D, Karis, J.P., Zwart, C.M., Swanson, K.R., Sarkaria, J., Wu, T., Mitchell, J.R. and Li, J., "Multi-parametric MRI and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma", 2015, e0141506. PloS one is an open access article distributed under the terms of the Creative Commons Attribution License. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Figure 2.2 is reproduced from Neuro-Oncology, 19/1, Hu, L.S., Ning, S., Eschbacher, J.M., Baxter, L.C., Gaw, N., Ranjbar, S., Plasencia, J., Dueck, A.C., Peng, S., Smith, K.A. and Nakaji, P., "Radiogenomics to characterize regional genetic heterogeneity in glioblastoma", pp.128-137, Copyright (2017), with permission from Oxford University Press, License Number: 4210950484148.

Chapter 1 and 2 are partly reproduced from Ranjbar, S. and Mitchell, J.R., "An Introduction to Radiomics: An Evolving Cornerstone of Precision Medicine", pp. 223-245. In: Biomedical Texture Analysis. Ed. by Depeursinge, A., Al-Kadi, O.S., and Mitchell, J.R., Copyright (2017), with permission from Elsevier, License Number: 4223750954707.

Chapter 4 is reproduced from American Journal of Neuroradiology, 38(5), Ramkumar, S., Ranjbar, S., Ning, S., Lal, D., Zwart, C.M., Wood, C.P., Weindling, S.M., Wu, T., Mitchell, J.R., Li, J. and Hoxworth, J.M., "MRI-Based Texture Analysis to Differentiate Sinonasal Squamous Cell Carcinoma from Inverted Papilloma.", pp.1019-1025. Copyright (2017), with permission from American Society of Neuroradiology, License Number: 4210950287177.

Chapter 5, 6, and 7 are reproduced from manuscripts that have not yet been published. All co-authors have granted their permissions to use the material in this thesis.