Title
Automatic detection of patient identification and positioning errors in radiotherapy treatment using 3D setup images

Author
Jani, Shyam

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Automatic detection of patient identification and positioning errors in radiotherapy treatment using 3D setup images

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Biomedical Physics

by

Shyam Shirish Jani

2015

ABSTRACT OF THE DISSERTATION

Automatic detection of
patient identification and positioning errors in radiotherapy treatment
using 3D setup images

by

Shyam Shirish Jani
Doctor of Philosophy in Biomedical Physics
University of California, Los Angeles, 2015
Professor James Michael Lamb, Chair

Radiation therapy is a complex healthcare operation that uses ionizing radiation for cancer treatment. The success of modern radiotherapy treatment depends on the correct alignment of the radiation beams with the target treatment region in the patient. In the conventional paradigm of image-guided radiation therapy (IGRT), 2D or 3D setup images are taken immediately prior to treatment and are used by radiation therapy technologists to localize the patient to the same position as defined from the reference planning CT dataset. However, numerous reports in the literature have described errors in during this step, which have led to incorrect treatments and potentially significant clinical harm to the patient. In addition, reported errors likely underestimate the true error rate, as many errors may pass by undetected or are simply not reported. The human factor has been shown to play a large role in these errors, where the setup and planning CT imaging registration is not interpreted or performed correctly as per standard practice.

The hypothesis of the proposed study was to address these human errors by developing a workflow that can algorithmically compare 3D setup and planning CT imaging using image similarity metrics. The proposed system, intended to work in an automated and real-time fashion immediately prior to radiotherapy delivery, has the potential to act as a robust second-check safety interlock to prevent any identification or misalignment errors from reaching the patient. As no additional equipment is required in the treatment room or for patient setup, this system adds virtually no additional complexity, time, or cost to the treatment process. It can be applicable to countries around the world and is particularly relevant for developing nations, where higher error rates have been reported due in part to a smaller number of trained personnel.

To simulate errors across multiple imaging platforms, we utilized both 3D-CBCT and 3D-MVCT images from our TrueBeam and TomoTherapy units, respectively. We gathered CBCT images of 83 head-and-neck (H&N), 100 pelvis, and 57 spine patients treated between 2011 and 2014, and MVCT images of 100 H&N, 100 pelvis, and 56 spine patients treated between 2012 and 2014. Our patient identification study involved the generation of same-patient and different-patient image pairs. Our patient misalignment study involved the translation of the setup image of a same-patient image pair away from the correctly registered alignment. H&N and pelvis image pairs were misaligned by 1cm increments up to 5cm in all six anatomical directions, while spine patients were misaligned to adjacent vertebral bodies.

Chapter 2 describes the development of the image similarity workflow. The system requires inputs of the fused image pair and a mask of the body contour, which was automatically generated using commercially-available software. The workflow involves several pre-processing steps, including image resampling, voxel filtering, CT number remapping for Tomo images, and image filtering. Image similarity is assessed by the use of three commonly-used similarity metrics and two custom-developed algorithmic comparisons. After a feature reduction and normalization step, these metrics are used to train and test five unique classification models as discussed in Chapter 3.

Aspects of model evaluation are also discussed in Chapter 3, including misclassification error, k-fold cross-validation, sensitivity, specificity, ROC curves, and more.

Chapter 4 summarizes the results from the workflow. For patient identification, our system can achieve accuracies ranging from 96.4% to 100% across all anatomical sites and both imaging modalities. Spinal misalignments can be detected with less than 5% error across both imaging modalities. Errors of 1.3% and 4.3% have been achieved for 1cm H&N and pelvis shifts, respectively, on MVCT images. For CBCT images, our models generate errors of 9.3/8.5% and 3.1%/3.2% for 1cm and 2cm H&N/pelvis shifts, respectively. Larger shifts result in increased accuracy as well as higher sensitivity and specificity parameters.

Chapter 5 provides an in-depth discussion about the workflow development and its important aspects. There are several potential ways to improve the algorithm in future studies, ranging from specific adjustments in algorithmic design to entirely new approaches of image similarity assessment. Future studies will allow for more robust error detection, contributing towards improved patient safety in radiation therapy treatments.

The dissertation of Shyam Shirish Jani is approved.

Daniel Abraham Low

Nzhde Agazaryan

Minsong Cao

Brent Liu

James Michael Lamb, Committee Chair

University of California, Los Angeles

2015

DEDICATION

To my family,

Shirish, Amita, and Ashish:

Thank you for everything.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AUC	area under the curve
CBCT	cone-beam computed tomography
CC	correlation coefficient
CV	cross-validation
FOV	field-of-view
H&N	head and neck
HU	Hounsfield units
IGRT	image-guided radiation therapy
KNN	k-nearest neighbors
kVCT	kilovoltage computed tomography
LDA	linear discriminant analysis
LR	logistic regression
LR-	negative likelihood ratio
LR+	positive likelihood ratio
MCC	Matthew's correlation coefficient
MCE	misclassification error
MI	mutual information
MVCT	megavoltage computed tomography
NB	naïve Bayes
PCA	principal components analysis
QA	quality assurance
QDA	quadratic discriminant analysis
ROC	receiver operator characteristic
ROI	region of interest
RP	right patient
RT	radiation therapy
SSIM	structural similarity
TBeam	TrueBeam
Tomo	TomoTherapy
TPS	treatment planning system
WP	wrong patient

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. James Lamb. You have taught me how to think like an independent scientist, something that I will carry with me for the rest of my career. I couldn't have been luckier to work with someone of your caliber. Thank you to the rest of my committee – Dr. Daniel Low, it was wonderful working with you over the past five years, and thank you for funding a large portion of my time at UCLA. To Dr. Nzhde Agazaryan and Dr. Minsong Cao: you have both taught me so much about the clinical world of medical physics, and I am grateful for having worked with you both on various clinical research projects. Dr. Brent Liu, thank you for the thoughtful discussions on this dissertation work.

UCLA Radiation Oncology has been an excellent environment for both my personal and academic development as a graduate student. The department truly embodies the culture of an academic institution and the staff has spent countless hours teaching me about various aspects of clinical practice. Thank you to Dr. Steve Tenn, Dr. John DeMarco, Dr. Anand Santhanam, Dr. Dan Ruan, Dr. Ke Sheng, Philip Chow, and Chul Lee for your physics expertise and involvement in many of my research and clinical projects. Thank you to Sherri Alexander and Mary-Ann Hagio for your help in various aspects of patient dosimetry and treatment planning. To Dr. David Thomas, Dr. Alan Wang, Dr. Peng Dong, Dr. George Sayre, and Dr. Amar Kishan, thank you for your friendships and various research collaborations. Justin Silliven, Farhad Sachinvala, Hussen Abdelhayi, and Maria Kebede were invaluable in teaching me about various aspects of patient setup. And finally, thank you to Tania Craig, Carmen Marticorena, and Emmie Caplan for being such fantastic and helpful administrative staff.

I am very grateful to have been a part of the Biomedical Physics graduate program. I want to extend a special thank you to our incredible program director, Dr. Michael McNitt-Gray, who always made time to give me valuable advice and guidance throughout my time at UCLA. Thank you to Terry Moore and Reth Im, our administrative staff, whose hard work and dedication made my life easier in so many ways. Thank you to Dr. Magnus Dahlbom, Dr. Keisuke Iwamoto, Dr. Daniel Ennis, Dr. Chris Cagnon, and Dr. James Sayre for your expertise both inside and outside the classroom. I have thoroughly enjoyed my time with the BMP student body, and I will always remember the fun times we had together over many events and activities around Los Angeles.

My time outside of graduate studies would not have been complete without the people around me. A very special thank you goes out to Punam, who has been such a meaningful and supportive presence by my side over the past four years. To my roommates past and present – Greg, Nate, Koosha, Mike, Myca, and Ellen – your lasting friendships have been life so fun and interesting. Though too many to list, all of my friends – both new and old – have played a huge role in enjoying Los Angeles and beyond. I am very grateful to this diverse city for providing me with many wonderful lifelong memories.

Finally, my accomplishments truly would not have been possible if it weren't for the unwavering support and guidance from my family. To my parents Shirish and Amita and my brother Ashish, thank you for everything over the past six years and more. Your impact on my growth has been immeasurable, and I am incredibly fortunate to have you in my life. I love you all so much.

VITA

EDUCATION

M.S. Biomedical Physics, University of California, Los Angeles 2012
B.S. Bioengineering, University of California, Berkeley 2009

AWARDS

Moses A. Greenfield Award (University of California, Los Angeles) 2014
Norm Baily Award (AAPM Southern California Chapter) 2014
Ursula Mandel Fellowship (University of California, Los Angeles) 2013 - 2014

PEER-REVIEWED PUBLICATIONS

1. Kishan AU, King CR, **Jani SS**, Kang JJ, Steinberg ML, Lamb JM. Pelvic nodal dosing with registration to the prostate: implications for high-risk prostate cancer patients receiving SBRT. *Int J Radiat Oncol Biol Phys* 2015; **91**(4): 832-9.
2. **Jani SS**, Lamb JM, Robinson CG, Dahlbom M, White BM, Low DA. Assessing margin expansions of internal target volumes in 3D and 4D PET: a phantom study. *Annals of Nuclear Medicine* 2015; **29**(1): 100-109.
3. White BM, Santhanam A, Thomas DA, Min Y, Lamb JM, Neylon J, **Jani S**, Gaudio S, Srinivasan S, Ennis D, Low DA. Modeling and incorporating cardiac-induced lung tissue motion in a breathing motion model. *Med Phys* 2014; **41**(4): 043501.
4. Thomas D, Lamb J, White B, **Jani S**, Gaudio S, Lee P, Ruan D, McNitt-Gray M, Low D. A novel fast helical 4D-CT acquisition technique to generate low-noise sorting artifact-free images at user-selected breathing phases. *Int J Radiat Oncol Biol Phys* 2014; **89**(3): 191-8.
5. **Jani SS**, Robinson CG, Dahlbom M, White BM, Thomas DH, Gaudio S, Low DA, Lamb JM. A comparison of amplitude-based and phase-based positron emission tomography gating algorithms for segmentation of internal target volumes of tumors subject to respiratory motion. *Int J Radiat Oncol Biol Phys* 2013; **87**(3): 562-9.
6. Low DA, White BM, Lee PP, Thomas DH, Gaudio S, **Jani SS**, Wu X, Lamb JM. "A novel CT acquisition and analysis technique for breathing motion modeling." *Phys Med Biol* 2013; **58**(11): L31-6.

ABSTRACTS & PRESENTATIONS

1. **Jani S**, O'Connell D, Chow P, Agazaryan N, Low D, Lamb J. Automatic detection of patient identification and patient positioning errors using 3D setup images. *Med Phys* 2014; **41**(6): 96.
2. **Jani S**, Kishan A, O'Connell D, King C, Steinberg M, Low D, Lamb J. Prediction of pelvic nodal coverage using mutual information between cone-beam and planning CTs. *Med Phys* 2014; **41**(6): 198.
3. O'Connell D, Chow P, Agazaryan N, **Jani S**, Low D, Lamb J. Prediction of dosimetric endpoints from patient geometry using neural nets. *Med Phys* 2014; **41**(6): 397.
4. Thomas D, Tan J, Neylon J, Dou T, **Jani S**, Lamb J, Low D. Investigating the minimum scan parameters required to generate free-breathing fast-helical CT scans without motion-artifacts. *Med Phys* 2014; **41**(6): 559.
5. Kishan AU, King CR, **Jani S**, Steinberg ML, Lamb J. Pelvic nodal dosing with registration to the prostate: implications for high-risk prostate cancer patients receiving SBRT. *Int J Radiat Oncolo Biol Phys* 2014; **91**(1): S410.
6. Low DA, Thomas DH, Lamb JM, Lee PP, Gaudio S, **Jani S**, Dou T, White B, Wu X. Comparison between existing and proposed 4DCT protocols. *Radiother Oncol* 2014; **111**: S196-S197.

7. Thomas D, White B, Gaudio S, **Jani S**, Lee P, Lamb J, Low D. A novel 4D CT acquisition and analysis technique to account for the effect of cardiac induced lung tissue motion during free breathing. *Med Phys* 2013; **40**(6): 411.
8. Thomas D, White B, Gaudio S, **Jani S**, Lee P, Lamb J, Low D. A novel 4D CT acquisition and analysis technique to generate low noise artifact-free images at user selected breathing phases. **Awarded Best in Physics (Joint Imaging-Therapy)**. *Med Phys* 2013; **40**(6): 456.
9. Gaudio S, Thomas D, White B, **Jani S**, Lee P, Lamb J, Low D. Breathing motion comparison inside and outside the lung. *Med Phys* 2013; **40**(6): 182.
10. Wu X, **Jani S**, Dahlbom M, Low D, Lamb J. Comparing the accuracy of the bilateral filter and Gaussian filter for PET image post-processing through a phantom study. *Med Phys* 2013; **40**(6): 144.
11. **Jani S**, Dahlbom M, White B, Thomas D, Gaudio S, Low D, Lamb J. Comparison of gating algorithms in 4D-PET for mobile tumor segmentation. *Med Phys* 2013; **40**(6): 106.
12. Low D, Thomas D, White B, Gaudio S, **Jani S**, Lee P, Lamb J. Development of a prospective gating algorithm for a novel 4DCT technique: retrospective data analysis. *Med Phys* 2013; **40**(6): 179.
13. Lamb J, **Jani S**, White B, Thomas D, Gaudio S, Robinson C, Low D. Ground-truth tests of deformable image registration using matched PET-CT image pairs. *Med Phys* 2013; **40**(6): 169.
14. Aliotta E, Thomas D, Gaudio S, White B, **Jani S**, Lee P, Lamb J, Low D. Improving image quality in 4D-CT scans using deformable registration and selective averaging. *Med Phys* 2013; **40**(6): 434.
15. White B, Thomas D, Lamb J, **Jani S**, Gaudio S, Min Y, Srinivasan S, Ennis D, Santhanam A, Low D. Modeling cardiac induced lung tissue motion for a quantitative breathing motion model. *Med Phys* 2013; **40**(6): 470.
16. White B, Santhanam A, Wang Z, **Jani S**, Lamb JM, Ennis D, Ruan D, Low D. Addition of cardiac motion in a breathing motion model. *Int J Radiat Oncolo Biol Phys* 2012; **84**(3): S746.
17. **Jani S**, Lamb JM, Dahlbom M, Robinson C, White B, Low D. 4D-PET maximum intensity projections improve accuracy of mobile tumor volume delineation. *Med Phys* 2012; **39**(6): 3970.
18. Lamb JM, Lee P, **Jani S**, Dahlbom M, White B, Low D. 4D-PET for abdominal tumor target volume generation. *Med Phys* 2012; **39**(6): 3691.
19. **Jani S**, Lamb JM, Dahlbom M, Robinson C, White B, Low D. 4D-PET maximum intensity projections improve accuracy of mobile tumor volume delineation. Poster presentation at the 13th Annual Biomedical Physics Research Colloquium; May 18, 2012; Los Angeles, CA.
20. Lamb JM, Robinson CG, **Jani S**, Laforest R, Bradley JD, Dehdashti F, White BM, Dahlbom M, Lee P, Low DA. Comparison of 4D-PET gating methods with regards to determining internal target volumes of mobile lung tumors. *Int J Radiat Oncolo Biol Phys* 2011; **81**(2): S816.
21. White B, Lamb JM. Zhao T, **Jani S**, Low D. Distribution of tissue divergence in the apex of the lung. *Int J Radiat Oncolo Biol Phys* 2011; **81**(2): S776-7.
22. White B, Zhao T, **Jani S**, Lamb J, Bradley J, Low D. Distribution of hysteresis magnitude during free breathing. *Med Phys* 2011; **38**(6): 3606.
23. **Jani S**, Lamb JM, Robinson C, *et al*. Utility of maximum intensity projections of gated PET images in determining internal target volumes of moving lung tumors: a phantom study. *Med Phys* 2011; **38**(6): 3374.
24. Weinstein G, Jani SK, Li K, **Jani S**, Pothilat J, Lee S, Glazebrook S. "In-vivo comparison of marker-based localization of prostate with conventional anatomical match using stereoscopic kVp x-rays during external beam radiotherapy." *Int J Radiat Oncolo Biol Phys* 2007; **69**(3): S358.

CHAPTER 1: INTRODUCTION

1.1: Medical patient safety: errors and costs

Patient safety is an emerging concept in the healthcare profession that has rapidly been growing in importance over the past two decades. Patient safety is a discipline that involves the reporting and analysis of errors in healthcare, and the application of scientific methods and research to prevent the occurrence of any adverse events. An adverse event is any unexpected medical occurrence that is unrelated to the intended treatment, ranging from minimal to serious harm to a patient. Healthcare systems should have the ability to minimize the incidence and impact of adverse events, as well as maximize recovery from such events. As an evolving field, patient safety is supported by a growing scientific framework and utilizes a transdisciplinary body of knowledge across several fields within and outside healthcare.

The concept of patient safety has been recognized numerous times in the scientific literature. As early as the 1960s, studies have assessed adverse patient reactions and have emphasized the need for care and caution with new medical procedures and measures [1, 2]. Although additional studies showed the extent of adverse events in the following two decades [3-6], the impact of these errors was not fully realized until the early 1990s. A study by Brennan *et al* in 1991 found adverse events in 3.7% in over 30,000 hospital records in New York over the course of a year, with 27.6% of these events due to substandard care or negligence from medical management [7]. 70.5% of these adverse events led to disability lasting for up to six months, and 2.6% caused permanently disabling injuries. A landmark study by Leape *et al* in 1993 found that more than two-thirds of adverse events were preventable and were primarily due to management errors [8]. They concluded that many of these errors were a result of medical care complexity, which involves hospital personnel, equipment, and procedures.

In 1999, the Institute of Medicine released a report detailing medical errors in the United States that attracted national media attention [9]. They estimated that between 44,000 to 98,000

people die in hospitals each year in the United States due to preventable errors alone – and these numbers exceeded the death rates from more commonly-feared threats such as breast cancer, AIDS, and motor vehicle crashes. The report concluded that faulty systems, processes, and conditions are the underlying causes behind the occurrence of mistakes and the failure to prevent them. To achieve a better record of safety, a four-tiered approach was recommended. First, a stronger national focus must be established to create leadership, research, tools, and protocols to enhance the knowledge base about safety, as healthcare is more than a decade behind other high-risk industries with respect to ensuring basic safety. A national public and mandatory reporting system must be developed to identify and learn from errors, and all healthcare practitioners should be encouraged to participate in voluntary reporting systems. Performance standards should be raised through the oversight of professional organizations and groups in order to form safety expectations among both providers and consumers. Finally, healthcare organizations must develop a strong culture of safety and implement such systems to ensure safe practice at the delivery level. The response to this report was rapid, with the Clinton administration issuing an executive order for government agencies to implement proven techniques for reducing medical errors. This order also included the creation of a task force to find novel strategies for reducing error. In December 2000, Congress appropriated $50 million to the Agency for Healthcare Research and Quality (AHRQ) to support many efforts targeting the reduction of medical errors [9].

Medical errors have been found to be prevalent in in other countries as well. One study in Australia revealed as many as 18,000 deaths and more than 50,000 disabled patients from preventable medical errors over the course of a year [10]. An expert group from the Department of Health reported on adverse events in the UK, estimating that over 850,000 incidents harm National Health Service hospitals annually, with an average of forty incidents contributing toward patient deaths in each institution [11]. A Canadian study looked at a range of hospitals throughout Canada and estimated an adverse error rate of 7.5%, estimating between 9,000 and 24,000 Canadians

dying annually from preventable mistakes [12]. Another Canadian study looking at a single hospital found a 12.7% incidence rate of adverse events, with more than one-third deemed to be avoidable [13]. A large-scale New Zealand study across 13 public hospitals over the course of a year found adverse events associated with 12.9% of admissions, with 35% of these classified as highly preventable [14]. A Denmark study across 17 hospitals found an adverse event rate of 9.0% of admissions, with a 40.4% associated preventability rate [15]. **Table 1** summarizes data from adverse events in acute care hospitals in Australia, Denmark, U.K., and the U.S.A. [7, 10, 12, 14].

	Study	Study focus (date ofadmissions)	Number of hospital admissions	Number of adverse events	Adverse event rate (%)
1	United States (Harvard Medical Practice Study)	Acute care hospitals (1984)	30 195	1 133	3.8
2	United States (Utah–Colorado study)	Acute care hospitals (1992)	14 565	475	3.2
3	United States (Utah–Colorado study)[a]	Acute care hospitals (1992)	14 565	787	5.4
4	Australia (Quality in Australian Health Care Study)	Acute care hospitals (1992)	14 179	2 353	16.6
5	Australia (Quality in Australian Health Care Study)[b]	Acute care hospitals (1992)	14 179	1 499	10.6
6	United Kingdom	Acute care hospitals (1999–2000)	1 014	119	11.7
7	Denmark	Acute care hospitals (1998)	1 097	176	9.0

Table 1: Summary of adverse events from acute care hospitals in various countries. (World Health Organization, Executive Board 109th session, provisional agenda item 3.4, 5 Dec. 2001, EB 109/9)
[a,b] Revised using the same methodologies as the Quality in Australian Health Care Study and Utah-Colorado study, respectively (harmonizing the four methodological discrepancies between the two studies). Studies 3 and 5 present the most directly comparable data for these two studies.

The economic impact of adverse medical events has been considerable, stemming primarily from medical expenses, additional hospitalizations, costs from litigations, lost income, and disability. In 1999, the AHRQ reported that the prevention of medical errors in the U.S. has the potential to save approximately $8.8 billion annually [16]. The Institute of Medicine reported total

annual hospital costs of $17-29 billion due to hospital expenses and additional costs such as lost income, household productivity, disability, and additional care necessitated by these errors [9]. A high economic impact can also be seen in other countries as well – for example, the Australian Patient Safety Foundation estimated $18 million of additional insurance costs due to large medical negligence lawsuits between 1997 and 1998 [17]. To settle claims from clinical negligence costs the National Health System in the U.K. £400 million annually [11]. Ultimately, no cost can be placed on the pain, suffering, and loss of independence for the affected patients, their families, and their short and long-term caretakers.

1.2: Theoretical frameworks of error causation

Analyses of many historical large-scale disasters (spacecraft, oil platforms, nuclear power plants, etc.) have shown that the root cause of accidents can be derived from multiple factors, commonly involving aspects such as workplace conditions, organizational decisions, and individual situational factors. In addition, increased complexity has been shown to correlate strongly with an increased potential for errors in a system or organization. Psychologist James Reason worked on the cognitive theory of latent and active error types, hypothesizing that most accidents result from both active errors (e.g. unsafe acts that are directly linked to a safety event) and latent errors (e.g. systematic conditions and practices that lead to such events) [18]. He developed the "Swiss cheese model" (**Figure 1**) to explain accident causation in a system with multiple defense barriers [19]. These barriers can be modeled by slices of Swiss cheese with many holes – representing weaknesses in the defense – that are continually opening, closing, and shifting their location. A system failure occurs when holes in multiple layers temporarily coincide to permit the passage of an accident opportunity.

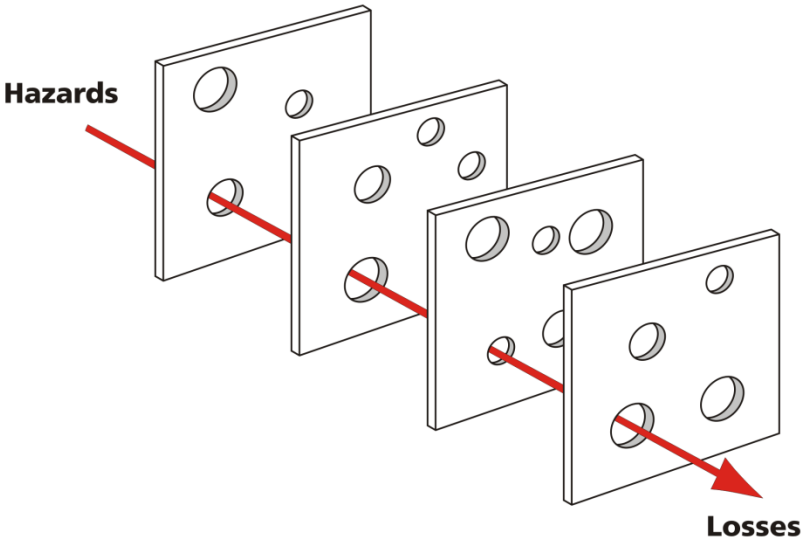


Figure 1: Reason's 'Swiss cheese' model. ("Swiss cheese model of accident causation" by Davidmack - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Swiss_cheese_model_of_accident_causation.png#mediaviewer/File:Swiss_cheese_model_of_accident_causation.png)

Human errors play a large role in the occurrence of accidents in the workplace. It was observed that 60%-80% of accidents on average are directly attributed to "operator error" [20]. From his analysis of many large-scale disasters in the 1980s, Reason also suggested that latent human errors were more significant than technical failures [21]. He classified human errors based on Rasmussen's three levels of performance (**Figure 2**) [22, 23]. Skill-based errors, or slips and lapses, refer to an unintended action. Rule-based mistakes are actions with the correct intention, but with the unintended outcome due to incorrect application to a rule. Knowledge-based mistakes are intended actions with the unintended outcome due to a lack of knowledge. Some specific human factors that fall under these categories are variations in training and experience, fatigue, inattention, and failure to acknowledge the prevalence or severity of errors [24-26].

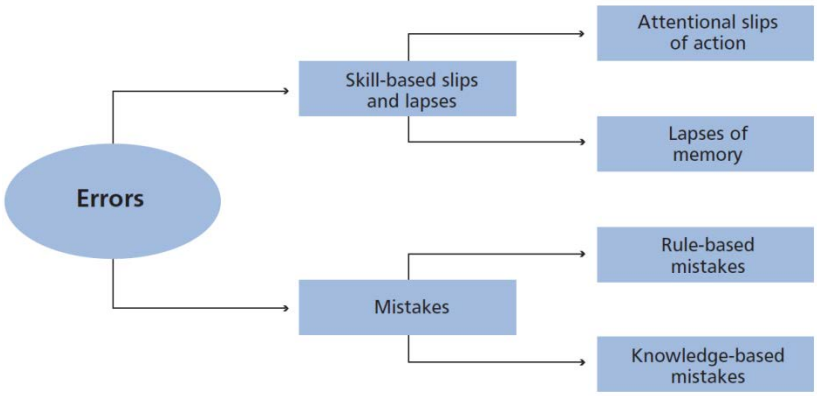


Figure 2: Reason's flowchart of human errors. (WHO Patient Safety Curriculum Guide, 2011)

1.3: External beam radiation therapy treatment: a brief overview

Radiation therapy (RT), which uses ionizing radiation for cancer treatment, is a fitting example of a complex healthcare operation. Radiation works by directly or indirectly ionizing the atoms of the DNA chain in cancerous cells, forming free radicals which thereby damage DNA and induce cellular death [27]. RT has been successfully used as a curative and palliative treatment for a wide array of cancerous and non-cancerous conditions. Over the years, the technology behind RT has continued to evolve to improve patient outcome and minimize damage to non-cancerous tissue. From a review of published evidence, 52% of cancer patients are excellent candidates for some form of RT treatment [28]. Combined with other treatment modalities, such as chemotherapy or surgery, RT treatment plays an important curative role in approximately 40% of cancer patients [29].

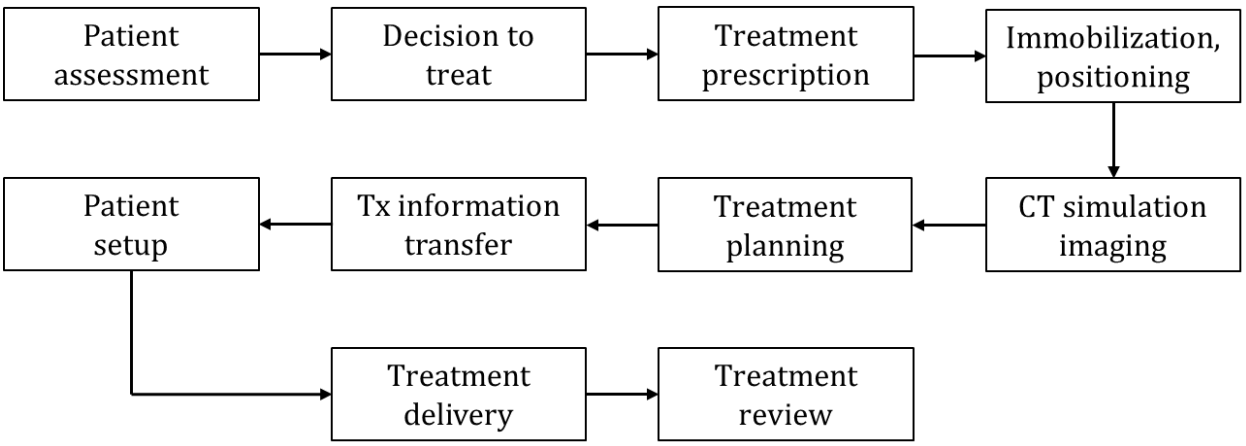


Figure 3: General graphic of the radiation therapy workflow.

A broad summary of the external beam RT workflow can be seen in **Figure 3**. A patient will first undergo a detailed consultation with a radiation oncologist to discuss the potential role of RT in the patient's treatment process. The proposed treatment, potential risks and side effects, and other possible treatment options will be discussed in full with the patient, and the patient will be

asked to sign a consent form. In the next stage, a three-dimensional (3D) computed tomography (CT) scan of the patient will be acquired as the image dataset for planning the RT treatment. During this CT simulation, radiation therapy technologists (hereby referred to as radiation therapists) will identify the target region on the body and mark appropriate locations on the skin (typically using permanent tattoos) in order to provide a frame of reference for patient setup during RT treatment. Immobilization devices – such as molded, patient-specific masks or body cushions – may be used to hold specific body parts in place depending on the treatment site.

The CT dataset from the simulation process, henceforth referred to as the planning CT, is then sent to the dosimetry team. Radiation oncologists and dosimetrists work together to design the patient's treatment plan. Dosimetrists will first contour relevant non-cancerous areas adjacent to the treatment region that will be affected by the RT treatment. The oncologist will then create contours of the appropriate treatment region(s) and provide a specific set of prescriptions (e.g. various dose tolerances for the treatment and non-cancerous regions) for the dosimetrist. The dosimetrist will then use these prescriptions while designing the patient's treatment plan through the use of advanced software to maximize radiation dose to the treatment region while minimizing dose to the non-cancerous areas. Medical physicists may also be responsible for part or all of this planning process.

After approval of the treatment plan by the oncologist, the patient is ready for treatment. Radiation therapists will position and immobilize the patient at the treatment machine in the same position as during the CT simulation process. The patient is then treated with RT. Over the course of the patient's treatment protocol, the oncologist(s) will monitor the patient through weekly checks and progress notes, while medical physicists will conduct weekly chart checks of the recorded treatment information and verify the correct delivery of the prescription and treatment plans.

1.4: Errors in RT

One goal in any healthcare system is to not only reduce the likelihood of adverse events, but to also increase the probability that any potential errors are quickly recognized and addressed. Despite the many quality assurance (QA) protocols and research studies that have been presented to address these concerns [30-37], errors can and still do occur in RT treatments. These errors can occur at many different points throughout the RT workflow as isolated or sequential events, including both equipment failures and operator errors/mistakes [38]. One study estimated 269 potential nodes of error in the planning and delivery stages of RT [39]. Due to the ever-increasing complexity in RT technology, there is ever-increasing potential for treatment errors to occur [40]. This may especially be the case in countries with lower income, due to a smaller number of trained QA personnel [41].

One of the most egregious types of errors – in RT or any healthcare profession – is simply the gross mistreatment of a patient, whether it be the wrong location or the wrong patient. Despite a completely error-free process up until radiation delivery, a patient mistreatment in RT could deliver significant clinical harm to a patient with irreversible consequences, and should therefore be avoided at all cost. This type of error is most likely to occur immediately prior to radiation delivery, or during the patient setup stage.

Numerous reports have been published detailing various errors related to radiotherapy mistreatments. In a 2010 *The New York Times* article, 621 mistakes were found across several hospitals in the state of New York [42]. In 284 occasions, RT treatment missed the intended target (in part or entirely) or treated the wrong target. 50 patients received RT treatment intended for a different patient entirely. In a study of internally-reported clinical incidences from 2007-2009 at a large academic center, 41 of 176 critical incidences were related to an incorrect patient setup, wrong patient treatment, or a geographic miss of the target [43]. In another study that evaluated 100 unintended RT exposures from previously-published reports around the world, a 21% error

incidence rate was found due to incorrect patient setup [44]. These reported error rates underestimate the true error rate, as many errors go by undetected and may not even be reported if noticed [36, 45]. Error rates are also likely to be higher in developing countries and nations due to inadequate knowledge or skills, or heavy staff workloads [41, 46-48]. An external audit of an oncology practice in a developing Asian nation revealed 52% of patients had received suboptimal radiation treatment, where 26.7% of these cases were inadequate for patient setup imaging [48].

1.5: Error prevention during patient setup

There are several methods currently in practice to correctly identify a patient prior to RT treatment. Radiation therapists will often ask the patient open-ended questions prior to setup, including their name, date of birth, and treatment site. A therapist may also visually identify the patient using a photo of the patient, which is often taken during the CT simulation process. Although less commonly used, some clinics use patient-specific identifiers (e.g. wristband with a barcode) for identifying the patient. However, these steps are still prone for accidents. Miscommunications between therapists or lack of attention are human errors that have led to treatments of wrong patients [49-51]. For example, one case study describes a mistreatment involving two back-to-back patients requiring RT of the same anatomical site [52]. A therapist loaded the first patient's treatment plan, but brought in the second patient for setup as the first patient was unavailable. The second patient was subsequently treated using the first patient's plan. This same scenario may also occur with two therapists instead, where miscommunication would be the primary cause of the wrong patient treatment.

Once a patient is set up on the treatment table, treatments requiring precision setup (such as intensity modulated RT (IMRT) or stereotactic body RT (SBRT)) use image guidance. Image-guided radiation therapy (IGRT) is the use of 2D or 3D imaging to localize the patient to the same imaging coordinates as defined from the reference imaging dataset (i.e. the planning CT). In a step

called image registration, these images are used to measure and correct positional errors for target and critical structures in real-time immediately prior to treatment, allowing for precise and accurate patient setup [53]. Examples of 2D IGRT include matching planar kilovoltage (kV) radiographs or megavoltage (MV) portal with 2D projections (digitally reconstructed radiographs, or DRRs) of the patient's planning kVCT [54]. 3D IGRT examples include cone-beam CT (CBCT) images or 3D MVCT images. Errors during this stage of image fusion are primarily related to operator error. Factors such as inattention, fatigue, lack of training, or inexperience have resulted in misaligned patients during treatment, or the allowance of an incorrect patient to be treated [49, 50].

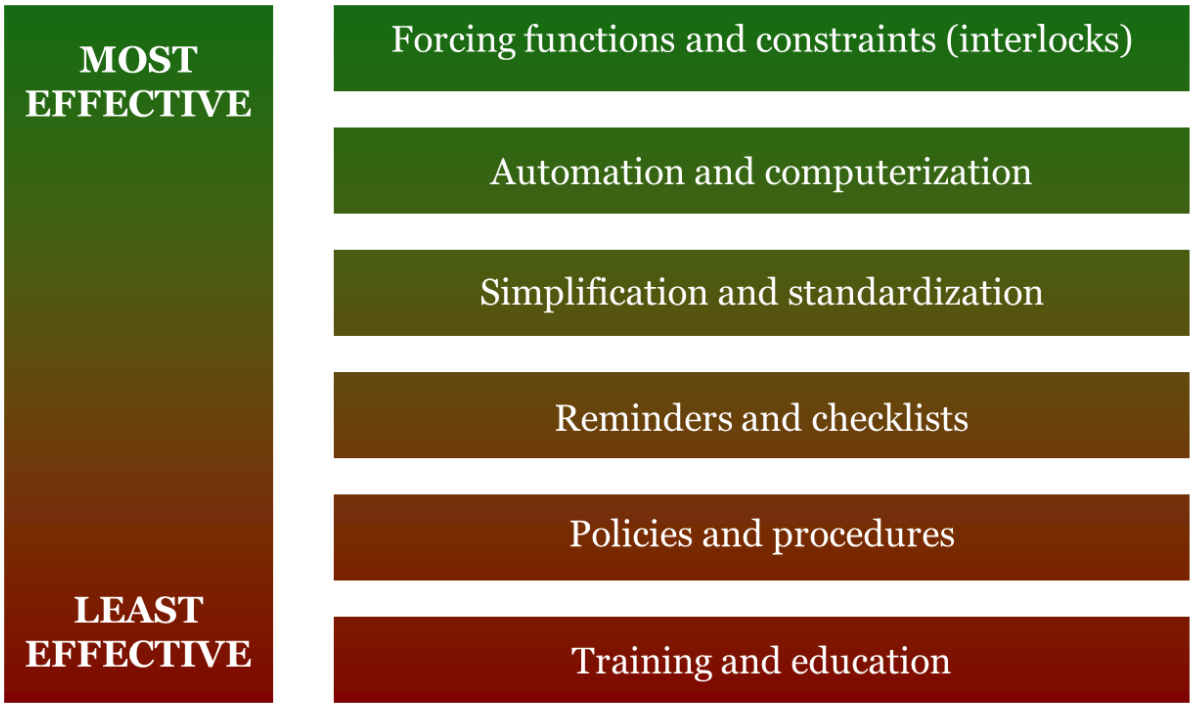


Figure 4: Hierarchy of hazard mitigation effectiveness. Adapted from Hendee [55].

Following the increased public attention to errors in radiation therapy, particularly those highlighted by the *New York Times* articles, a safety-specific meeting in June 2010 was organized by two large organizations in medical physics, the American Association of Physicists in Medicine and

the American Society of Radiation Oncology. Experts inside and outside the field of radiation oncology were gathered to analyze the causes of errors and develop protocols to help prevent the occurrence of future errors. One aspect discussed extensively was the hierarchy of hazard mitigation effectiveness (**Figure 4**), which describes the short-term effectiveness of various implementations in error reduction [55]. As treatment approaches should be fault-tolerant and prevent error occurrence before they can reach the patient, aspects such as automation and forcing functions tend to be more effective than implementing training/education and policies [55].

1.6: Facial recognition and real-time tracking

Facial recognition is the general concept of automatically identifying an individual through some static digital image or a video source. The process involves face detection, feature extraction and normalization, and identification or verification of the individual [56]. This method has a potential application in the clinic as a secondary check for patient identification, as a daily image or video could be taken of the patient and compared to a database of images taken during the CT simulation stage. However, facial recognition can have difficulty performing in a number of variable conditions, including the viewing angle of the face, poor lighting, low resolution images, or objects partially covering the subject's face [57].

There are few studies involving real-time error tracking in the RT framework. One study used two mounted video cameras in the treatment room, a computer-controlled tabletop, and a dento-maxillary fixation device to monitor head and neck (H&N) cancer treatments [58]. A recent study used mounted charge-coupled device (CCD) cameras to track infrared markers placed on an immobilization device or on the patient directly [59]. Another recent study proposed a safety framework using 3D cameras to detect the treatment of a wrong patient or anatomic site [60]. There is significant value in having a real-time monitoring system for preventing patient mistreatments. This setup, however, along with any system involving camera-based recognition,

would require the costly acquisition of complex equipment and is unlikely to be feasible across all radiation treatment centers. In addition, the need to attach something to the patient (as described in [58, 59]) could introduce another source of human error, as this step relies on the correct placement of said markers by the radiation therapist. Finally, these additional steps and equipment add both time and complexity to an already lengthy and complex process, which could hinder the daily clinical workflow.

1.7: Drawbacks of IGRT

One of the inherent flaws in currently-available IGRT systems is that they do not provide information on the quality of image fusion. The system will not alert the therapist about any fundamental discrepancies between these two images. After fusion, the only information available to the therapist is the 3D couch shifts necessary to align the setup image to the planning kVCT image. **Figure 5** shows how a flashlight can successfully be registered to a spine using built-in automatic registration software of a 2D image matching system. Although this situation would never arise in a clinical workflow, it demonstrates one fundamental flaw of these systems: they will always output some numerical shift for the therapists, regardless of the input images.

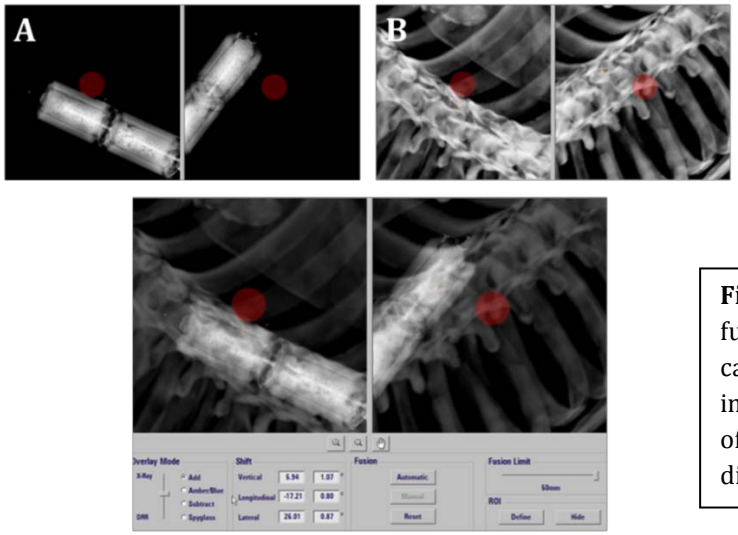


Figure 5: Example of the ExacTrac fusion: a flashlight (A) and spine (B) can be successfully fused (bottom image) without alerting the therapist of any error, despite the obvious dissimilarity of the two images.

Most IGRT systems have a built-in automatic registration system to help guide the therapist in the process of image fusion. However, this software is sometimes not used in clinical practice. The personal preference of many therapists is to fuse images manually, and they tend to trust their own judgment more than the automatic software. In addition, therapists will oftentimes use the magnitude of the output shift as a general benchmark for patient alignment. This value can vary depending on the treatment site, but a value of ≤1cm is commonly used as an indicator for correct patient setup. However, the automated registration software may display shifts under 1cm for an incorrect fusion, potentially leading to an incorrectly positioned patient during treatment. **Figure 6** shows an example of this using automatic registration software for a previously-treated prostate patient. A setup image correctly matched to its corresponding planning kVCT at coordinate (0,0,0) was intentionally misaligned by 2cm in the superior direction to (0,0,2). The automatic registration results in an incorrect match at the coordinates (0.1,0,1.4). If a therapist had aligned the patient incorrectly during image fusion, the automatic registration would have displayed shifts less than 1cm, which may have resulted in therapist approval and a subsequent patient mistreatment.

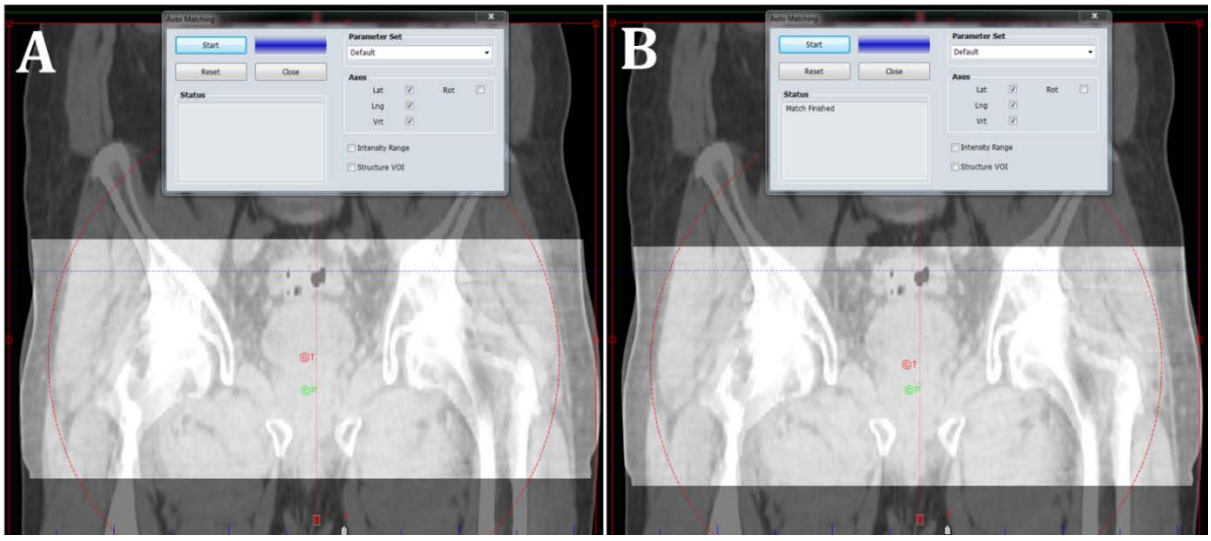


Figure 6: Example of an auto-registration failure. In image (A), the setup is intentionally misaligned by 2cm superior to the planning CT image. After running the auto-registration, the resultant shifts are 0.6cm in the inferior direction and 0.1cm in the vertical direction, and not by 2cm inferiorly to recreate the original correct alignment.

1.8: IGRT as a means for error detection

Given the history of errors in healthcare, specifically those in RT, it is clear that human error plays a large role in the occurrence of adverse events. Despite the continued implementation of safety protocols, the increased complexity of RT technology introduces a large potential for operator error. The use of technology such as real-time tracking has great potential for the future of RT and healthcare safety, but suffers from both complexity and costs that would not be feasibly adaptable by many clinics around the world. To address these concerns, a question therefore arises: how can we build an accurate, automated, and inexpensive system for the purpose of patient safety?

In 2013, Lamb *et al* proposed a novel idea to use IGRT as a means of patient safety [61]. In the comparison of a patient setup image and a corresponding planning kVCT image, they hypothesized that the uniqueness of a patient's internal anatomy would be able to discriminate between two images of the same person or of different people. As most modern treatment machines possess IGRT technology, this concept would be readily applicable to the majority of RT clinics. They performed a pilot study using the ExacTrac (BrainLAB) system (shown in **Figure 5**), where initial patient positioning is performed via cameras that generate and detect infrared radiation reflected off markers placed on the patient's skin [62]. Final patient positioning occurs with radiographic image guidance via a pair of amorphous silicon detectors affixed to the ceiling and two floor-mounted x-ray sources [63]. Two kilovoltage radiographs centered on the radiation isocenter are acquired and compared to DRRs from the planning kVCT dataset. The ExacTrac system generates these DRRs in the same projected field of view as the kilovoltage images, and creates an 'ideal' pair of DRRs for the patient's initial positioning. The system then iteratively recomputes DRRs and uses quadratic convergence to optimize the comparison of these images, outputting appropriate shifts at the treatment console [64]. By extracting the similarity metric used in this system, histograms could be formed between same-patient and different-patient image pairs.

Dr. Lamb *et al* found that this measure was able to identify cranial, prostate, and misaligned spinal vertebral bodies with misclassification errors of 0%, 0.45%, and 1.4% respectively. **Figure 7** shows example 1D histograms of correct and incorrect image pairs for the prostate case.

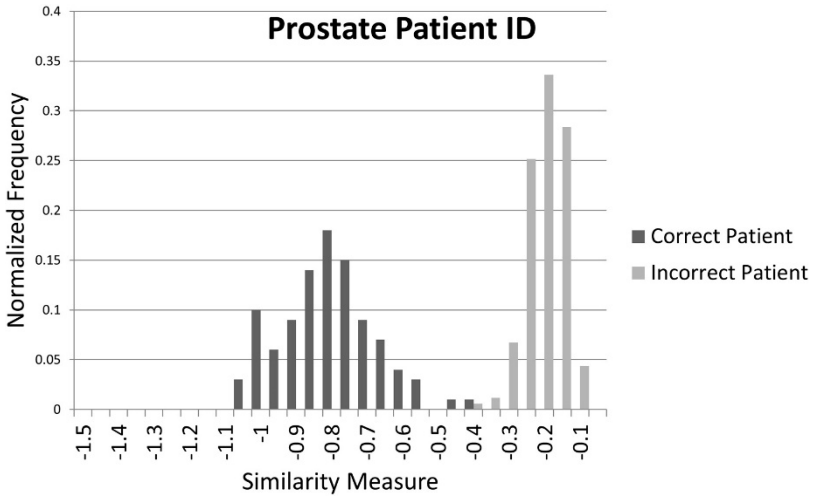


Figure 7: Histogram showing the large separation between histograms of correct and incorrect prostate image pairs using a gradient-based correlation coefficient metric (Lamb *et al* [61])

1.9: Hypothesis and specific aims

This dissertation will build upon the pilot study of Lamb *et al*, with the primary hypothesis that the general technique can be extended to 3D setup images. To explore how adaptable this technique is to different imaging platforms, setup images from two different modalities (kV and MV) have been used. To further generalize the technique, three commonly-treated anatomical regions have been explored: H&N, pelvis, and spine. Finally, we hypothesize that in addition to detecting patient identification errors, this technique can also be used to detect errors in patient alignment.

There are numerous benefits to implementing this type of system in clinical practice. First, it has the potential to become a robust, automated second layer of patient safety following image fusion by the therapists. There is minimal cost to implement such a system, as no extra cameras or patient equipment are required. In addition, there is no reliance on the therapists to set up markers or devices on the patient, thus reducing any detrimental impact on the time of the clinical workflow.

There is also a great potential for cost savings for both the patient and the clinic by avoiding negligence lawsuits, lost income, decreased quality of life, increased need for care, and more. Should this system prove robust enough, the question also arises if the presence of two therapists is always needed at the treatment console as per standard practice [50], allowing for large annual savings by the clinic. Finally – and perhaps most importantly – this system would have a particularly strong impact in developing nations, where it has been shown that lack of knowledge and high staff workload have frequently led to inadequate patient treatments [46-48, 65].

Chapter 2 will outline the first specific aim: to develop algorithmic workflows for 3D kVCT-based and 3D MVCT-based setup imaging that detect patient identification and misalignment errors.

Chapter 3 will outline the second and third specific aims. Specific aim 2 was to build classification models to optimize the accuracy of the proposed algorithmic workflows. Specific aim 3 was to evaluate and validate the proposed models using setup images from three different anatomical sites.

Chapter 4 will show the results of the algorithmic workflow as described in chapters 2 and 3 for both patient identification and patient misalignment studies.

Chapter 5 will provide a discussion of the results with considerations to clinical implementation. Future studies and directions will also be proposed and discussed.

CHAPTER 2: WORKFLOW DEVELOPMENT

2.1: kV-CBCT imaging

We used setup images from two different 3D image guidance technologies commonly used for target localization prior to treatment. The first was a Varian TrueBeam linear accelerator using 3D kilovoltage CBCT images. CBCT imaging uses a conical geometry between the imaging source and a 2D detector, allowing for the acquisition of a volumetric dataset with a single gantry rotation [66]. CBCT imaging allows for high resolution images at doses comparable to or lower than conventional multi-detector CT scans [66-70]. However, the wider collimator leads to increased scatter radiation, leading to reduced contrast resolution and increased noise [66, 67]. Streak and cupping artifacts also contribute to image degradation [71]. Antiscatter grids, a set of lead leaves that are fitted over the panel detectors, are used to reduce cupping artifacts and scattered photons, although at the expense of increased image noise [68, 72]. An increased imaging dose or a reduction in image resolution would be necessary to offset this increased noise [66]. CBCT images are reconstructed through a three-step process: pre-processing of the projections, iterative filtered backprojection, and image post-processing.

2.2: MVCT imaging

The TomoTherapy unit combines features of a linear accelerator and a helical CT scanner by delivering radiation and generating CT images from the same MV x-ray beam, allowing for both IGRT and RT using the same treatment geometry [69, 73-75]. Compared to diagnostic CT images, MVCT images have lower longitudinal resolution as well as lower contrast resolution due to increased Compton interactions – and hence lower interaction probability – from the higher energy photons [76-78]. In addition, MVCT images suffer from increased noise due to the low efficiency (~1%) of the xenon detector arrays [69, 79]. The number of photons can be increased to compensate for image quality, but at the expense of increased patient dose.

MVCT images are reconstructed using a filtered backprojection technique, which inverts the 2D Radon transform of an object [80]. The MVCT images are reconstructed slice-by-slice along the longitudinal direction based on the fan-beam geometry with curved detectors [81]. Images are processed by a variety of methods, including dark-current subtraction, reference-channel normalization, logarithmic conversation of transmission data, and spectral correction [79].

2.3: Patient data acquisition

From here onwards, TrueBeam and TomoTherapy will be abbreviated as 'TBeam' and 'Tomo,' respectively. All images in the study were acquired from patients treated at UCLA's Department of Radiation Oncology between 2011-2014 (TBeam) and 2012-2014 (Tomo). Planning kVCTs were captured using a SOMATOM Sensation CT scanner (Siemens Medical Solutions; Munich, Germany), Biograph 64 TruePoint PET/CT scanner (Siemens Medical Solutions; Munich, Germany), or Brilliance CT Big Bore (Philips; Amsterdam, Netherlands). For TBeam patients, the software package ARIA (Varian Medical Systems; Palo Alto, CA) – specifically the Offline Review tool – was used to export planning kVCTs and CBCTs to a local personal computer (PC). Registration files containing the registration of the planning kVCT and CBCT that was actually performed by the therapists at the time of treatment were available and also exported. All files were exported in DICOM format. Planning kVCT and MVCT images of Tomo patients were exported directly from the TomoTherapy treatment planning system (TPS). Exportable registration files were not available from the Tomo system at the time of this work.

One setup image and one planning kVCT image were exported for each patient. In order to provide consistent image characteristics and viewing windows across different patients, the following criteria were used to guide the selection of the setup image:

1) Field-of-view (FOV):
    a. *H&N*: superior-most slice must not be inferior to the maxillae; inferior-most slice must not be superior to maxillae.
    b. *Pelvis*: superior-most slice must be inferior to the L5 vertebrae; FOV must encompass the pubic symphysis.
    c. *Spine*: FOV must encompass thoracic vertebrae; inferior-most slice must not be inferior to L2.
2) Time of acquisition: the minimum time elapsed between the planning kVCT and setup image was selected, provided the other criteria were also satisfied.
3) Image quality: absence of extreme image artifacts for CBCT images (qualitatively determined). Four images were excluded for this reason.

Due to a limited number of spinal treatments with 3D setup imaging, lung patient images were also included in the spine dataset, provided the above criteria were satisfied for the setup image. **Tables 2** and **3** lists various attributes and properties of setup images from both machines.

		# pts	SO	B64	BBB	kVp	Exposure (mAs)	x/y res (mm)	z res (mm)
TBeam	*H&N*	83	72	5	6	120*	[43, 501]**	[0.63, 1.37]***	[1.5, 5]****
	Pelvis	100	57	38	5	120†	[71, 451]††	[0.88, 1.60]†††	[1.5, 5]††††
	Spine	57	36	6	16	120‡	[64, 464]‡‡	[0.82, 1.6]‡‡‡	[1.5, 3]‡‡‡‡

Table 2: Relevant image parameters for the TBeam planning CT datasets. Entries with brackets indicate the minimum and maximum values.
pts = number of patients; SO = Sensation Open; B64 = Biograph 64; BBB = Brilliance Big Bore; kVp peak kilovoltage; res = resolution
*: one pt had 90kVp; **: 67/83 pts had 120mAs; ***: 36/83 pts had 0.9766mm; 72/83 pts had res>0.9766mm; ****: 75/83 had 3mm
†: one pt had 140kVp; ††:57/100 had 250mAs, 32/100 had 400 mAs; †††: 33/100 had 1.17mm; 19/100 had 0.97mm, 17/100 had 0.90mm; ††††: 86/100 had 3mm
‡: one pt had 140 kVp; ‡‡: 21/57 had 300kVp, 15/57 had 250 kVp; ‡‡‡: 34/57 had 0.98mm; ‡‡‡‡: 30/57 had 1.5mm, 25/57 had 3mm

		# pts	SO	B64	BBB	kVp	Exposure (mAs)	x/y res (mm)	z res (mm)
Tomo	*H&N*	100	N/A°	N/A°	N/A°	N/A°	N/A°	[1.56, 3.20]*	[1.5, 3]**
	Pelvis	100	N/A°	N/A°	N/A°	N/A°	N/A°	[1.27, 3.20]†	[1.5, 3]††
	Spine	56	N/A°	N/A°	N/A°	N/A°	N/A°	[1.17, 3.20]‡	[1, 5]‡‡

Table 3: Relevant image parameters for the Tomo planning CT datasets. Entries with brackets indicate the minimum and maximum values.
pts = number of patients; SO = Sensation Open; B64 = Biograph 64; BBB = Brilliance Big Bore; kVp peak kilovoltage; res = resolution
°: These values were unavailable from the DICOM headers.
*: 60/100 had 1.95mm, 12/100 had 2.15mm, and 12/100 had 2.23mm; **: 95/100 had 3mm, 4/100 had 1.4mm
†: 35/100 had 1.95mm, 14/100 had 1.80mm; ††: 98/100 had 3mm
‡: 33/56 had 1.95mm; ‡‡: 31/56 had 1.5mm, 22/56 had 3mm;

2.4: Generation of image pairs

MIM Software (v6.1; Cleveland, OH) is a commercially-available software package that was used to archive, fuse, register, and save all image pairs. All TBeam DICOM files were directly imported to MIM from the local PC. Tomo image pairs were imported into MIM from a network server used as the export destination for the TomoTherapy TPS. All MIM-related work was performed on a local PC.

Right patient (RP) image pairs refer to a planning kVCT and setup image that come from the same patient. TBeam RP image pairs were fused using the exported registration file for all three anatomical sites. Lung images were manually adjusted to simulate a therapist alignment for a spine treatment. As no registration file was available for the Tomo dataset, RP image pairs were manually registered using cross-sectional views in the coronal, axial, and longitudinal planes. In order to best simulate a realistic image fusion prior to treatment, a generalized set of rules was devised for each anatomical site based on the guidance of several therapists and oncologists at our institution:

1) H&N: spinal column, upper cervical vertebrae (C1/C2), base of skull
2) Pelvis: pubic symphysis, pelvic circle
3) Spine: vertebral bodies

In cases when all target sites in the above guidelines could not be accurately aligned (e.g. due to patient deformation), an average fusion between the mismatching sites was performed as per therapist instruction.

Wrong patient (WP) image pairs were manually aligned as per the instructions above for both imaging modalities. The selection of patient matches was generated using a random sampling script written in MATLAB (R2014a; Mathworks; Natick, MA). Only unique WP image pairs were generated – i.e. if patient A's planning kVCT and patient B's setup image were paired together, then patient B's planning kVCT with patient A's setup image was not permitted as a comparison. Two WP image pairs were generated for each unique planning kVCT, resulting in *N* total RP matches and *2N* total WP matches.

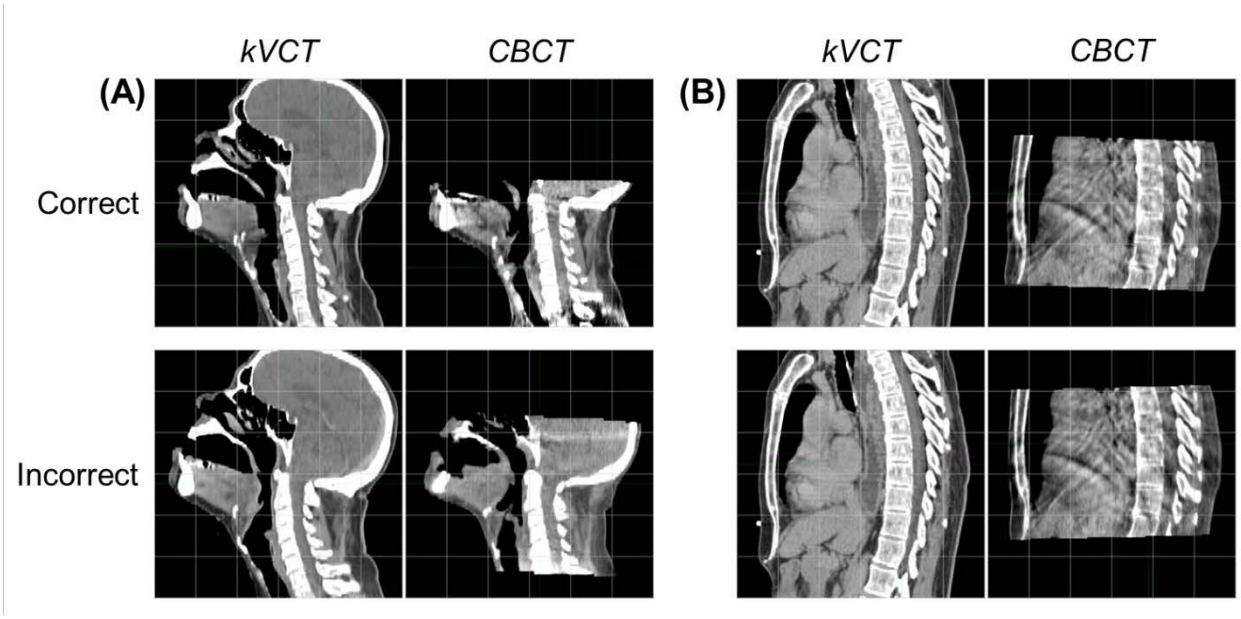


Figure 8: Example image pairs. (A) shows RP/WP matches using TBeam H&N images; (B) shows a vertebral misalignment using TBeam spine images.

Spine vertebral misalignments were also performed in MIM. The RP image pair was uploaded for each patient, and the setup image was realigned by one vertebral body superiorly and inferiorly to the correct position. H&N and pelvis misalignments were simulated in MATLAB by translating the setup image of an RP pair away from the correctly registered alignment. The setup

image was shifted in 10mm increments ranging from -50mm to 50mm in all six anatomical directions (i.e. anterior-posterior (AP), left-right (LR), and superior-inferior (SI)). Example side-by-side image pairs of RP/WP and vertebral misalignments are shown in **Figure 8**.

After all image fusions, the setup image was resaved with the same longitudinal resolution and the same number of axial slices as the planning kVCT. A body contour of the setup image was also saved and exported as a single DICOM format (to be explained further in section 2.5). Custom workflows were developed in MIM to expedite the process of opening/fusing/saving images as well as automatically generating and saving the contour. **Figure 9** shows an example workflow for a Tomo RP match. In step 1, the appropriate image pair is uploaded. In step 2, the workflow pauses to allow the user to manually adjust the images as necessary. Steps 3 and 4 create and save the fused secondary image as a separate DICOM file. Step 5 selects the fused secondary image for contouring, step 6 creates a body contour using a built-in MIM function, and step 7 saves the contour file.

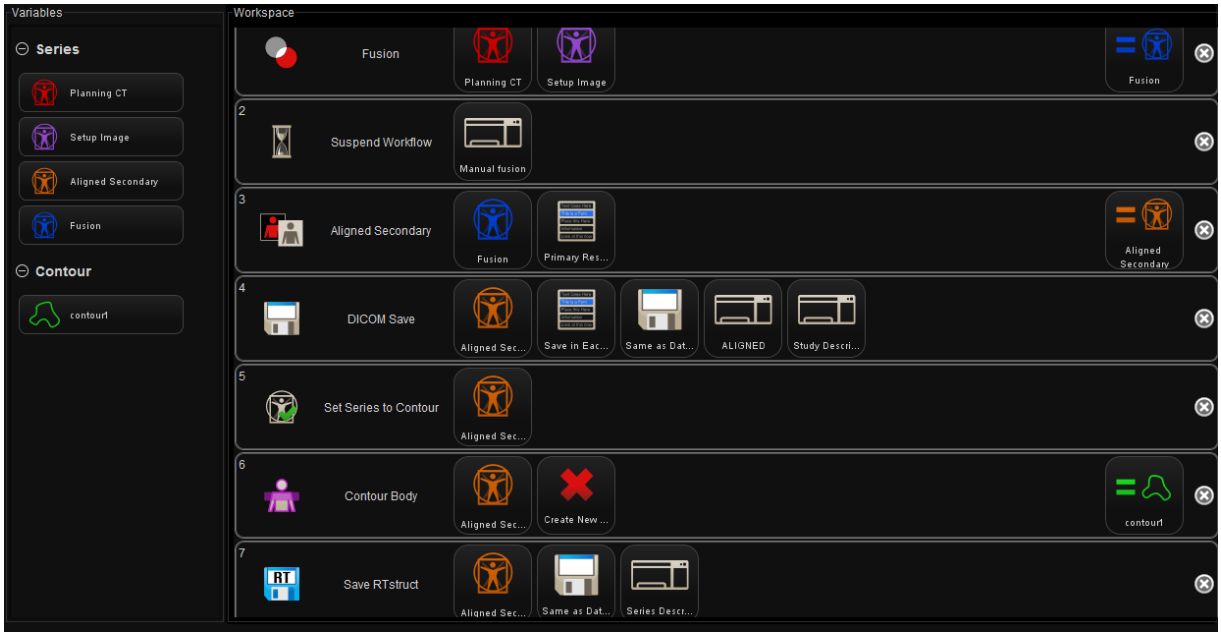


Figure 9: Graphic of the MIM workflow used to load and fuse images pairs, save the aligned setup image, and generate and save a body contour of the setup image.

2.5: Algorithm development: image pre-processing

The algorithmic workflow was developed by custom software written in the MATLAB environment. The PC used was equipped with an Intel Core i5-650 3.20GHz quad-core processor and 8GB RAM, and used the Windows 7 operating system. Initially, the registered image pair was used as the only input to the algorithmic workflow. The DICOM header of each image was also loaded and their axial resolutions were compared. The setup image was downsampled to the resolution of the planning kVCT if the former had a finer resolution; this step was to place both images on the same coordinate system and allow for subsequent image similarity comparisons.

In our first approach, we generated an initial mask for the image comparison by extracting the real-space coordinates of the setup image from the DICOM headers. A rectangular region of interest (ROI) was generated around these coordinates, and this bounding box was transformed into a binary mask inside which image comparison would take place. Upon further analysis of an initial image similarity assessment (see sections 2.6.1, 2.6.2), we found a large amount of spatial mismatch occurring between the couches of the setup and planning kVCT images (**Figure 10**). Different couches are often used for the simulation and treatment phases in RT, which may introduce inherent differences in couch size, material, and attenuation properties. These discrepancies degrade the accuracy of an RP match, which subsequently reduce the ability for accurate discrimination between correct and incorrect image pairs. We also found spatial mismatches at the edges of the bounding box due to interpolation errors and image noise, especially when using images pre-processed with a gradient operator (explained further in section 2.6).

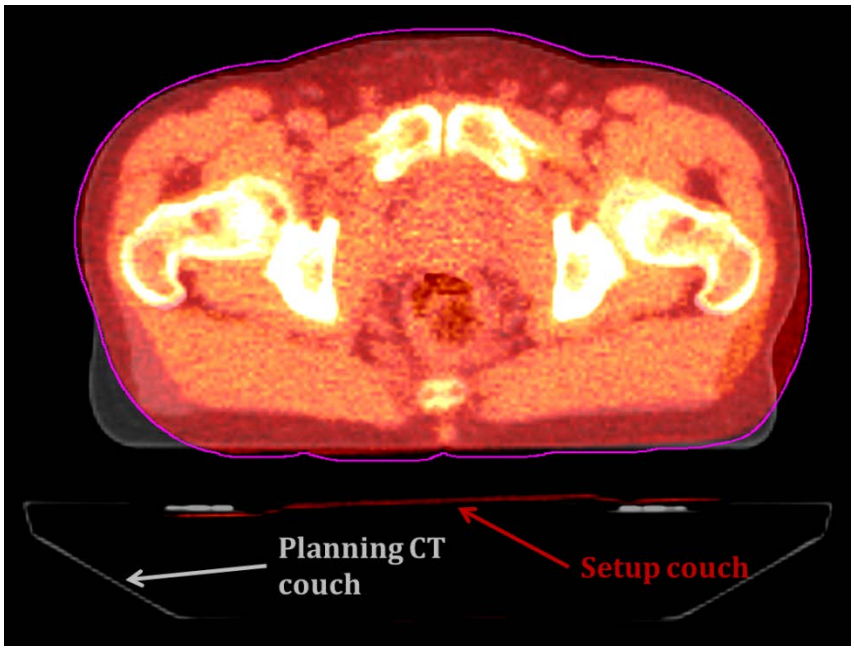


Figure 10: Different-sized couches used in the setup image (red arrow) and planning CT image (light gray arrow) degrade the strength of a same-patient comparison. A body contour, shown as the pink outline, circumvents this issue.

A Tomo setup image is shown in this example.

To correct for this issue, we investigated the use of a body contour (**Figure 10**) as a more accurate initial mask for image comparison. MIM's built-in 'Body Contour' function was used to generate a single DICOM file containing the real-space coordinates of the setup image's body contour. The open-source CERR software was used to convert these files into MATLAB-friendly binary masks [82]. In short, the program first uses the setup image and its corresponding mask (both in DICOM format) to create a mask structure in CERR format. A 3D mask can then be extracted from this structure. One unique mask corresponding to each setup image in the RP/WP/vertebral shift scenarios was generated and used alongside the image pair as inputs to the algorithmic workflow.

Another key element to the algorithmic design was the thresholding of voxels with pre-specified Hounsfield units (HUs). The HU scale represents a linear transformation of a linear attenuation coefficient measurement relative to water. After CT reconstruction, each voxel is normalized and truncated to integer values using the following expression [83]:

$$HU = 1{,}000 * \frac{\mu - \mu_{water}}{\mu_{water}} \quad (1)$$

μ and μ_{water} represent the linear attenuation coefficient of the voxel and of water, respectively. HU values can range from approximately -1000 (air) to 3000 (dense bone) [83]. We chose to threshold out air voxels (HU<-700) with the intention of removing any mismatching air pockets between RP and WP image pairs. Relevant examples in the studied anatomical sites include bowel gas, air cavities in the H&N region, and lung motion. **Figure 11** illustrates an example of mismatching bowel gas pockets that, when unaccounted for, significantly reduced the strength of an RP match.

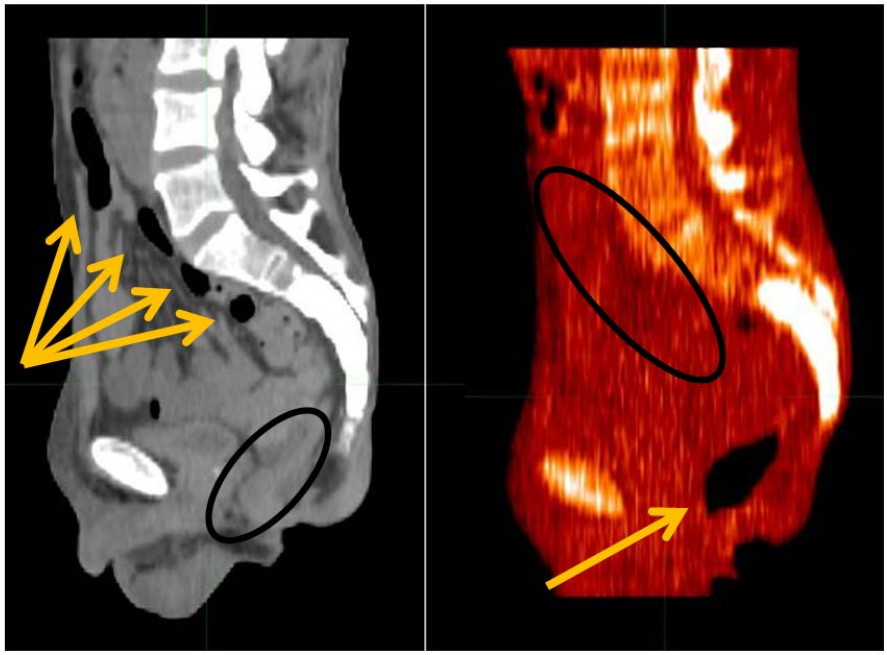


Figure 11: Example of variable bowel gas presence in an RP Tomo image pair. Yellow arrows and black ovals indicate mismatching bowel gas locations between the two images.

In addition to thresholding out air voxels, we also eliminated high density voxels corresponding to metallic implants commonly found in RT patients (**Figure 12**). Common examples include dental, hip/femur, and sternum implants. Due to their high attenuation, metallic objects often cause streaking artifacts as a result of beam hardening, scatter, and edge effects [84]. We empirically determined an absolute threshold of HU>2200 after examining several images with said artifacts.

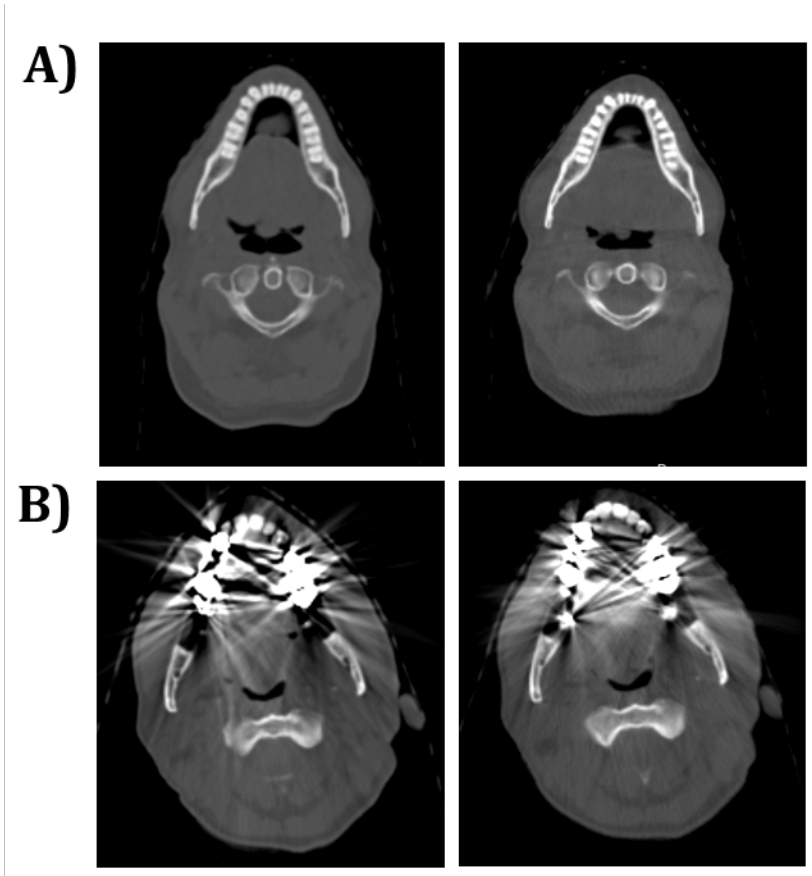


Figure 12: Examples of artifacts from metallic objects. (A) shows a planning kVCT (left) and registered CBCT (right) without any dental implants. (B) shows a registered planning CT / CBCT pair with the presence of dental implants, where streaking artifacts are readily apparent.

For Tomo images, we also implemented a voxel intensity correction due to inherent CT number differences between kVCT and MVCT images. CT numbers for these imaging energies match well for soft tissue, but begin to differ for tissues with increasing atomic number (e.g. bone). kVCT numbers are larger than MVCT numbers due to the higher occurrence of the photoelectric effect, which is proportional to Z^3/E^3; photons with MV energy interact primarily via Compton scattering, which is roughly independent of the material's atomic number.

To address this discrepancy, data was used from a study looking at deformable registration of kilovoltage planning kVCT images and daily MVCT setup images for the abdominal region [85]. Dr. Deshan Yang from Washington University in St. Louis provided additional information about their experimental procedure as described below. Six MVCT and kVCT image pairs were used to quantify their CT number relationship. The kVCT image was registered to the MVCT using

deformable image registration. **Figure 13a** shows the CT numbers of 5% of the total voxel pairs randomly selected from one image pair; with increasing tissue density, one can see the higher CT numbers of the kVCT image relative to the MVCT image. Phantom experiments were also performed to measure CT numbers for both kVCT and MVCT systems, showing similar results (**Figure 13b**).

A third-order polynomial [86] was subsequently fit to the points from the image pair comparison study (coefficient of determination: $R^2 = 0.9973$):

$$\frac{I_{MV}}{1000} = 0.1023 * \left(\frac{I_{kV}}{1000}\right)^3 - 0.5701 * \left(\frac{I_{kV}}{1000}\right)^2 + 1.4664 * \frac{I_{kV}}{1000} + 0.0003 \qquad (2)$$

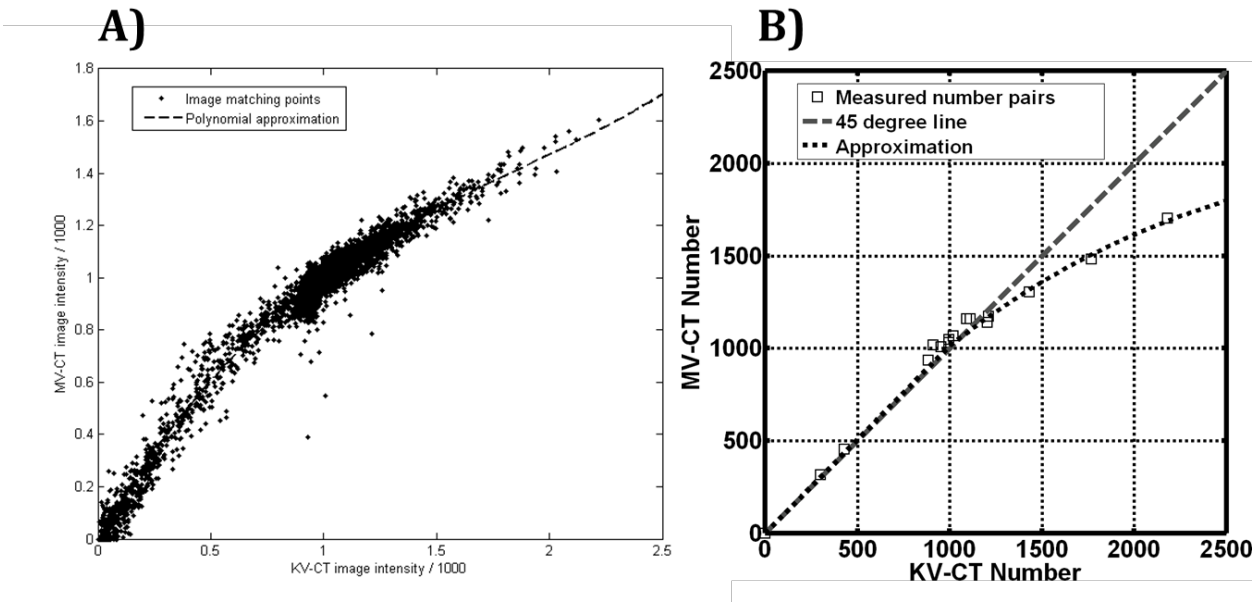


Figure 13: A) shows a plot with the kVCT/MVCT CT numbers of 5% of the total measured voxel pairs on a single image pair. The dashed line represents a third-order polynomial fit to all the voxel pairs. Figure B) shows a similar plot from a validation phantom experiment measuring CT numbers on kVCT and MVCT units. (both images courtesy of Dr. Deshan Yang)

After the aforementioned preprocessing steps, RP/WP/vertebral shift image pairs were compared using metrics described in the following section. As mentioned earlier, H&N and pelvis shifts were simulated in MATLAB by first loading an RP image pair. The setup image and the corresponding body mask were then translated in 10mm real-space increments (up to 50mm) in all

six anatomical directions. One set of image similarity comparisons was generated for each translational shift.

2.6: Algorithm development: similarity metrics

Many similarity metrics have been developed and used in the literature for both medical and non-medical image comparison. We utilized three commonly-used metrics as well as two custom-designed metrics, explained in the following sub-sections.

2.6.1: Correlation coefficient

Pearson's correlation coefficient (CC) is a global measure of the statistical linear correlation or dependence between two variables [87]. For a population, it is calculated as the covariance of both variables divided by the product of their individual standard deviations. For a sample, estimates of the above parameters are used in the following formula:

$$CC = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad \textbf{(3)}$$

X_i and Y_i are the intensity values for all voxels n for the masked setup and planning kVCT images, and $\bar{X}$ and $\bar{Y}$ are the sample means of each image. The value of CC can range from -1 to 1, where 1 represents total positive correlation between the variables (i.e. a perfect image match), 0 indicates no correlation (i.e. images are completely different), and -1 represents total negative correlation. CC has been found to be useful in assessing image similarity [88, 89] as well as a metric for image registration [90-93].

2.6.2: Mutual information

Mutual information (MI) is a measure of how much one random variable tells us about another. It can be calculated by the following expression:

$$MI = \sum_{x,y} P_{XY}(x,y) * \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \quad (4)$$

P_{XY} is the joint probability distribution of variables X and Y, while P_X and P_Y are the individual probability distributions. MI is closely related to the Kullback-Leibler divergence, a measure of the distance between two distributions [94, 95]. MI can be computationally expensive, but has been frequently used as a metric of image similarity for image registration, particularly multimodal registration [96-105].

2.6.3: Structural similarity

The structural similarity (SSIM) metric is a recently developed measure of the similarity between two images [106]. The underlying premise behind its design is that it attempts to better mimic the human visual system (HVS) for image quality assessment. The HVS is highly adapted to extract structural information, where pixels have strong interdependencies if they are spatially close. As such, local patterns of pixel intensities can be compared in varying window sizes for two input variables as a metric for assessing similarity. SSIM is calculated from the following formula on local square window sizes:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (5)$$

μ_x and μ_y are the averages of variables x and y, respectively; σ_x and σ_y are the standard deviations of x and y, respectively; σ_{xy} is the covariance of x and y; c_1 and c_2 are stabilizing variables. The final SSIM metric is an average of all SSIM values from each local window over the entire image. SSIM can range from -1 to 1, with 1 indicating a perfect similarity between two variables. The SSIM index has been used numerous times to assess image quality and similarity [107-112].

2.6.4: Point-to-ROI approach: rationale

During the setup stage of the RT process, a patient may not be positioned in precisely the same position as they were during CT simulation. Minor patient deformations – both externally and internally – may create systematic mismatches between the setup image and planning kVCT, which could degrade similarity assessment between an RP image pair.

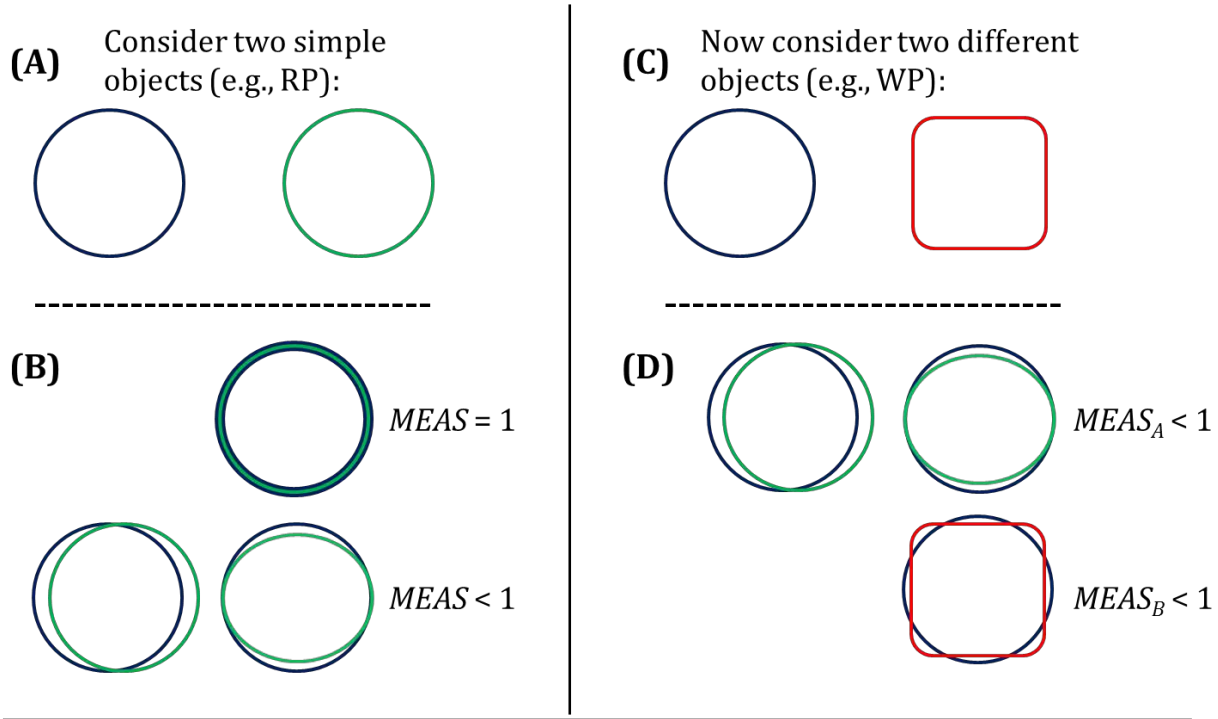


Figure 14: Simple example of spatial mismatches. (A) two simple objects that are the same (e.g. RP match); (B) an exact alignment of these objects will result in a perfect similarity measure while any minor translational or rotational offsets will produce a similarity measure less than 1; (C) two similar, but different objects (e.g. WP match); (D) the measure between these dissimilar objects and the spatially offset objects in (B) could potentially be difficult to compare for identification purposes.

A simple example of this is illustrated in **Figure 14**. Consider two identical circles (e.g. a simulation of an RP) as seen in **14A**. Should a perfect overlap exist between these two objects (**14B**), an image similarity metric (*MEAS*) would theoretically calculate a value of 1. If there are any minor deformations (shown in **14B** as a minor translation or rotation), the similarity measure would result in a value less than 1. In **14C**, consider two different objects simulating a WP image pair. For some situations (**14D**), it may be unclear how the similarity metric assessing two of the

same objects with a spatial mismatch ($MEAS_A$) would compare with two different objects entirely ($MEAS_B$). This introduces a potential source of error if any spatial offsets exist in the final fusion of the setup and planning kVCT image pair.

Our goal was to construct a metric that would account for the presence of potential spatial mismatches in a meaningful way. Our conceptual design was inspired by the gamma index, a routinely-used tool used in RT for IMRT QA [113, 114]. In short, this tool was designed to account for large differences that can occur when comparing two dose distributions that have relatively small misalignments. A search is conducted for each point in the reference dose distribution that includes normalized parameters for both dose and distance (**Figure 15**). The gamma index calculates the minimum distance between both distributions through both distance and dose parameters, overcoming any poor comparisons due to shallow or steep dose gradients in the dose distributions.

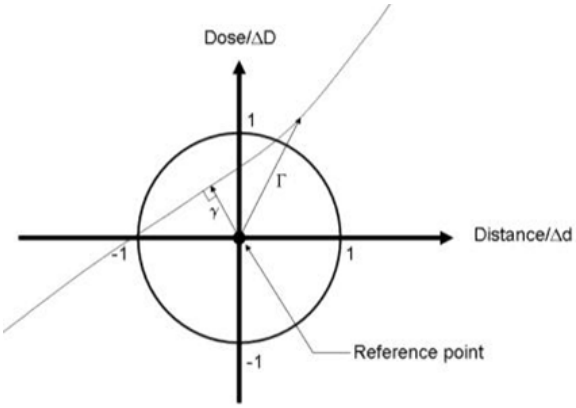


Figure 15: Graphic of the gamma index. For each reference point, the gamma γ is calculated against the evaluated distribution (gray line), and the circle represents the 'pass' criteria. (Figure taken from [114])

We developed a metric loosely based off this approach. To focus the comparison between boundaries of anatomical regions (i.e. areas where deformations would be most readily detectable), images were first pre-processed using a two-step approach. First, a bilateral filter was applied to the setup and planning kVCT images [115], which replaces the intensity value of each pixel by a weighted average of its surrounding pixels. The filter has two kernels – one is a spatial weight, which depends on Euclidean distance, and the other is a range kernel, which depends on intensity difference. Together, these kernels allow for image denoising and smoothing (via the spatial kernel)

while simultaneously maintaining edge preservation (via the range kernel). Both kernels are typically based off a Gaussian distribution. The filter is calculated as follows:

$$I^{filtered}(x) = \frac{1}{k}\sum_{x_i \in \Omega} I(x_i) * f_r(\|I(x_i) - I(x)\|) * g_s(\|x_i - x\|) \quad \textbf{(6)}$$

$I^{filtered}$ is the resultant filtered image as a function of pixel location x, I is the input image, f_r is the range kernel for intensity smoothing, g_s is the spatial kernel, Ω is the window centered in x, and k is a normalizing term for energy preservation: $k = \sum_{x_i \in \Omega} f_r(\|I(x_i) - I(x)\|) * g_s(\|x_i - x\|)$. For our parameter selection, we used spatial domain standard deviations in the axial plane and longitudinal direction of 3mm and 1mm, respectively, to approximate the image resolution. The intensity domain standard deviation was empirically chosen to be 10 grayscale units.

The second step was to apply a Sobel gradient operator [116], which creates an image that emphasizes edges and transitions between different boundaries. It is a discrete differentiation operator that calculates an approximation of the image intensity gradient. The input image is convolved with an integer-valued filter in the x, y, and z axes. The following is an example of the Sobel operator in the z-axis:

$$G_z'(:,:,1) = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix};\ G_z'(:,:,0) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix};\ G_z'(:,:,-1) = \begin{bmatrix} -1 & -2 & -1 \\ -2 & -4 & -2 \\ -1 & -2 & -1 \end{bmatrix} \quad \textbf{(7)}$$

The final value of each voxel can be computed by taking the gradient magnitude in each direction: $G = \sqrt{G_x^2 + G_y^2 + G_z^2}$. In order to compare only large edges and object boundaries in the image, we chose to threshold out smaller gradients that would be indicative of smaller intensity variations within some anatomical region (e.g. soft tissue contrast) that would not provide discriminatory power between correct and incorrect image pairs. We used an absolute threshold of 5,000 intensity units.

2.6.5: Point-to-ROI approach: description

The general approach of the point-to-ROI approach (hereby referred to as 'gradient-based') is outlined in **Figure 16**. Each voxel in the pre-processed planning kVCT was compared to a volumetric ROI (5x5x5) centered on the corresponding voxel in the setup image. Two image similarity metrics were developed from this point-to-ROI method of comparison. In the first, the intensity difference was calculated between voxel P_i on the planning kVCT and each voxel P_j in the corresponding ROI on the setup image (**Figure 16a**). The minimum value of these N differences was recorded in a separate vector D. This was repeated for all voxels Z on the pre-processed kVCT image (**Figure 16b** and **c**):

$$D(i) = \min_{j \in [1,\dots N]} |P_i - P_j|, \quad i = 1, \dots Z \qquad (8)$$

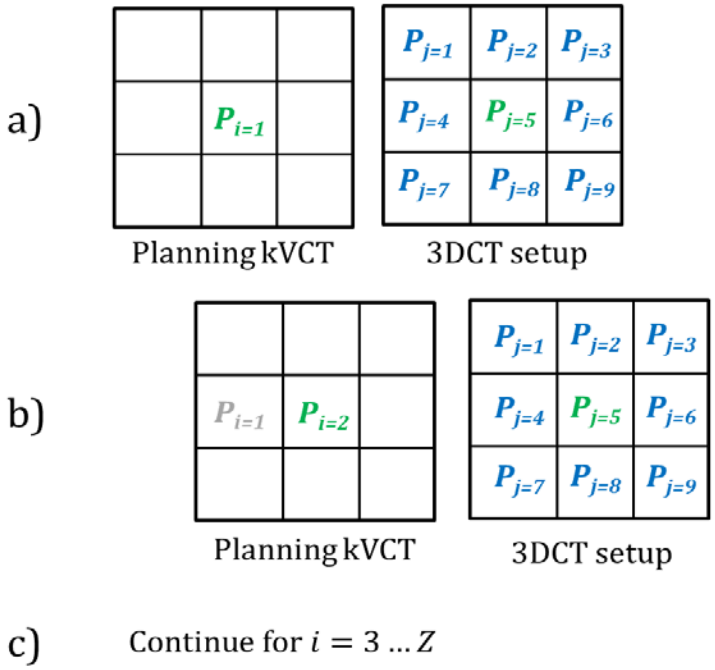


Figure 16: General description of the gradient-based metric. In a), voxel P ($i=1$) is compared to a corresponding ROI centered around the corresponding voxel (both shown as green). Some metric is calculated. In b), the same calculation occurs between planning CT voxel $i=2$ and its new corresponding ROI centered about the new corresponding voxel (shown as green). This continues for all voxels in the planning CT (c).

Since gradient norms ignore any directional component, a tissue interface with a large gradient (e.g. soft tissue-bone) would be treated equivalently to an interface with the same tissue types on opposite ends of the interface (e.g. comparing soft tissue-bone to bone-soft tissue). A small difference could be found between these two interfaces, but would falsely represent a

correct match. As such, for each P_i, the voxel location of P_j corresponding to the minimum intensity difference was saved. The intensity difference of these corresponding voxels in the unprocessed planning kVCT and setup image was also recorded.

In order to extract additional spatial information from the comparison, each image was decomposed into its gradient vector components $\vec{v}_x$, $\vec{v}_y$, and $\vec{v}_z$. The second similarity metric calculated the dot product between each voxel *i* of the planning kVCT and voxels *j* in the corresponding ROI of the setup image. The maximum of these values was recorded into a separate vector *G*:

$$G(i) = \max_{j\in[1,\dots N]}\left(v_{x,i} * v_{x,j} + v_{y,i} * v_{y,j} + v_{z,i} * v_{z,j}\right), \;\; i = 1, \dots Z \qquad (9)$$

2.7: Model outputs

The CC, MI, and SSIM metrics produced a single value for each image pair comparison. The gradient-based metrics produced a vector output when using the gradient-based metrics – *Zx2* for the intensity different metric and *Zx1* for the gradient dot product metric, where *Z* refers to the number of voxels in the pre-processed planning kVCT image. Example vector histograms can be seen in **Figure 17**, where the RP dot product and intensity difference profiles are overall greater and smaller than the WP profiles, respectively. In order to consolidate these vector histograms into a smaller set of values for the classification step, a set of commonly-used descriptors were extracted from each vector histogram. Separate metrics were calculated for each column of the *Zx2* histogram from the intensity difference metric. The following descriptors were generated from the histogram of each image pair: mean, max, 5th to 95th percentiles in 5% increments, 3rd moment, and 4th moment.

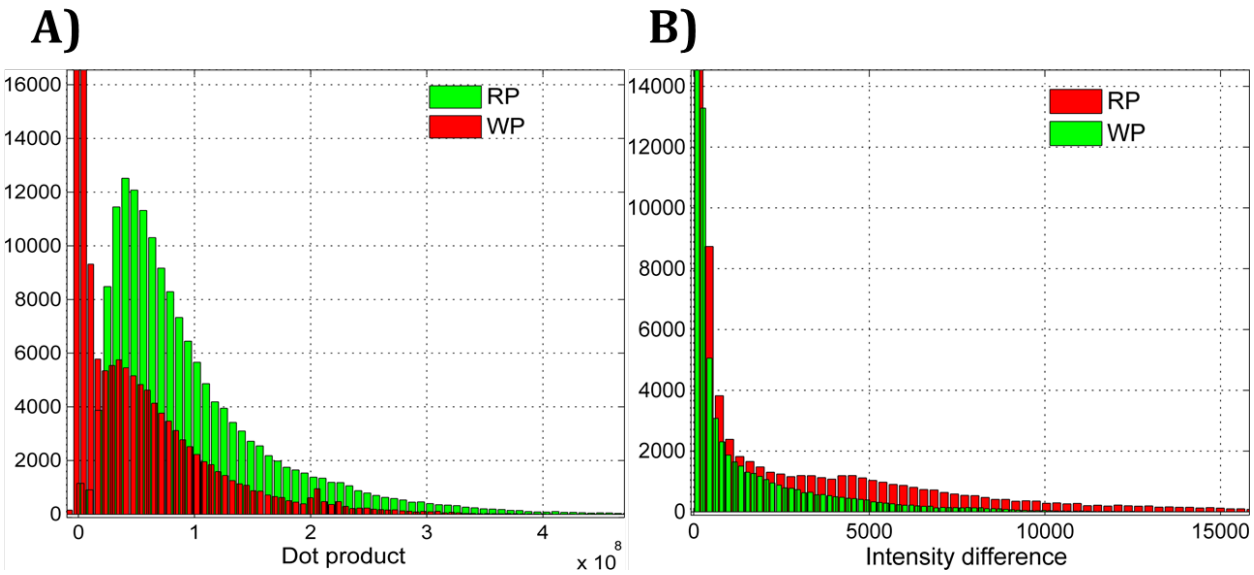


Figure 17: Example output histograms from the gradient-based metrics. The plot axes are truncated to better illustrate differences between RP and WP histograms. A) shows a dot product histogram of a randomly-selected TBeam pelvis patient (truncated WP bars reach up to 3.3e4). B) shows an intensity difference histogram of a randomly-selected TBeam H&N patient (truncated RP and WP bars reach 8.7e4 and 6.6e4, respectively).

2.8: Feature selection

An additional feature selection step was implemented in preparation for image classification. Problems can arise in the performance of recognition problems when using input data that exist in a high dimensional space [117]. Principal components analysis (PCA) is a statistical technique used to transform a set of variables into a set of linearly uncorrelated values called the principal components. Through this linear transformation, the dimensionality of the input data can be reduced while retaining a pre-specified proportion of the variation present in the original data. PCA has been frequently used in the medical imaging literature for both dimensionality reduction and classification [118-129].

A brief description of the PCA mathematics is described in the following paragraphs. First, the initial training set $\vec{x}_i, i = 1, \ldots M$ is gathered, with M representing the total number of images.

Each image is of size *N*. Each image is then standardized by subtracting the mean and dividing by the standard deviation:

$$\vec{\omega}_i = \frac{\vec{x}_i - \bar{x}_i}{\sigma_i} \quad \textbf{(10)}$$

The covariance matrix is then computed:

$$C = AA^T \text{ (where } A = [\vec{\omega}_1\ \vec{\omega}_2\ \dots\ \vec{\omega}_M]) \quad \textbf{(11)}$$

As an *NxN* matrix, solving this equation for the eigenvectors (u_i) and eigenvalues (λ_i) is computationally expensive. As there are far less data points in the image space than the dimension of the image space (i.e. $M << N^2$), there exist only *M-1* meaningful eigenvectors. As such, the eigenvectors ($\hat{u}_i$) and eigenvalues ($\hat{\lambda}_i$) can be solved for an *MxM* matrix $D = A^T A$. It has been shown that the eigenvectors ($\vec{u}_i$) and eigenvalues (λ_i) of the covariance matrix *C* can then be calculated through the following expressions [130]:

$$\vec{u}_i = A\hat{u}_i \text{ and } \lambda_i = \hat{\lambda}_i \quad \textbf{(12)}$$

In the dimensionality reduction step, the minimum number of eigenvalues that describe some minimum threshold of the data's variance will be computed to determine the highest *K* eigenvectors used for the final dataset:

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{M} \lambda_i} > threshold \quad \textbf{(13)}$$

The principal components can now be calculated:

$$\vec{z}_j = \begin{bmatrix} \vec{u}_1^T \\ \vdots \\ \vec{u}_K^T \end{bmatrix} * \vec{\omega}_i, j = 1, \dots K \quad \textbf{(14)}$$

A threshold of 0.95 was used to select the highest *K* eigenvectors. PCA was performed on the entire set of gradient-based vectors, and the CC/MI/SSIM metrics were then added to comprise the final set of possible input values for classification. **Figure 18** describes the complete experimental workflow developed thus far.

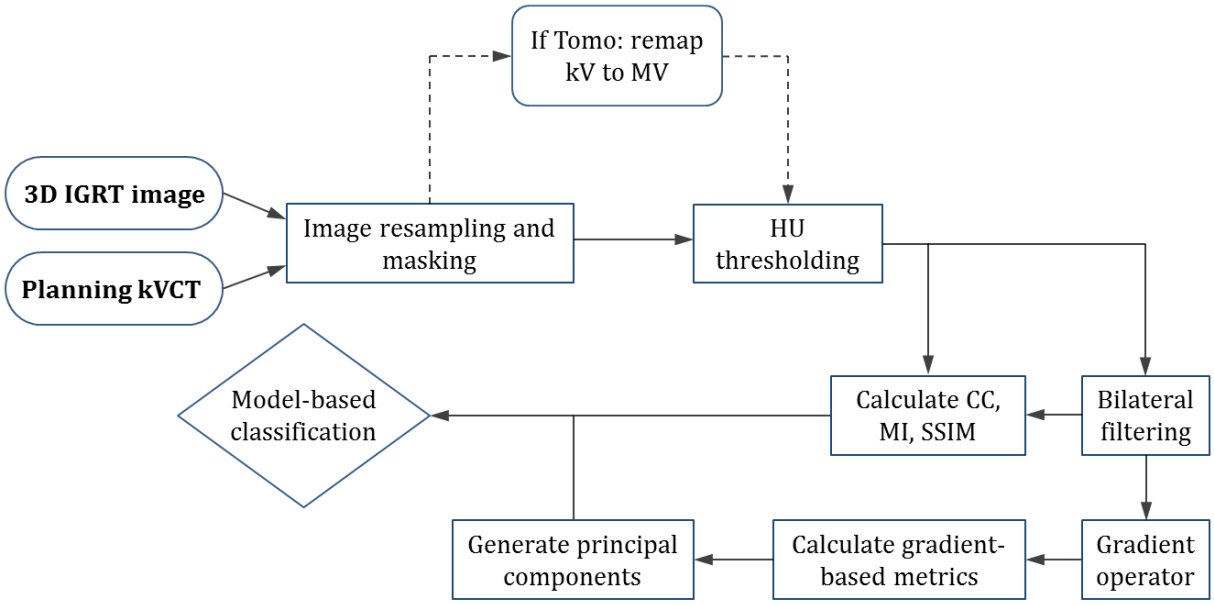


Figure 18: Experimental workflow prior to classification.

CHAPTER 3: CLASSIFICATION AND MODEL EVALUATION

3.1: The task of classification

In the fields of statistics and machine learning, classification is used to address the problem of making some decision on the basis of available information. This is typically performed through the use of a classification model, which uses a set of known observations to build a set of criteria that will label any unknown observation into a predefined category. The known observations in this study include the CC, MI, SSIM, and PCA-reduced gradient-based metrics, which are the potential *features* used for the classification model. The predefined categories, or *classes*, are 'right' and 'wrong' and refer to either a same-patient image pair ('right') or a wrong patient / misaligned patient image pair ('wrong').

For a real-world application, Michie *et al* discuss many issues of concern for a potential classifier [131]. Accuracy describes the reliability of the classifier and is represented as the proportion of correctly classified observations. The *training* dataset is used to construct the model parameters, and a *test* dataset is then used to generate the model accuracy. As some errors may carry more weight than others, special efforts may be implemented to control the accuracy for these errors. The speed of an algorithm is essential to be practically implemented in a real-world setting – for example, a classifier with 85% accuracy could be preferable over one with 90% accuracy if its implementation is significantly faster. The learning time can also be important depending on the classifier used, as new rules may need to be quickly learned or existing rules should be able to be rapidly adjusted. Understandability or comprehensibility of a classifier is important when it requires a human operator to apply the procedure. In addition, the operator must also believe in the system's robustness; in the partial nuclear meltdown at Three Mile Island, the computer system's recommendation for shutdown was overridden by a human operator who did not believe that recommendation was well-founded [131, 132].

Many classification algorithms have been used for medical image analysis. The following sections describe the specific classifiers utilized for the present study.

3.2: k-Nearest Neighbors

k-Nearest Neighbors (KNN) is a non-parametric algorithm used for learning and classification [133]. Non-parametric indicates that the algorithm makes no assumptions about the underlying probability distribution of the data being assessed. This is useful when the data distribution is unknown, which is generally the case in real-world examples, or does not visually match a known theoretical distribution. KNN is also a "lazy" learning algorithm, meaning that it has a minimal training phase and defers most of the calculation for the testing phase. As such, the final classification decision is made on the entire training dataset. Although KNN is considered to be one of the simplest machine learning algorithms, it is easy to understand and implement while performing well in many situations [134]. Some examples of its use include cancer cell nuclei classification in microscopic images [135], pattern classification in mammographic images [136], and texture-based classification of atherosclerotic carotid plaques in ultrasound images [137].

Assume that we are given a training set D comprised of objects ($\mathbf{x}$, c), where $\mathbf{x}$ is a set of Nx1 data objects and c is the known class label of each object. Given some test object T = ($\mathbf{x}$', c') (where $\mathbf{x}$' is the data of the test object, and c' is the unknown class of that object), the distance is computed between T and all objects in the training set D. The closest k neighbors in D are then segregated into a separate list D_T. The test object is then classified by the majority class vote of those k objects:

$$c' = \underset{c}{\operatorname{argmax}} \sum_{(\mathbf{x}_k, c_k) \in D_T} I(c = c_k) \qquad \textbf{(15)}$$

c is a class label, ($\mathbf{x}_k$, c_k) is the data and class label of the k^{th} nearest neighbors, and I is some indicator function that returns 1 if true, 0 if false [134]. To provide a simple visual example of the algorithm's output, consider the commonly-used Iris dataset [138]. The data is comprised of three

varieties of the iris flower: setosa, versicolor, and virginica. The length and width of the petals and sepals were measured for 50 flowers of each type. A scatterplot using the petal length and width is shown in **Figure 19**.

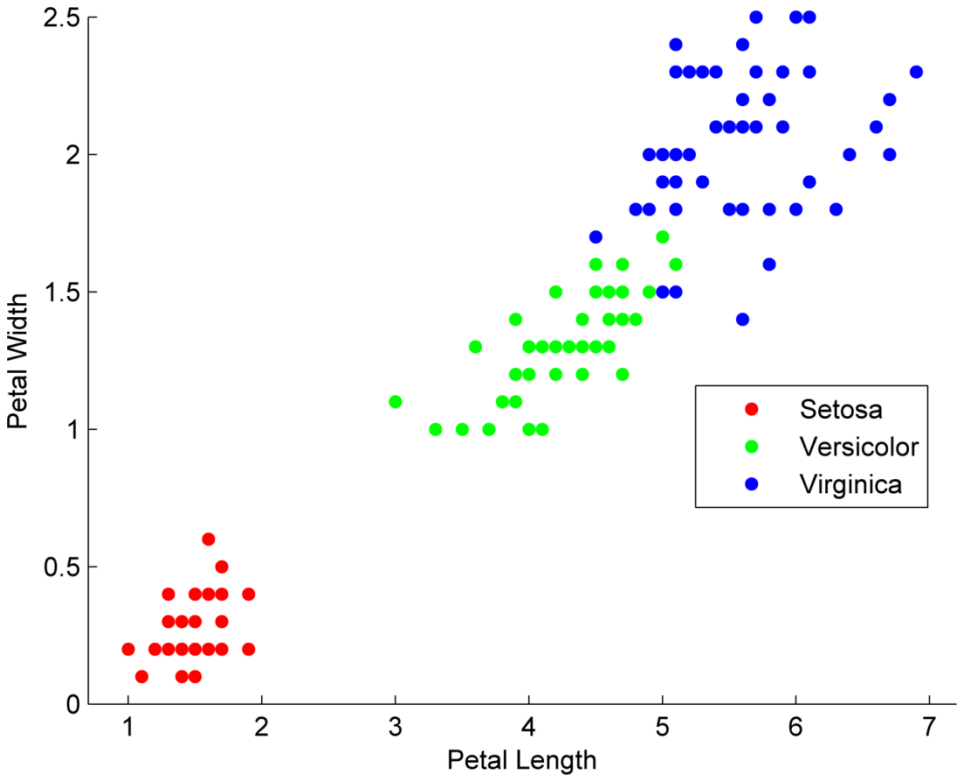


Figure 19: The commonly-used iris dataset. Shown is a scatterplot of the three flower varieties using the petal width and length measurements [cm].

In **Figure 20**, an unknown point (shown as the black **X**) is classified by finding the majority class of its nearest *k* neighbors (in this case, *k*=5). The point is determined to be the versicolor type, given that 4 of the 5 neighbors are verisicolor.

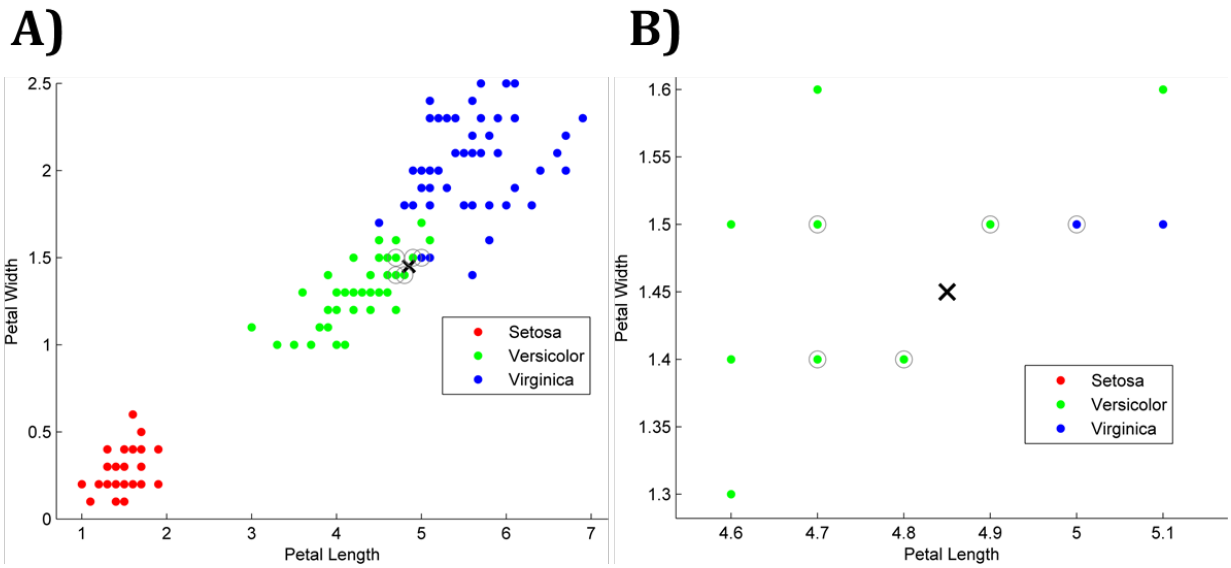


Figure 20: Example of KNN classification on the iris dataset. A) shows an unlabeled point X and the 5 nearest neighbors. B) shows a zoomed version to better visualize the nearest neighbors. The petal length and width are measured in [cm].

There are several important considerations for the implementation of KNN. First and foremost is the value of *k*. Choosing a *k* that is too small introduces a potential sensitivity to noise and may result in incorrect classifications. On the other hand, too large of a *k* may result in the inclusion of too many points from other classes, degrading the overall classification accuracy. Oftentimes, *k* is selected as an odd number for a binary classification problem in order to avoid ties in the voting process. The approach of determining the test point's class label is another consideration. The simplest method is to take a majority vote of the nearest neighbors as described above. If closer neighbors more reliably indicate the class of the unknown test object, a weighting factor – typically the inverse or squared inverse of the distance metric – can be assigned to each vote to improve classification. The choice of distance measure can also be important, as the most desirable measure will be one where smaller distances indicate a greater probability of having the same class [134].

3.3: Discriminant analysis

Discriminant analysis is a method that aims to find a linear combination of features that best separate two classes of observations [139]. The separation can be a line for 2D observations, a plane for 3D observations, or a hyperplane for *N* dimensional observations. It has found many uses in the imaging literature, ranging from facial recognition and image classification to tumor detection and brain activity classification [61, 140-145].

Two types of discriminant analysis classifiers were used in this study: linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). Generally, discriminant analysis relies on the calculation of class-conditional probabilities using Bayes' theorem [146]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \textbf{(16)}$$

A and B are certain independent events, $P(A)$ and $P(B)$ are the probabilities of A and B, $P(A|B)$ is the probability of A given the occurrence of B (in other words, given that B is true), and $P(B|A)$ is the probability of B given the occurrence of A.

For discriminant analysis, assume two possible values for class C (0 or 1). Given some data x, we are interested in calculating the probability that it belongs to class 0 or 1. From Bayes' theorem, we can write the following general expression:

$$P(C = 0|x) = \frac{f_0(x)P(C = 0)}{f_0(x)P(C = 0) + f_1(x)P(C = 1)} \quad \textbf{(17)}$$

$P(C=1)$ and $P(C=0)$ are prior probabilities of the classes. As this is typically an unknown parameter in practice, it can be estimated by dividing one over the total number of available classes (uniform approach). Another method (empirical approach) is to divide the number of observations in each class C over the total number of observations in the training set.

$f_C(x)$ is the probability distribution for the data x, given that x is from class C. LDA assumes that both class distributions are normally distributed, and that the data from each class has the

same covariance matrix $\hat{\Sigma}$. Given $\hat{\mu}_C$ for each class and the joint covariance matrix $\hat{\Sigma}$, the class density functions can therefore be calculated from the normal density formula:

$$\hat{f}_C(x) = \frac{1}{2\pi\left(\left|\hat{\Sigma}\right|\right)^{1/2}} * e^{-\frac{1}{2}(x-\hat{\mu}_C)^T \hat{\Sigma}^{-1}(x-\hat{\mu}_C)}, \;\; C = [0,1] \qquad \textbf{(18)}$$

$\left|\hat{\Sigma}\right|$ is the determinant of the covariance matrix. Once these are calculated, the probability estimate of an unknown observation $\hat{x}$ can be estimated by updating equation (18) with the class density functions:

$$P(C = 0|\hat{x}) = \frac{\hat{f}_0(x)\hat{P}(C = 0)}{\hat{f}_0(x)\hat{P}(C = 0) + \hat{f}_1(x)\hat{P}(C = 1)} \qquad \textbf{(19)}$$

$\hat{P}$ refers to the prior probabilities calculated above. Bayes' rules state that the observation $\hat{x}$ should be labeled to the class with the higher posterior probability, i.e. $P(C = 0|\hat{x})$ and $P(C = 1|\hat{x})$. In other words, this is equal to maximizing the product of the prior probability and the class density function:

$$\hat{C}(\hat{x}) = \underset{C}{\operatorname{argmax}}\left[\hat{x}^T\hat{\Sigma}^{-1}\hat{\mu}_C - \frac{1}{2}\hat{\mu}_C^T\hat{\Sigma}^{-1}\hat{\mu}_C + \log(\hat{P}_C)\right] \qquad \textbf{(20)}$$

$\hat{C}$ refers to the class label of input observation $\hat{x}$. A linear boundary function can then be derived and used as the linear discriminant function. Returning to the iris dataset, linear classifiers between the neighboring classes can be created as shown in **Figure 19**. A line can perfectly separate the setosa and versicolor classes, while an imperfect boundary is drawn between the versicolor and virginica classes.

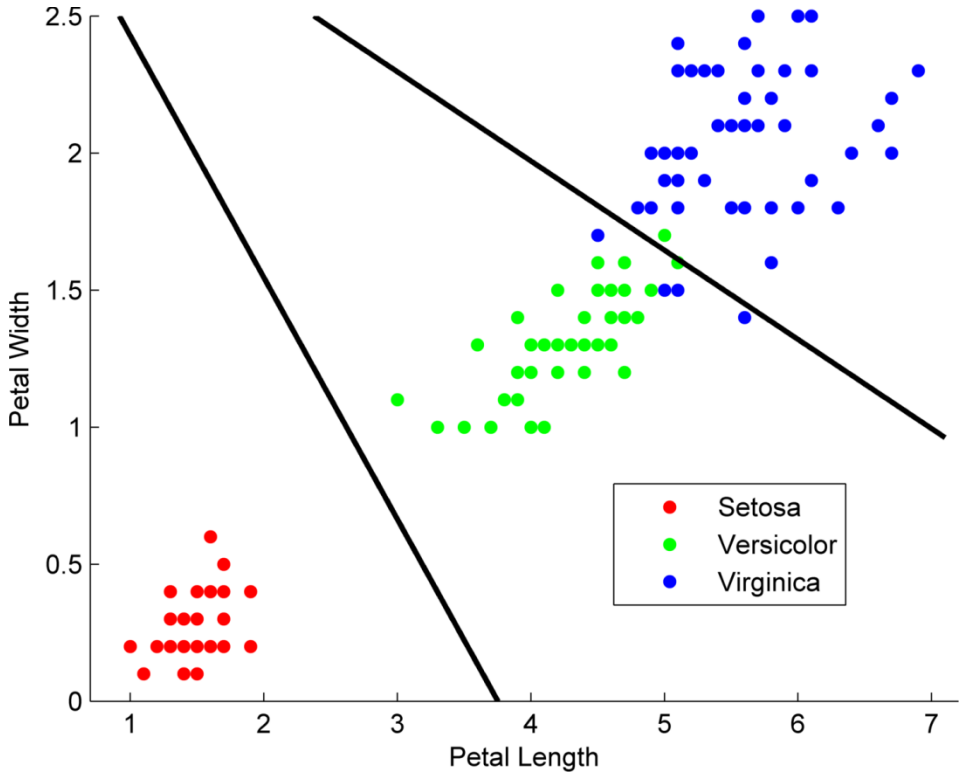


Figure 21: Example of LDA classification on the iris dataset [cm]. Two linear boundaries are shown between the setosa/versicolor and the versicolor/virginica classes.

In QDA, classes are not assumed to have the same covariance matrix $\hat{\Sigma}$ as in LDA. Rather, class has its own covariance matrix $\hat{\Sigma}_C$. The quadratic discriminant function is as follows:

$$\hat{C}(\hat{x}) = \underset{C}{\operatorname{argmax}} \left[-\frac{1}{2}\log|\hat{\Sigma}_C| - \frac{1}{2}(\hat{x} - \hat{\mu}_C)^T \hat{\Sigma}_C^{-1}(\hat{x} - \hat{\mu}_C) + \log(\hat{P}_C) \right] \quad \textbf{(21)}$$

After running QDA on the iris dataset using this quadratic function instead of the linear function, the resultant quadratic boundaries between the neighboring classes can be seen in **Figure 22**.

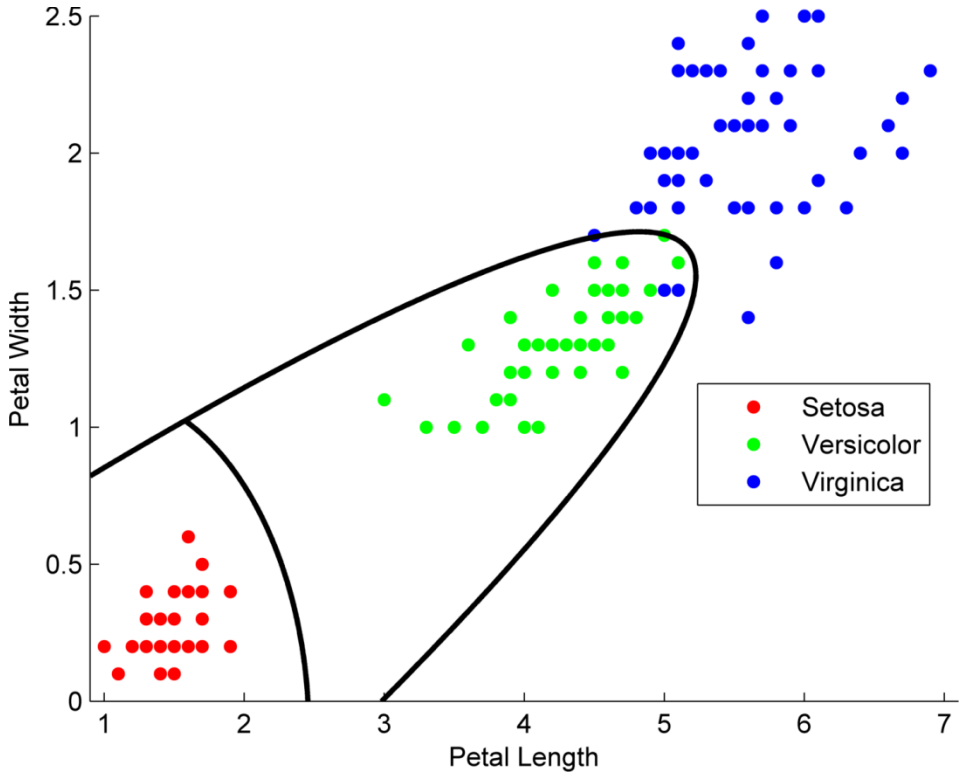


Figure 22: Example of QDA classification on the iris dataset [cm]. Two quadratic boundaries are shown between the setosa/versicolor and the versicolor/virginica classes.

With sufficient data, both LDA and QDA have been found to perform very well in practice [131, 147]. LDA may not perform well if the covariances of the class datasets are very dissimilar. QDA will generally perform better in these situations, though at some computational expense. LDA/QDA performance can also be affected by severely non-normal or noisy distributions, and LDA implicitly assumes the mean of the data is the discriminating factor, not the variance. However, standardized PCA has been shown to help by noise reduction from correlated features and incorporating the data's variance as a feature selection step. Normality is not an absolute assumption for the input dataset, and discriminant analysis can still be useful given that it performs well as a classifier [148].

3.4: Naïve Bayes

The Naïve Bayes (NB) classifier is a simple probabilistic, supervised classifier. The term supervised refers to the construction of a rule or classification procedure from a given set of data with known class labels [131, 134]. NB is relatively easy to construct, as it does not require complicated methods for estimating model parameters. It is also relatively easy to interpret and can perform surprisingly well despite its simplicity, as the independence assumption leads to low variance in its probability estimates [149]. And although this assumption (hence the term "naïve") is often not applicable to real-world data, it has been shown to perform quite well in practice despite the presence of strong inter-feature dependencies [134, 149, 150]. It can also perform well with smaller sample sizes. It has been used in CT imaging as both a prediction and classification tool for head injuries and texture analysis for emphysema [151-153].

The NB algorithm is formulated from Bayes' theorem (see equation 17). Assuming some class variable C and a set of feature vectors x_1 through x_n, Bayes' theorem states:

$$P(C|x_1, \dots, x_n) = \frac{P(C)P(x_1, \dots, x_n|C)}{P(x_1, \dots, x_n)} \quad \textbf{(22)}$$

Assuming independence between the features, we can simplify the equation to the following:

$$P(C|x_1, \dots, x_n) = \frac{P(C)\prod_{i=1}^{n} P(x_i|C)}{P(x_1, \dots, x_n)} \quad \textbf{(23)}$$

The denominator $P(x_1, \dots, x_n)$ is constant, and so we can construct the classifier from the expressions:

$$P(C|x_1, \dots, x_n) \propto P(C)\prod_{i=1}^{n} P(x_i|C) \quad \textbf{(24)}$$

$$\hat{C} = \underset{C}{\operatorname{argmax}}\left(P(C)\prod_{i=1}^{n} P(x_i|C)\right) \quad \textbf{(25)}$$

Each of the distributions $P(x_i|C)$ are estimated separately. As such, the n dimensional multivariate problem is simplified to n univariate estimation problems. Returning to the iris dataset and modeling the predictor distribution of each iris class as Gaussian, we can visualize the class distributions as seen in **Figure 23**.

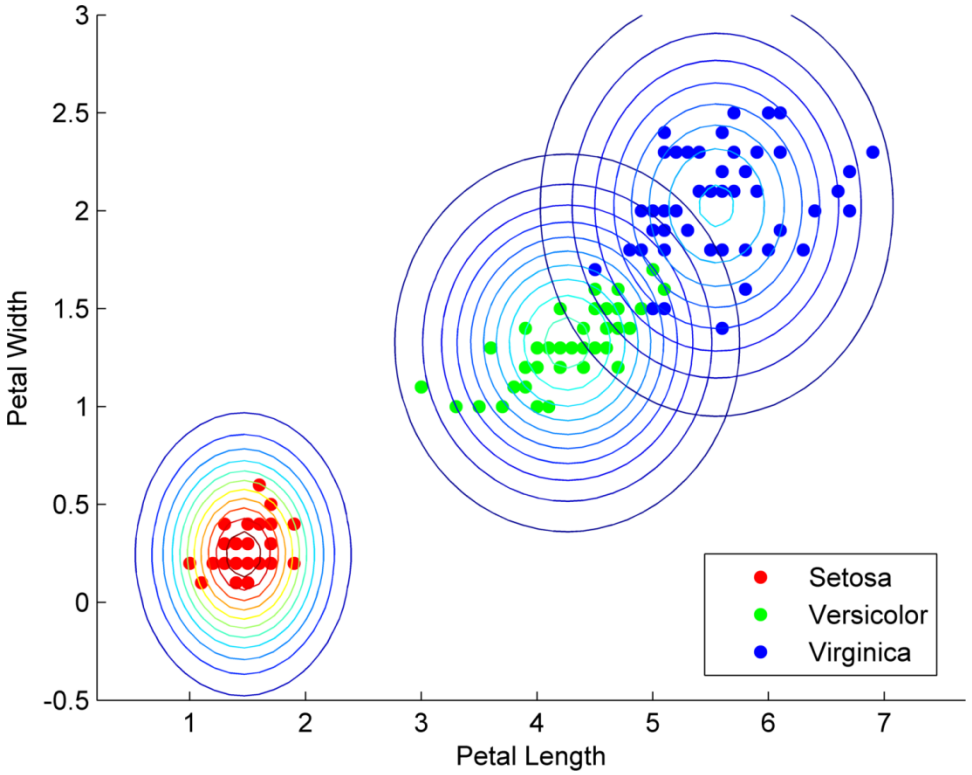


Figure 23: Example of Gaussian distributions of each class in the iris dataset [cm].

3.5: Logistic regression

Logistic regression (LR) is a model similar to ordinary linear regression, except that it estimates the probability of an event occurrence rather than some predictive change in a dependent variable. While ordinary regression models a continuous outcome variable, LR classification can model a dichotomous outcome variable using some set of independent variables [131]. It has found many uses in the imaging literature – some examples include classification of head injury with CT

imaging [154], predicting coronary heart disease [155], and prostate cancer detecting using MRI imaging [156].

The derivation of the LR model begins with a definition of the odds ratio, which is defined as the odds of the occurrence *p* relative to the occurrence *q*. Assuming a binary situation where *p* and *q* sum to 1 (i.e. *q* is "not *p*"):

$$odds = \frac{p}{q} = \frac{p}{1-p} \tag{26}$$

The essence of logistic regression is to provide a method for modeling a binary response variable. To change the range of proportions from the bounded interval (0,1) to the unbounded interval (-∞,∞), the natural log of this odds ratio is taken. This logit function [157] can be equated to a linear regression line, which is also an unbounded function:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{27}$$

P is the probability of a binary outcome, β_0 represents a constant, and $\beta_i x_i$ represents one of *n* independent variables for modeling the binary outcome. Solving this equation for *P* will give a constrained output in the interval (0,1) with an unconstrained output [158]:

$$\frac{P}{1-P} = e^{(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)} \tag{28}$$

$$P = \frac{e^{(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}{1 + e^{(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}} = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}} \tag{29}$$

The general form of this equation is called the logistic function, as seen in **Figure 24**:

$$f(z) = \frac{1}{1 + e^{-z}} \tag{30}$$

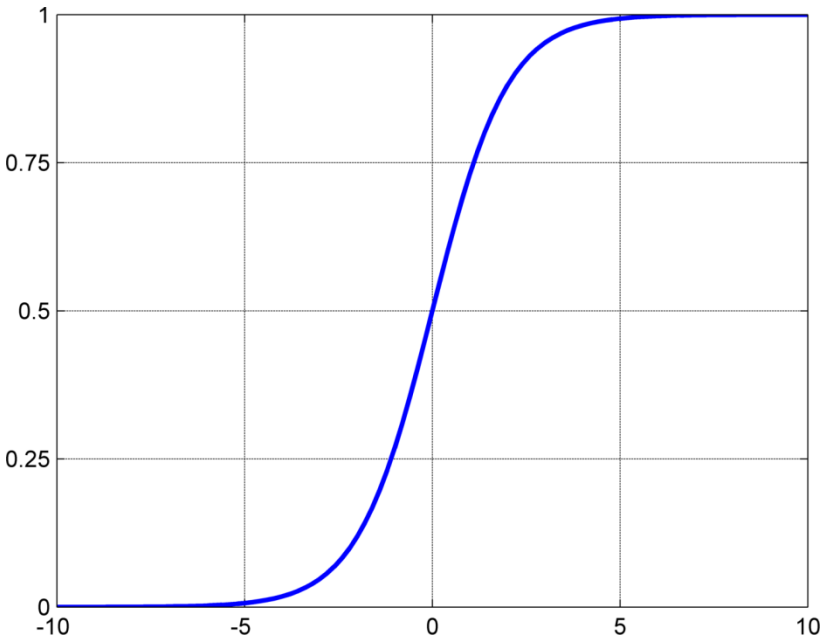


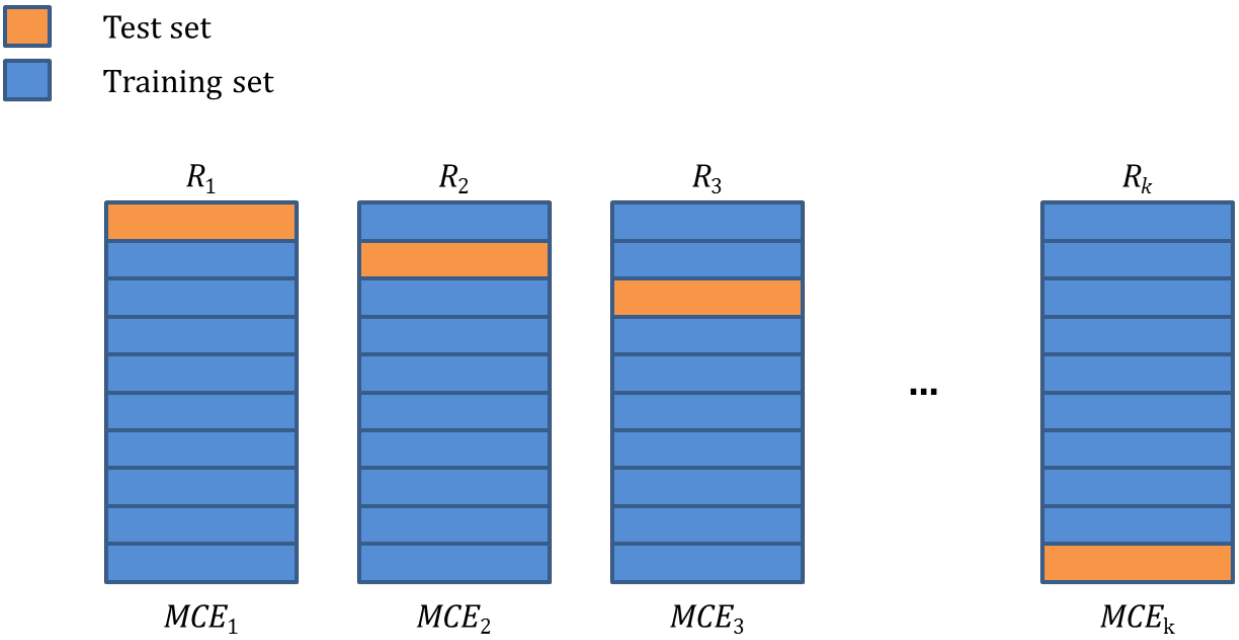
Figure 24: Plot of the basic logistic equation.

LR makes no assumptions about the underlying distributions of the independent variables, allowing for additional flexibility in variable selection. However, these variables should not be highly correlated with one another. In addition, large sample sizes are required in order to have a high power for the model goodness-of-fit [157]. Its use for classification has often been compared to that of LDA – although LR allows for additional flexibility due to fewer assumptions about the data distribution, both models have been shown to perform similarly in practice [157, 159-161]

3.6: Model evaluation

The accuracy of a model as measured on a training dataset could be different from what is measured on unseen data. In other words, the model parameters that are optimized for the training data (and thereby result in a high accuracy) could perform differently for new unseen data. This latter accuracy, where the true classification rate is usually unknown, is of greater practical importance. However, finite samples by definition are approximations to the behavior of a

population. To simulate this unseen population behavior, the collective training set can be partitioned into two groups: one subset acts as the new training set upon which the model is trained, and the other is used as 'unseen' test data to estimate the retrained classifier's accuracy. This provides an unbiased estimate of the model accuracy and prevents model overfitting because the entire dataset isn't used for model training. This process is termed cross-validation (CV) [131, 146].



$$Final\ MCE = \frac{1}{k}\sum_{i=1}^{k} MCE_i$$

Figure 25: Diagram of k-fold CV. The data is randomly split into k unique folds (in this example, k = 10). For each round R_k, the model is trained using the blue data and tested using the orange data, resulting in a misclassification error (*MCE*) for each round. The final accuracy is an average of the k rounds.

There are multiple ways to implement CV. Assume a dataset with N observations. *Holdout CV* partitions data into two subsets of pre-specified ratios for training and testing. *K-fold CV* randomly splits the data into k unique partitions, or "folds," where each fold has N/k observations. The model is trained and tested k times, with each fold being used for testing and the remaining N - N/k points used for model training (**Figure 25**). In this method, each data point is used exactly once

for testing and *k*-1 times for training. In *leave-one-out CV (LOOCV)*, the dataset is split into *N* folds with the test group of size 1 and the training group of size *N* - 1.

LOOCV is approximately unbiased for the true misclassification error (MCE), but may have high variance due to the strong similarity between the training datasets [146]. In addition, there is additional computational burden due to repeating the CV technique *N* times. Implementing k-fold CV with a relatively low *k* (e.g. *k* = 5) would have lower variance, but could be biased depending on the classifier performance as a function of the training set size. As such, we elected to implement k-fold CV for our data with a value of *k* = 10 as a recommended compromise [146, 162, 163]. The CV was also stratified such that each fold contained the same label proportions as the original dataset, as this has been shown to reduce estimation bias [163].

For each fold, the MCE is calculated as the number of misclassifications in the testing dataset over the number of data points in the testing set. For a binary classifier, misclassifications can fall into one of two categories: a "true" label mislabeled as "false," and vice versa. The former is considered a *false negative*, and the latter a *false positive*. These specific types of misclassifications can be identified using a 2x2 contingency table for each fold (**Figure 26**).

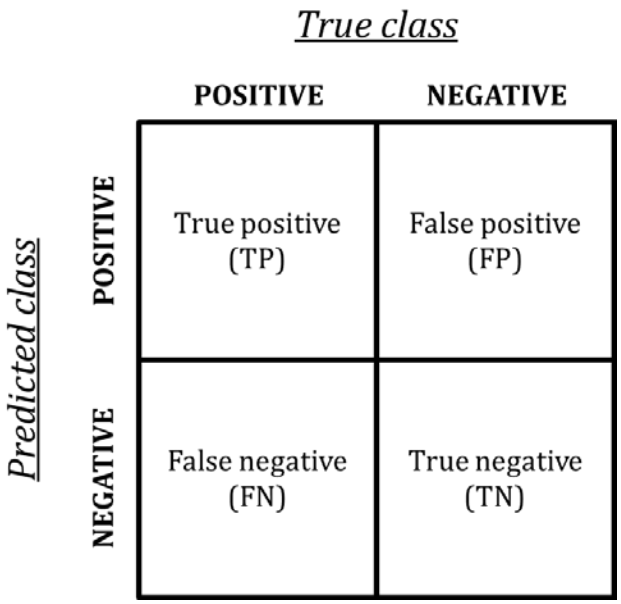


Figure 26: Sample 2x2 contingency table. From here, the true positives, true negatives, false positives, and false negatives can be summed for various population-based classifier measures.

These contingency tables can be used to calculate a number of population-based measures that estimate the model's ability to perform in practice. *Sensitivity* is the fraction of actual positives that are correctly identified as positive. *Specificity* is the fraction of actual negatives that are correctly identified as negative. High sensitivities can help rule out a condition (i.e. WP or misaligned patient) when a classifier outputs a negative result, and high specificities can help rule in a condition (i.e. RP) with a positive result. These parameters can be calculated from the following relations:

$$Sensitivity = \frac{TP}{P_{total}} = \frac{TP}{TP + FN}; \quad Specificity = \frac{TN}{N_{total}} = \frac{TN}{TN + FP} \tag{31}$$

Sensitivity and specificity are synonymous to the true positive rate and false positive rate, respectively. There is an inherent tradeoff between these two parameters – an increase in sensitivity will result in a decreased specificity, and vice versa. The choice of which value to optimize depends strongly on the ultimate objective of the classifier, as well as the relative importance placed on each parameter. A given binary classifier can take on varying combinations of sensitivity and specificity, depending on the selection of a given decision threshold. This effect can empirically visualized using a *receiver operator characteristic*, or *ROC curve*, which illustrates possible combinations of the correct and incorrect decision frequencies as a function of the discrimination threshold [164]. An example curve can be seen in **Figure 27**.

An ROC curve allows for the selection of an optimal model depending on the classifier's purpose, and is directly related to the cost/benefit analysis of diagnostic decision making. One commonly-used metric to describe the curve is the *area under the curve*, or the *AUC*, which quantifies the overall ability of the classifier to discriminate between those who do and do not have a given condition. It also represents the probability that the classifier will rank a randomly-selected test subject higher than a randomly-selected control, assuming the test subject has a higher test value than a control[165]. An AUC of 0.5 represents a test that has no discriminative power, i.e. a

random guess (dashed black line in **Figure 27**). An AUC of 1 indicates a perfectly-performing classifier with no false positives or false negatives.

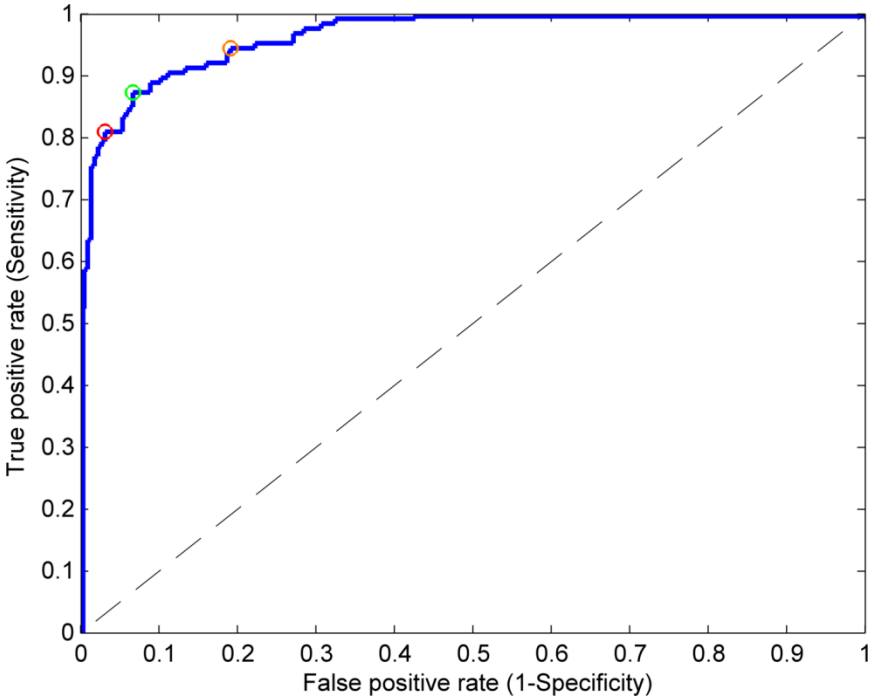


Figure 27: Sample ROC curve. Dashed line represents a line of no discrimination, or equal probability of predicting a positive or negative outcome. Circled points represent various sensitivity and specificity combinations corresponding to different decision boundaries, which can be chosen depending on the classifier's purpose.

Another metric to measure the overall quality of a binary classifier is the Matthew's correlation coefficient (MCC) [166]. It has been shown to be insensitive to class size, and is generally regarded as a balanced, single-value measure to describe a 2x2 contingency matrix [167, 168]:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad \textbf{(32)}$$

A natural extension of a classifier is to determine the probability of a condition's presence for a given individual. One limitation of the sensitivity and specificity parameters is that given a positive test result, they do not inform the user of the probabilistic likelihood of having the condition. A positive predictive value, defined as the proportion of people with a positive result that actually have the disease, can be useful to address this question; however, it is numerically unstable due to its reliance on the actual prevalence of the condition in the population [169, 170]. Our study included N RP and $2N$ WP values for the patient identification study, which is not reflective of the

actual proportion of wrong patients imaged prior to RT treatment. As such, the post-test probability can be estimated more accurately by calculating the *likelihood ratio* of a test. The likelihood ratio uses both the sensitivity and specificity of a test to estimate the likelihood of a test result to change the probability that a condition actually exists [169, 171]. The likelihood ratio can be calculated for both positive and negative test results:

$$LR+ = \frac{sensitivity}{1 - specificity}; \quad LR- = \frac{1 - sensitivity}{specificity} \tag{33}$$

Values of $LR+$ greater than 1 indicate that a positive test result is more likely to occur in people with a condition than those without the condition; values of $LR-$ less than 1 mean that individuals without a particular condition are more likely to have a negative test result than those with the condition [171]. Larger and smaller values of $LR+$ and $LR-$, respectively, indicate a more reliable test result for assessing the likelihood of the presence or absence of a condition. These values inherently reflect the tradeoff between sensitivity and specificity, and do not have any instability relative to the condition's prevalence in the population.

3.7: Classification summary

Features from the patient identification and misalignment workflows, for all three sites and both imaging modalities, were used to train and test the five classification models described in this chapter. The CC/MI/SSIM features were first used to train and test the models. The principal components of the gradient-based metrics were then added to the CC/MI/SSIM features, and models were retrained and retested. All training/testing of models was performed 100 separate times, and the aggregate results (MCE, sensitivity, specificity, MCC, LR+, LR-, AUC) with the appropriate confidence intervals were calculated. If perfect classification was achieved (i.e. 100% sensitivity/specificity and 0% MCE), LR+ and LR- values were approximated using sensitivity/specificity combinations of 100%/99.99% and 99.99%/100%, respectively (i.e. resultant LR+ and LR- values of 10,000 and 0.0001, respectively). Two-sample t-tests were

performed using MATLAB's Statistics Toolbox to compare MCE differences for select datasets (e.g. inclusion vs. exclusion of the gradient-based features, bilateral filtering vs no filtering).

To determine optimal parameters for the KNN, we tested *k* values ranging from 1 to 17, three weighting functions (no weighting, inverse weight, and inverse squared weighting), and 11 distance measures (City block, Chebychev, correlation, cosine, Euclidean, Hamming, Jaccard, Mahalanobis, Minkowski, standardized Euclidean, and Spearman). The final model parameters that resulted in the lowest MCE were k=7, a standardized Euclidian distance, and inverse squared weighting.

For LDA implementation, Bartlett's test was used to test if RP and WP samples came from populations with equivalent variances [172]. A linear and quadratic discriminant function was used to fit the LDA and QDA models, respectively. Pseudoinverses of the covariance matrices were used for QDA classification to account for the presence of any singular covariance matrices.

In NB classification, a Gaussian distribution may not be a fitting distribution choice for the feature data. Lilliefors test can be used to test the null hypothesis that the datasets come from a normal population [173]. Each individual feature was tested for normality using Lilliefors test, and a Gaussian or kernel distribution was selected accordingly for model training. A kernel distribution is a nonparametric representation of a variable's density function and can be used for skewed distributions or those with multiple peaks [174].

To optimize LR parameters, models were initially trained using just the one feature and the model's deviance parameter was then calculated. The deviance is a measure of the model's lack of fit to the input data, where smaller values indicate a better model fit [158]. Individual features were subsequently added in a stepwise fashion, with the model retrained and the deviance parameter recalculated. If the new deviance was smaller than the previous deviance, the feature was included in the final set of features.

CHAPTER 4: RESULTS

4.1: Patient identification: MI/CC/SSIM

Initially, separate classifications were performed using just one feature (MI, CC, or SSIM). To examine the effect of image filtering, MI/CC/SSIM metrics were also calculated on images smoothed with a bilateral filter. MCEs comparing the two can be seen in **Tables 4** and **5** using the five classification algorithms across all treatment sites and both machines.

TOMO	MI					CC					SSIM				
	LR	*NB*	*KNN*	*LDA*	*QDA*	*LR*	*NB*	*KNN*	*LDA*	*QDA*	*LR*	*NB*	*KNN*	*LDA*	*QDA*
H&N	2.3	2.6	3.1	2.3	2.2	0.7	0.9	1.0	0.4	0.6	1.4	1.3	2.6	1.5	1.3
H&N (B)	1.3	1.4	1.7	1.9	1.4	0.7	0.9	0.9	0.8	0.5	2.0	2.5	2.8	2.3	2.4
Pelvis	0.9	1.1	1.1	1.6	1.0	1.7	1.7	2.6	1.7	1.7	2.3	2.0	2.7	1.3	2.4
Pelvis (B)	1.6	1.0	1.6	1.0	1.7	4.3	4.2	6.5	4.2	4.1	2.2	2.1	3.0	1.7	1.9
Spine	1.2	1.7	1.8	1.8	1.5	2.4	1.7	1.4	2.3	2.3	3.0	3.0	3.7	3.0	2.5
Spine (B)	3.4	2.7	3.2	2.4	3.5	2.4	2.7	1.9	2.4	2.3	3.0	3.0	5.1	3.0	2.9

(B) = bilateral filtered images; MI = mutual information; CC = cross-correlation coefficient; SSIM = structural similarity; LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis

Table 4: Average misclassification errors for Tomo image pairs for patient identification. Classification was performed using a single feature of MI, CC, or SSIM. Each feature was extracted from image pairs with and without smoothing from a bilateral filter (B). All values are shown as a percentage. 95% confidence intervals ranged from 0% to 0.1% across all values.

For Tomo image pairs without bilateral filtering, CC had the lowest MCE (1.5±0.7) compared to MI (1.7±0.6) and SSIM (2.3±0.8) across all algorithms and sites. The use of image filtering had a variable effect on classification accuracy, depending on both treatment site and similarity metric. H&N saw a 1% improvement in MCE using MI, no change using CC, and a 0.8% increase in MCE using SSIM across all algorithms. The pelvis site showed an MCE increase of 0.3% and 2.8% using MI and CC, respectively, and no change in SSIM. The spine site had an average MCE increase of 1.4%, 0.3%, and 0.4% for the MI, CC, and SSIM metrics, respectively. For all sites, H&N had the lowest overall MCE for both excluding and including the bilateral filtered data. LDA had the best overall performance across all classifiers.

TBEAM	MI					CC					SSIM				
	LR	*NB*	*KNN*	*LDA*	*QDA*	*LR*	*NB*	*KNN*	*LDA*	*QDA*	*LR*	*NB*	*KNN*	*LDA*	*QDA*
H&N	5.3	5.9	8.5	7.6	5.5	4.9	4.7	6.9	4.6	5.1	3.4	4.3	5.1	6.5	4.0
H&N (B)	5.7	5.2	7.5	8.0	5.6	4.6	4.4	7.6	4.2	4.5	4.2	4.5	6.4	4.2	4.2
Pelvis	2.8	2.9	7.0	4.9	2.8	2.0	2.4	3.2	2.3	2.5	3.2	2.9	3.3	5.7	3.3
Pelvis (B)	5.0	5.3	8.7	5.4	5.0	1.9	1.7	2.7	1.7	1.8	1.8	2.6	2.6	2.3	1.8
Spine	3.4	3.5	3.2	3.5	2.9	2.5	2.3	3.6	2.3	3.4	1.6	1.8	2.9	2.4	1.8
Spine (B)	4.1	4.1	6.4	5.1	4.2	2.6	2.4	3.7	2.3	2.9	3.5	3.7	5.2	4.8	3.6

(B) = bilateral filtered images; MI = mutual information; CC = cross-correlation coefficient; SSIM = structural similarity; LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis

Table 5: Average misclassification errors for TBeam image pairs for patient identification. Classification was performed using a single feature of MI, CC, or SSIM. Each feature was extracted from image pairs with and without smoothing from a bilateral filter (B). All values are shown as a percentage. 95% confidence intervals ranged from 0% to 0.2% across all values.

TBeam image pairs generally produced higher MCEs than Tomo image pairs. CC and SSIM performed similarly across unfiltered image pairs (3.5±1.4) while MI had poorer performance (4.6±1.9). With the inclusion of filtering, H&N had a slight improvement for the MI and CC metrics (0.2% decrease in MCE) with no change in SSIM. Pelvis images saw a 1.8% MCE increase for MI, 0.5% decrease for CC, and 0.1% decrease for SSIM. Spine images had a 1.5% and 2.1% increase for MI and SSIM, respectively, and no change for CC. LR had the best overall classification performance.

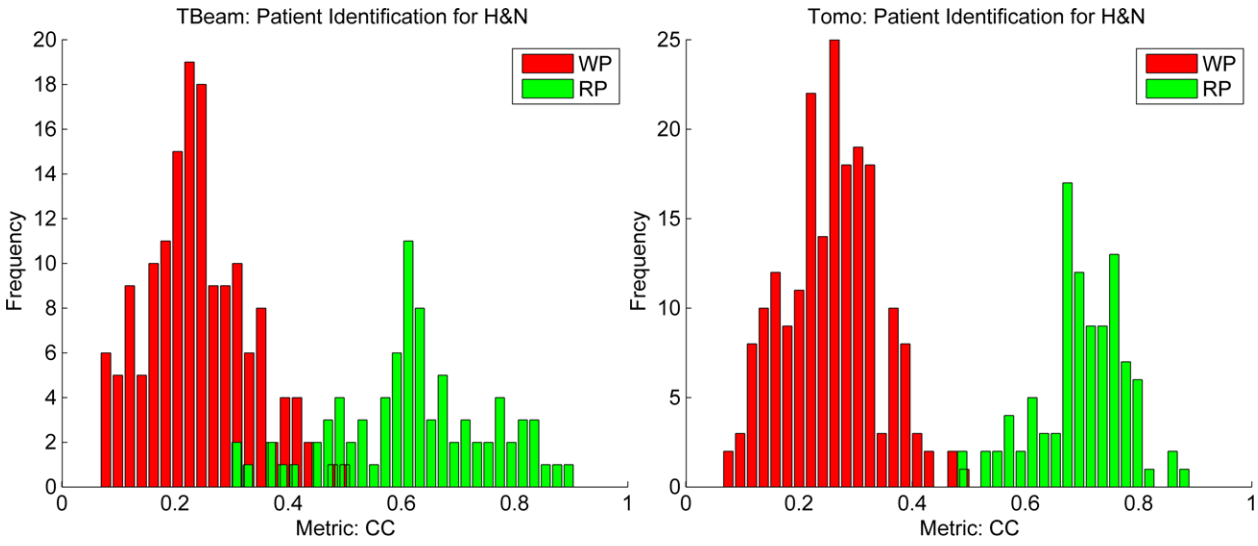


Figure 28: CC frequency histograms of the H&N site for the TBeam and Tomo machines. (WP = wrong patient, RP = right patient, CC = cross-correlation coefficient)

Figures 28-30 show 1D histograms of the CC metric frequency distribution for collected data in each anatomical site for both machines.

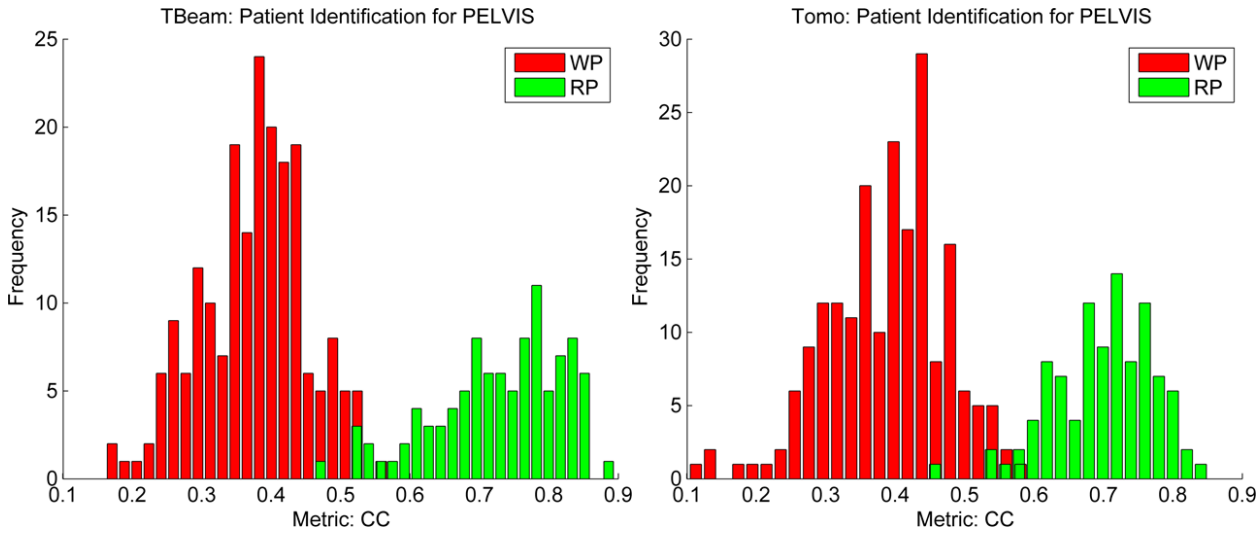


Figure 29: CC frequency histograms of the pelvis site for the TBeam and Tomo machines. (WP = wrong patient, RP = right patient, CC = cross-correlation coefficient)

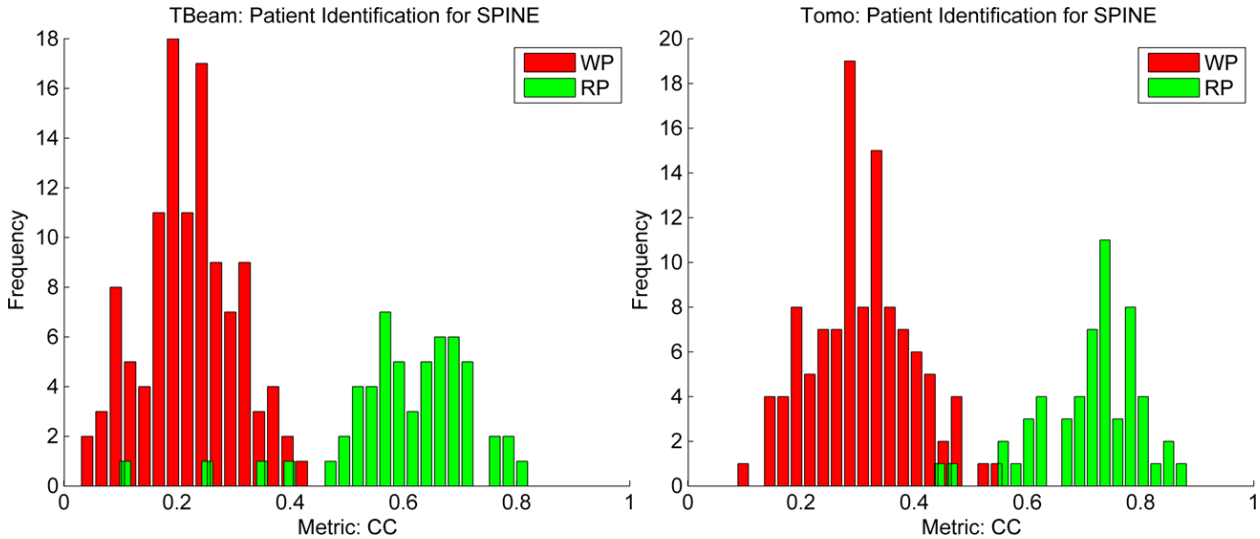


Figure 30: CC frequency histograms of the spine site for the TBeam and Tomo machines. (WP = wrong patient, RP = right patient, CC = cross-correlation coefficient)

The spread of the RP histograms is larger for the TBeam images than the Tomo images. For example, H&N values from TBeam images had a mean of 0.63 ± 0.13 while the Tomo values had a spread of 0.71 ± 0.08. Low TBeam CC values for RP image pairs corresponded strongly to images with high noise and artifacts. High CC values from both incorrect image pair histograms generally corresponded to patients with similar sizes and anatomical features.

Figure 31 shows Tomo pelvis values comparing inclusion or exclusion of bilateral filtering. **Figure 32** shows Tomo spine histograms using the MI and SSIM metrics.

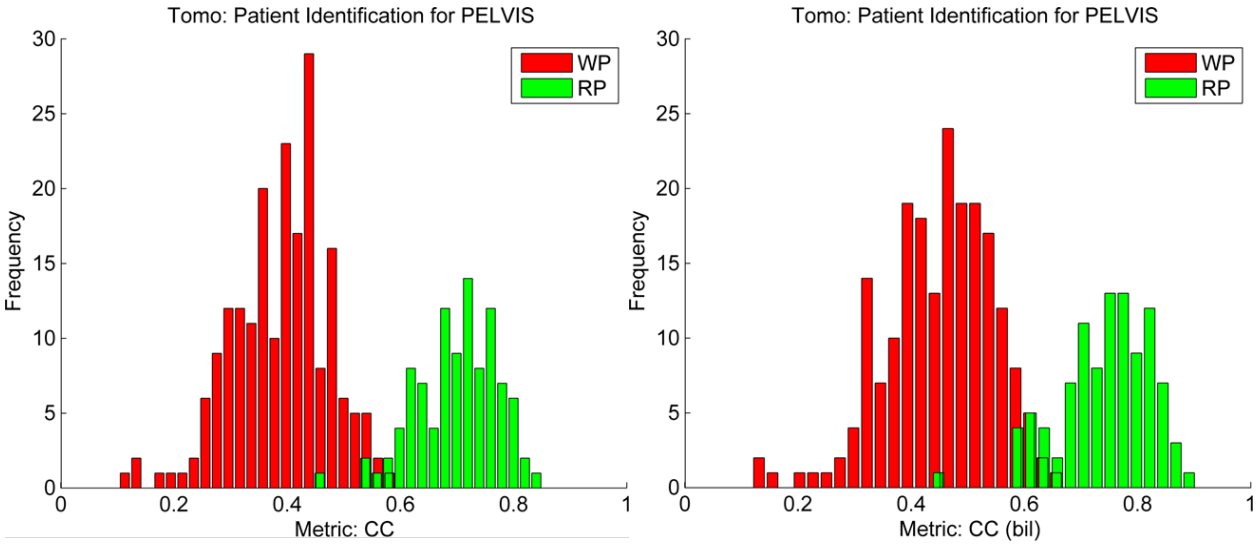


Figure 31: CC frequency histograms of the pelvis site for Tomo, with and without bilateral filtering. (WP = wrong patient, RP = right patient, CC = cross-correlation coefficient, bil = bilateral filtering)

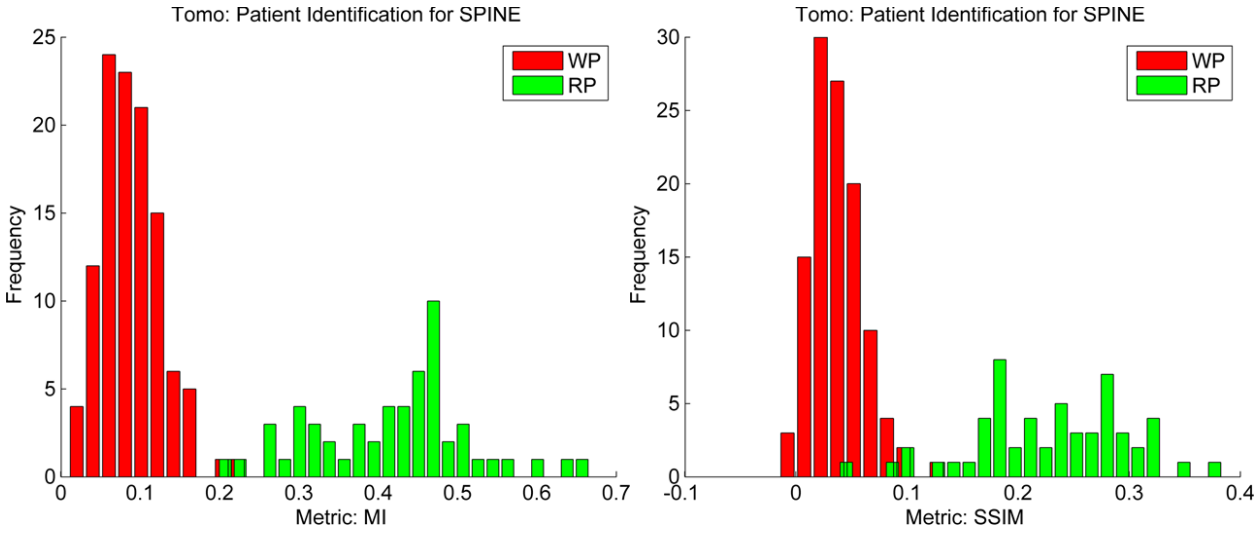


Figure 32: MI and SSIM frequency histograms of the spine site for Tomo.
(WP = wrong patient, RP = right patient, MI = mutual information, SSIM = structural similarity)

The MI/CC/SSIM metrics were then used together as features for classification. Both unfiltered and bilateral-filtered metrics were run as two separate cases. Classification results can be seen in **Table 6**. Sensitivity and specificity ranged from 98% to 100% across all Tomo image pairs, and 95.2% to 98.7% across all TBeam image pairs. An overall improvement was seen by including all the metrics for classification. H&N and pelvis sites showed improved errors through the use of bilateral filtering, while the spine site had a higher error.

		MI/CC/SSIM					MI/CC/SSIM (bil)				
		LR	***NB***	***KNN***	***LDA***	***QDA***	***LR***	***NB***	***KNN***	***LDA***	***QDA***
Tomo	*H&N*	1.3	0.7	0.7	1.0	0.7	0.4	0.7	0.7	0.3	0.3
	Pelvis	2.0	1.6	1.7	1.7	1.4	0.7	0.3	0.4	1.0	0.9
	Spine	0.09	0.1	0	1.6	0.6	1.2	0.9	1.1	1.2	1.1
TBeam	*H&N*	4.6	5.0	5.0	5.4	3.4	3.8	3.7	3.2	3.7	2.7
	Pelvis	1.9	1.9	2.0	2.8	1.3	1.8	1.7	2.3	2.0	1.9
	Spine	2.3	2.3	1.8	1.8	2.1	3.1	3.4	2.9	2.9	3.1

(bil) = bilateral filtered images; MI = mutual information; CC = cross-correlation coefficient; SSIM = structural similarity; LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis;

Table 6: Average misclassification errors for Tomo and TBeam image pairs for patient identification. Classification was performed using the MI, CC, and SSIM features with and without smoothing from a bilateral filter (bil). All values are shown as a percentage. 95% confidence intervals ranged from 0% to 0.05% across all values.

4.2: Patient identification: including gradient-based features

Tables 7 and **8** summarize classification results from the inclusion of the gradient-based metrics for the Tomo machine, excluding or including bilateral filtering of the MI/CC/SSIM metrics, respectively. Additional parameters of sensitivity, specificity, MCC, LR+, LR-, and AUC are also shown.

Tomo: Grad + MI/CC/SSIM						
		LR	***NB***	***KNN***	***LDA***	***QDA***
H&N	*MCE*	0.03±0.02	1.0±0.05	0.6±0.03	0.7±0.02	0.7±0
	Sens	100±0	99.1±0.1	99.3±0.08	99.2±0.09	99.2±0.08
	Spec	99.9±0.06	98.9±0.1	99.6±0.09	99.5±0.1	99.5±0.1
	MCC	0.999±4e-4	0.979±1e-3	0.987±6e-4	0.987±4e-4	0.986±6e-4
	LR+	9010±585	1772±734	4767±970	4663±970	4863±972
	LR-	0.0001±0	0.009±1e-3	0.007±8e-4	0.008±9e-4	0.008±8e-4
	AUC	1±0	0.999±1e-5	0.998±5e-4	0.999±9e-6	0.990±1e-5
Pelvis	*MCE*	3.6±0.1	1.7±0.06	1.3±0.04	1.7±9e-3	3.0±0.1
	Sens	98.5±0.08	98.7±0.1	98.5±0.1	98.0±0.1	98.1±0.2
	Spec	92.5±0.3	98.9±0.1	99.0±0.4	98.8±0.2	95.6±0.3
	MCC	0.921±3e-3	0.975±1e-3	0.973±9e-4	0.966±2e-4	93.9±2e-3
	LR+	13.8±0.6	1975±766	2374±821	1963±767	26.8±3.2
	LR-	0.016±9e-4	0.013±1e-3	0.016±1e-3	0.02±1e-3	0.02±2e-3
	AUC	0.975±1e-3	0.999±1e-4	0.993±3e-4	0.999±8e-5	0.991±1e-3
Spine	*MCE*	0.8±0.07	0.2±0.06	1.0±0.08	0±0	0.2±0.07
	Sens	99.8±0.08	99.8±0.09	98.8±0.2	100±0	99.8±0.09
	Spec	98.2±0.2	99.9±0.09	99.3±0.2	100±0	99.9±0.1
	MCC	0.983±2e-3	0.996±1e-3	0.980±2e-3	1±0	0.996±1e-3
	LR+	1048±588	8909±611	5531±973	10000±0	9107±560
	LR-	0.003±8e-4	0.002±9e-4	0.01±2e-3	0.0001±0	0.002±9e-4
	AUC	0.999±4e-4	0.999±9e-4	0.999±2e-5	1±0	0.999±3e-4

Grad = gradient-based features; MI = mutual information; CC = cross-correlation coefficient; SSIM = structural similarity;
LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis;
MCE = misclassification error; sens = sensitivity; spec = specificity; MCC = Matthew's correlation coefficient; LR+ = positive likelihood ratio; LR- = negative likelihood ratio; AUC = area under the curve

Table 7: Classification algorithm outputs for Tomo image pairs for patient identification. Classification was performed using all features with the MI/CC/SSIM metrics without bilateral filter smoothing. All values are shown as μ±2σ averaged across 100 trial runs, where σ is one standard deviation.

Tomo: Grad + MI/CC/SSIM (bil)						
		LR	***NB***	***KNN***	***LDA***	***QDA***
H&N	*MCE*	0.5±0.04	0.9±0.03	0.7±0	0.3±0.03	0.7±0
	Sens	99.5±0.02	99.1±0.09	99.2±0.08	99.7±0.06	99.3±0.08
	Spec	99.6±0.1	99.2±0.1	99.5±0.1	99.8±0.07	99.5±0.1
	MCC	0.990±1e-3	0.982±7e-4	0.986±6e-5	0.995±7e-4	0.986±7e-5
	LR+	6434±937	2773±866	4764±971	7436±852	4071±953
	LR-	0.005±2e-4	0.009±9e-4	0.008±8e-4	0.003±6e-4	0.007±8e-4
	AUC	0.999±1e-4	0.999±7e-5	0.995±4e-4	0.999±4e-6	0.990±4e-5
Pelvis	*MCE*	3.6±0.1	1.7±0.05	1.3±0.06	1.7±0.02	3.8±0.1
	Sens	98.4±0.08	98.7±0.1	98.5±0.1	98.1±0.1	97.3±0.3
	Spec	92.5±0.3	98.9±0.1	99.0±0.1	98.7±0.2	94.7±0.4
	MCC	0.920±2e-3	0.975±7e-4	0.973±1e-3	0.966±5e-4	0.922±2e-3
	LR+	13.6±0.5	1576±697	2172±795	1863±751	21.2±2
	LR-	0.02±8e-4	0.01±1e-3	0.02±1e-3	0.02±1e-3	0.03±3e-3
	AUC	0.974±1e-3	0.996±2e-4	0.998±5e-4	0.993±3e-4	0.990±1e-3
Spine	*MCE*	0.7±0.09	0.7±0.05	0.6±0.01	0±0	1.1±0.05
	Sens	99.7±0.08	99.1±0.1	99.2±0.1	100±0	98.6±0.1
	Spec	98.3±0.2	99.5±0.1	99.6±0.1	100±0	99.2±0.2
	MCC	0.983±2e-3	0.986±9e-4	0.988±4e-4	1±0	0.977±1e-3
	LR+	1545±700	6626±926	7222±877	10000±0	5233±977
	LR-	0.003±8e-4	0.009±1e-3	0.008±1e-3	0.0001±0	0.01±1e-3
	AUC	0.999±2e-4	0.985±4e-4	0.999±3e-4	1±0	0.997±6e-4

Grad = gradient-based features; (bil) = bilateral filtered images; MI = mutual information; CC = cross-correlation coefficient; SSIM = structural similarity;
LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis;
QDA = quadratic discriminant analysis;
MCE = misclassification error; sens = sensitivity; spec = specificity; MCC = Matthew's correlation coefficient; LR+ = positive likelihood ratio; LR- = negative likelihood ratio; AUC = area under the curve

Table 8: Classification algorithm outputs for Tomo image pairs for patient identification. Classification was performed using all features with the MI/CC/SSIM metrics with bilateral filter smoothing. All values are shown as μ±2σ averaged across 100 trial runs, where σ is one standard deviation.

Inclusion of the gradient-based features saw a general improvement across all algorithms ($p<0.001$). Without bilateral filtering, H&N showed the largest error reduction with the LR (1.3% to 0.03%) and LDA (1.0% to 0.7%) algorithms ($p<0.001$). Errors were comparable for the pelvis site, although an increase in error was seen for the LR and QDA algorithms ($p<0.001$). The spine site showed the largest improvement with the LDA algorithm (1.6% to 0%) ($p<0.001$). Inclusion of the gradient-based features with bilateral filtering of CC/MI/SSIM had no significant effect on the H&N

results. Pelvis features showed an overall increase in error across all algorithms ($p<0.001$). The spine site had an overall reduction in error across all algorithms, with the largest decrease from the LDA algorithm (1.2% to 0%) ($p<0.001$). For the H&N, pelvis, and spine sites, the best performing classifiers were LR/LDA, KNN, and LDA, respectively. Across all anatomical sites, LDA had the highest accuracy and the best sensitivity / specificity / MCC / LR+ / LR- / AUC parameters.

Tables 9 and **10** summarize classification results from the inclusion of the gradient-based metrics for the TBeam machine, excluding or including bilateral filtering of the MI/CC/SSIM metrics, respectively.

		TBeam: Grad + MI/CC/SSIM				
		LR	***NB***	***KNN***	***LDA***	***QDA***
H&N	*MCE*	4.1±0.2	5.6±0.08	4.2±0.06	3.6±0.05	3.7±0.1
	Sens	96.3±0.2	95.1±0.3	95.4±0.2	96.2±0.2	96.3±0.2
	Spec	95.1±0.3	93.6±0.3	96.6±0.3	96.8±0.2	96.3±0.3
	MCC	0.910±4e-3	0.887±2e-3	0.916±1e-3	0.927±1e-3	0.924±2e-3
	LR+	21.4±1.4	16.4±1.3	234±275	134±195	31.3±3.4
	LR-	0.04±2e-3	0.05±3e-3	0.05±2e-3	0.04±2e-3	0.04±2e-3
	AUC	0.967±2e-3	0.988±3e-4	0.984±7e-4	0.998±7e-5	0.981±5e-4
Pelvis	*MCE*	3.8±0.09	1.8±0.03	1.7±0.02	2.3±0.05	2.2±0.05
	Sens	97.7±0.07	98.4±0.1	98.0±0.1	97.5±0.2	98.4±0.2
	Spec	93.3±0.3	98.7±0.2	98.7±0.2	98.0±0.2	97.1±0.2
	MCC	0.916±2e-3	0.970±8e-4	0.965±5e-4	0.952±1e-3	0.956±1e-3
	LR+	15.3±0.6	1372±657	1372±657	852±531	44±6
	LR-	0.02±7e-4	0.02±1e-3	0.02±1e-3	0.03±1e-3	0.02±2e-3
	AUC	0.962±1e-3	0.997±5e-3	0.983±6e-4	0.998±5e-5	0.991±1e-4
Spine	*MCE*	7.0±0.2	1.8±0.06	2.9±9e-3	2.2±0.05	2.3±0.09
	Sens	95.1±0.2	97.9±0.2	96.7±0.2	97.5±0.2	97.7±0.2
	Spec	89.1±0.5	98.7±0.2	97.8±0.3	98.4±0.2	97.7±0.3
	MCC	0.844±5e-3	0.964±1e-3	0.941±3e-4	0.956±1e-3	0.953±2e-3
	LR+	9.2±0.5	3240±913	1636±719	2341±824	1339±659
	LR-	0.06±2e-3	0.02±1e-3	0.03±2e-3	0.03±2e-3	0.02±2e-3
	AUC	0.930±2e-3	0.982±2e-4	0.961±2e-4	0.982±2e-4	0.968±4e-4

Grad = gradient-based features; MI = mutual information; CC = cross-correlation coefficient; SSIM = structural similarity;
LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis;
QDA = quadratic discriminant analysis;
MCE = misclassification error; sens = sensitivity; spec = specificity; MCC = Matthew's correlation coefficient; LR+ = positive likelihood ratio; LR- = negative likelihood ratio; AUC = area under the curve

Table 9: Classification algorithm outputs for TBeam image pairs for patient identification. Classification was performed using all features with the MI/CC/SSIM metrics without bilateral filter smoothing. All values are shown as $\mu \pm 2\sigma$ averaged across 100 trial runs, where σ is one standard deviation.

TBeam: Grad + MI/CC/SSIM (bil)						
		LR	***NB***	***KNN***	***LDA***	***QDA***
H&N	*MCE*	3.2±0.1	3.2±0.06	3.5±0.06	3.4±0.03	4.3±0.06
	Sens	97.6±0.1	97.1±0.2	96.2±0.2	96.3±0.2	96.0±0.2
	Spec	95.2±0.2	96.5±0.2	97.1±0.2	97.1±0.2	95.4±0.3
	MCC	0.929±3e-3	0.936±1e-3	0.930±1e-3	0.932±7e-4	0.913±1e-3
	LR+	22.1±1.5	31.1±2.8	336±335	437±384	123±196
	LR-	0.03±1e-3	0.03±2e-3	0.04±2e-3	0.04±2e-3	0.04±2e-3
	AUC	0.978±1e-3	0.985±6e-4	0.985±8e-4	0.998±9e-5	0.969±4e-4
Pelvis	*MCE*	2.0±0.09	2.0±0.06	1.8±0.03	2.7±6e-3	1.8±0.05
	Sens	98.3±0.1	98.3±0.1	97.9±0.1	96.8±0.2	98.6±0.2
	Spec	97.3±0.2	98.3±0.2	98.8±0.2	98.1±0.2	97.6±0.2
	MCC	0.954±2e-3	0.965±1e-3	0.964±7e-4	0.946±3e-4	0.963±1e-3
	LR+	44±4	463±384	1670±716	955±560	250±274
	LR-	0.02±1e-3	0.02±1e-3	0.02±1e-3	0.03±2e-3	0.01±2e-3
	AUC	0.991±9e-4	0.997±1e-3	0.996±6e-4	0.999±2e-5	0.990±2e-4
Spine	*MCE*	6.6±0.2	2.3±0.07	2.9±0.02	1.9±0.05	2.6±0.09
	Sens	95.4±0.1	97.5±0.2	96.5±0.2	97.7±0.2	97.5±0.2
	Spec	89.4±0.6	98.1±0.3	98.0±0.3	98.8±0.2	97.4±0.3
	MCC	0.852±5e-3	0.952±1e-3	0.941±4e-4	0.962±1e-3	0.947±2e-3
	LR+	9.8±0.6	1940±769	1938±769	3539±934	937±561
	LR-	0.05±1e-3	0.03±2e-3	0.04±2e-3	0.02±2e-3	0.03±2e-3
	AUC	0.917±2e-3	0.979±4e-4	0.962±4e-4	0.982±5e-5	0.967±5e-4

Grad = gradient-based features; (bil) = bilateral filtered images; MI = mutual information; CC = cross-correlation coefficient; SSIM = structural similarity;
LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis;
QDA = quadratic discriminant analysis;
MCE = misclassification error; sens = sensitivity; spec = specificity; MCC = Matthew's correlation coefficient; LR+ = positive likelihood ratio; LR- = negative likelihood ratio; AUC = area under the curve

Table 10: Classification algorithm outputs for TBeam image pairs for patient identification. Classification was performed using all features with the MI/CC/SSIM metrics with bilateral filter smoothing. All values are shown as μ±2σ averaged across 100 trial runs, where σ is one standard deviation.

Inclusion of the gradient-based features resulted in similar trends for the TBeam dataset. H&N saw an overall improvement in MCE across all algorithms, with the largest improvement from LDA (5.4% to 3.6%) ($p<0.001$). The pelvis site saw minor improvements in KNN (2.0% to 1.7%) and LDA (2.8% to 2.3%), but similar to the Tomo dataset, saw an increase in error with the LR algorithm (1.9% to 3.8%) ($p<0.001$). The spine dataset generally did not benefit from the gradient-

based features, with error increases of 1.1% and 4.7% for the NB and LR classifiers, respectively. Using the CC/MI/SSIM metrics with bilateral filtering resulted in an average decrease in error across all algorithms for the H&N (0.7%) and pelvis (0.3%) sites ($p<0.001$), and no significant change for the spine site. The best performing classifiers for H&N, pelvis, and spine sites were LDA/QDA/LR, KNN/QDA, and NB/LDA, respectively. LDA had the best performance across all anatomical sites due to the lowest overall error and highest MCC/LR+/LR-/AUC values. **Figures 33-35** show example ROC curves for all three anatomical sites with inclusion of the gradient-based features using the LDA algorithm. ROC curves were relatively similar across the classification algorithms, as evidenced by the relatively comparable AUC values.

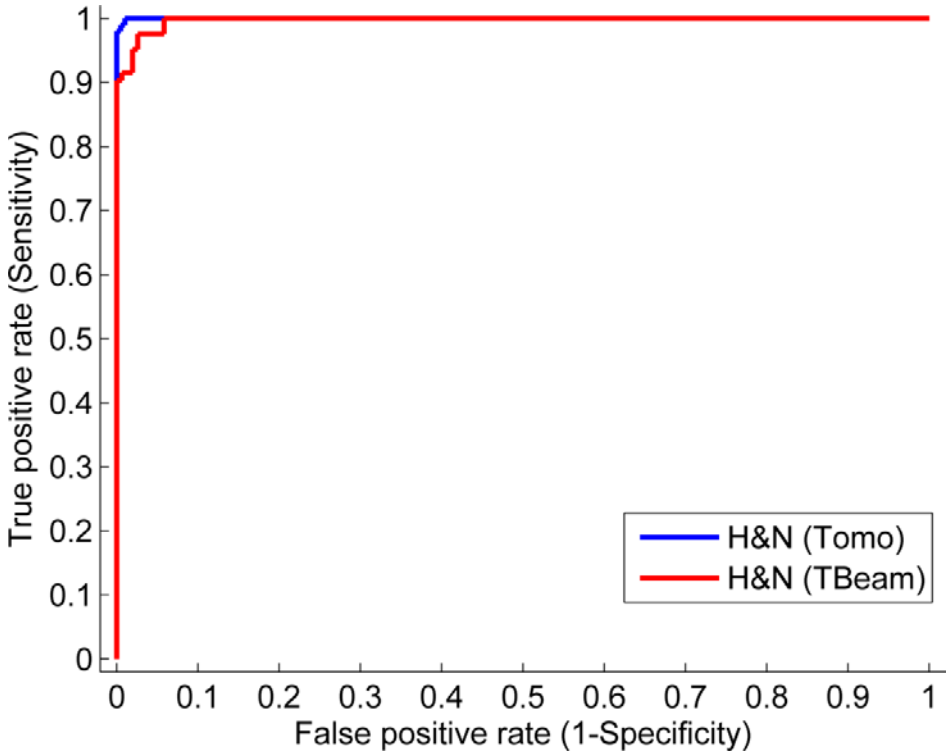


Figure 33: ROC curves for H&N images for patient identification on Tomo and TBeam machines. Curves are shown with inclusion of the gradient-based features.

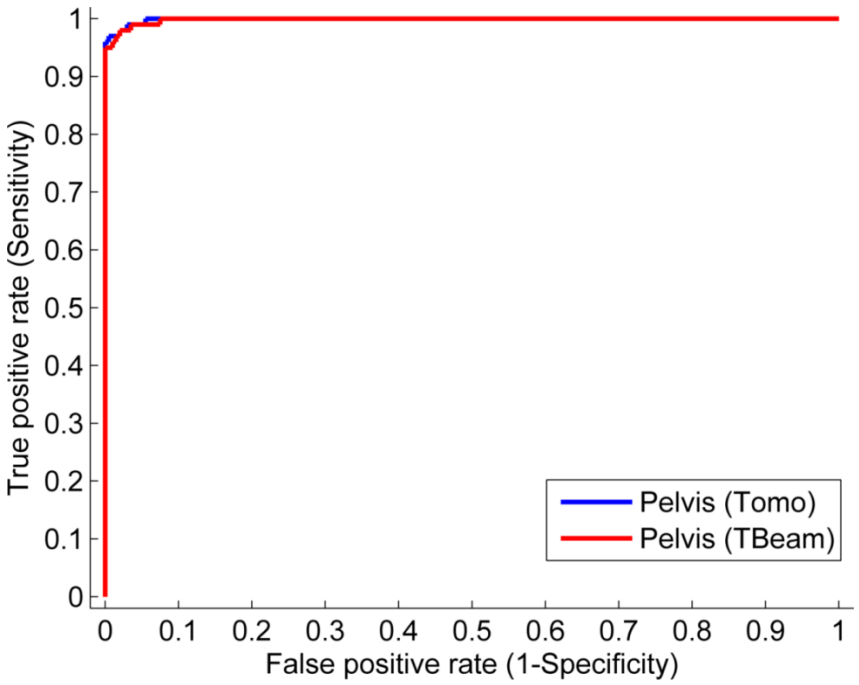


Figure 34: ROC curves for pelvis images for patient identification on Tomo and TBeam machines. Curves are shown with inclusion of the gradient-based features.

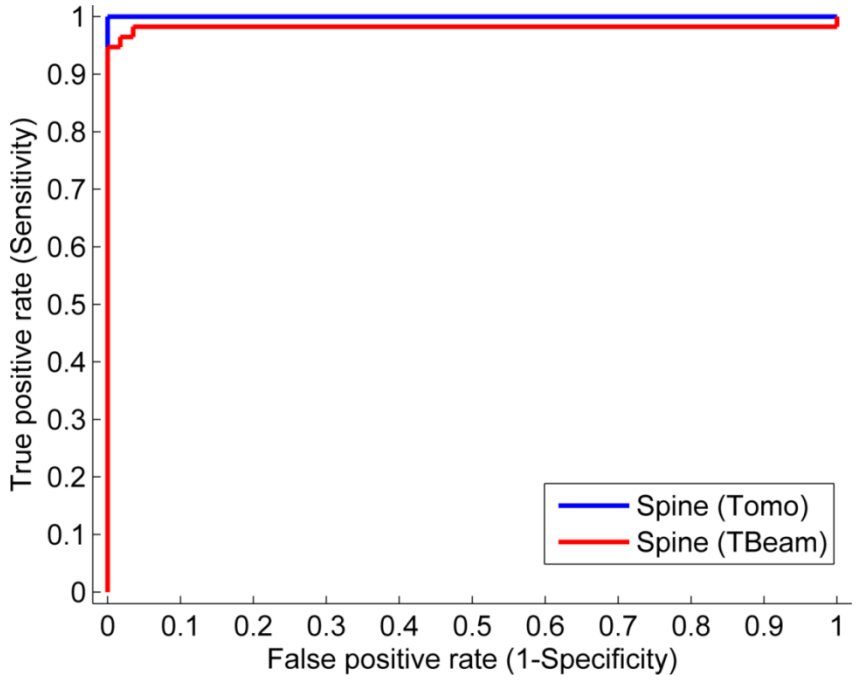


Figure 35: ROC curves for spine images for patient identification on Tomo and TBeam machines. Curves are shown with inclusion of the gradient-based features.

For further investigation, models were also trained and tested using the gradient-based features and individual MI/CC/SSIM metrics. For Tomo images, H&N accuracies showed minor improvements using the CC metric, with the largest MCE improvement in LDA (0.7% to 0.09%) ($p<0.001$). Pelvis and spine images showed no improvement in MCE across all algorithms. For TBeam images, use of the SSIM metric with bilateral filtering showed an average decrease in MCE of 0.9% across the LR/NB/KNN/LDA algorithms ($p<0.001$). The lowest achievable error was 3.1% using the LDA classifier, with sensitivity/specificity values of 96.6% and 97.4%. For the pelvis site, use of CC with bilateral filtering resulted in a 1% and 0.5% decrease in MCE for the LR and LDA algorithms, respectively ($p<0.001$). The spine site also produced decreased MCE estimates by an average of 0.7% across all algorithms ($p<0.001$). The lowest achievable error was 1.8% using LDA (98.0% sensitivity and 98.6% specificity), and the largest improvement was seen for the LR classifier (7.0% to 4.2%) ($p<0.001$).

4.3: Patient alignment: MI/CC/SSIM

Tables 11-13 display average MCEs for the patient misalignment study using a single feature for classification. Six features were examined: MI, CC, and SSIM with and without bilateral filtering. For Tomo image pairs, bilateral filtering reduced classifier accuracy for 1cm H&N and pelvis shifts across all three metrics ($p<0.001$). Bilateral filtering for vertebral shifts increased MCE for the MI and SSIM metrics ($p<0.001$), but decreased with CC ($p<0.001$). Across all sites, the SSIM metric resulted in the smallest average MCE while MI resulted in the largest. KNN was the poorest performing classifier with larger errors than each of the other classifiers across all sites and metrics ($p<0.001$). NB, CC, LDA, and QDA performed comparably for the H&N and pelvis sites. LR was the best overall classifier for vertebral shifts, followed by LDA and QDA. Shifts of 4cm and 5cm had sub-1% MCEs (except 1.4% for a 4cm pelvis shift using KNN).

MI		TOMO							TBEAM						
		H&N			Pelvis			Spine	H&N			Pelvis			Spine
		1cm	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*	*1cm*	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*
No bil	**LR**	3.0	0.5	0.02	6.5	1.9	0.9	9.4	12.8	5.2	2.7	13.2	7.9	5.2	11.1
	NB	3.1	0.3	0	6.6	1.9	0.9	9.3	13.2	4.9	2.6	13.4	8.1	5.4	10.4
	KNN	5.0	0.6	0.02	9.4	2.1	1.4	15.7	15.5	7.4	3.8	18.2	10.9	7.3	16.5
	LDA	3.0	0.6	0.3	6.7	1.7	0.9	9.3	12.7	6.2	3.1	13.3	7.8	5.1	10.2
	QDA	2.9	0.5	0.4	6.7	1.9	0.8	9.3	12.7	4.8	2.6	13.4	7.8	5.3	10.8
Bil	**LR**	6.3	0.3	0.1	6.9	1.5	0.6	12.6	13.4	7.6	2.7	14.0	9.6	5.5	15.5
	NB	6.3	0.3	0.1	6.7	1.5	0.6	10.7	13.5	7.6	2.8	14.0	9.6	5.6	15.8
	KNN	8.0	0.4	0.1	9.9	1.7	1.0	13.5	18.2	8.7	4.4	20.8	14.2	8.5	19.7
	LDA	6.4	0.3	0.1	6.9	1.6	0.6	12.4	13.3	7.7	3.7	14.0	9.4	5.5	15.3
	QDA	6.4	0.3	0.1	7.2	1.5	0.7	12.6	13.4	7.6	2.7	14.0	9.5	5.5	17.1

MI = mutual information; bil = bilateral filtered images; VS = vertebral shift; LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis

Table 11: Average misclassification errors after 100 trial runs using a single feature (MI) with and without bilateral filtering for the patient misalignment study. 4cm and 5cm results were excluded for the sake of brevity. All values are shown as a percentage. 95% confidence intervals ranged from 0.01%-0.2% and 0%-0.2% for TBeam and Tomo image pairs, respectively.

CC		TOMO							TBEAM						
		H&N			Pelvis			Spine	H&N			Pelvis			Spine
		1cm	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*	*1cm*	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*
No bil	**LR**	2.2	0.03	0	7.9	1.6	0.9	5.7	11.8	4.1	1.9	9.5	3.6	2.0	8.6
	NB	2.2	0.1	0	8.1	1.8	1.0	6.5	11.8	4.3	1.9	10.1	3.7	2.0	8.1
	KNN	3.7	0	0	11.0	2.9	1.1	7.8	16.4	6.7	2.7	14.7	6.0	3.3	13.3
	LDA	2.3	0	0	8.1	1.8	1.0	5.9	11.9	4.1	1.9	10.1	3.7	2.0	7.8
	QDA	2.3	0.02	0	8.0	1.8	1.0	5.9	11.9	4.3	1.9	10.1	3.7	2.0	8.0
Bil	**LR**	3.6	0.3	0	9.3	3.1	2.0	4.8	11.7	4.3	1.7	10.0	4.3	2.1	9.9
	NB	3.5	0.1	0	9.6	3.0	2.2	6.5	11.7	4.3	1.8	10.5	4.3	2.1	9.0
	KNN	3.6	0.1	0	14.6	3.9	3.0	6.3	18.8	7.2	2.9	14.9	6.1	3.1	14.6
	LDA	3.5	0	0	10.2	3.0	2.2	4.8	11.7	4.1	1.7	10.5	4.3	2.0	9.5
	QDA	3.4	0.07	0	10.2	3.0	2.1	4.9	11.7	4.0	1.9	10.5	4.4	2.1	9.5

CC = cross-correlation coefficient; bil = bilateral filtered images; VS = vertebral shift; LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis

Table 12: Average misclassification errors after 100 trial runs using a single feature (CC) with and without bilateral filtering for the patient misalignment study. 4cm and 5cm results were excluded for the sake of brevity. All values are shown as a percentage. 95% confidence intervals ranged from 0.01%-0.1% for all image pairs.

Similar to the patient identification study, TBeam image pairs produced average error values higher than Tomo image pairs. Bilateral filtering resulted in higher errors across all sites and metrics ($p<0.001$) except H&N with the CC metric ($p>0.05$). The SSIM metric resulted in the lowest

average MCE for H&N, while the CC metric had the lowest MCE for the pelvis and spine regions. KNN had the worst performance of the five classifiers. LR, LDA, and QDA had the best performance across all sites. MCE values for 4cm and 5cm shifts ranged from 0.7%-4.8% and 0.2%-3.7%.

SSIM		TOMO							TBEAM						
		H&N			Pelvis			Spine	H&N			Pelvis			Spine
		1cm	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*	*1cm*	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*
No bil	**LR**	2.0	0.4	0.3	5.3	0.9	0.8	5.8	9.3	2.8	1.7	10.3	4.1	2.5	11.4
	NB	2.0	0.5	0.2	5.1	0.8	0.7	6.8	9.4	2.9	1.8	10.4	4.2	2.6	10.1
	KNN	2.8	0.8	0.3	7.0	1.3	1.0	7.8	13.2	4.0	2.7	16.2	6.5	4.1	14.3
	LDA	2.0	0.5	0.4	5.6	1.0	0.6	7.2	9.1	3.2	2.6	10.1	5.2	3.1	10.4
	QDA	2.0	0.4	0.1	5.6	1.0	0.6	5.9	9.1	2.5	1.6	10.1	4.0	2.6	10.3
Bil	**LR**	4.3	0.6	0.1	11.5	3.6	1.6	7.9	11.8	3.8	1.9	13.2	5.6	2.4	13.4
	NB	3.9	1.0	0.01	11.8	3.6	1.6	8.9	11.9	3.8	2.0	12.9	5.6	2.4	13.2
	KNN	5.3	0.5	0.1	17.0	5.2	2.8	11.5	17.0	6.6	3.3	19.1	9.3	3.4	17.7
	LDA	3.9	1.0	0.1	11.4	3.6	2.3	8.8	11.8	3.7	2.1	13.3	5.6	2.5	13.2
	QDA	4.0	0.9	0.5	11.3	3.6	2.3	7.8	11.7	3.9	1.9	13.3	5.6	2.4	13.2

CC = cross-correlation coefficient; bil = bilateral filtered images; VS = vertebral shift; LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis

Table 13: Average misclassification errors after 100 trial runs using a single feature (SSIM) with and without bilateral filtering for the patient misalignment study. 4cm and 5cm results were excluded for the sake of brevity. All values are shown as a percentage. 95% confidence intervals ranged from 0%-0.2% and 0%-0.1% for TBeam and Tomo image pairs, respectively.

Figures 36-39 show 1D histograms of CC and SSIM frequency distributions for the H&N and pelvis sites.

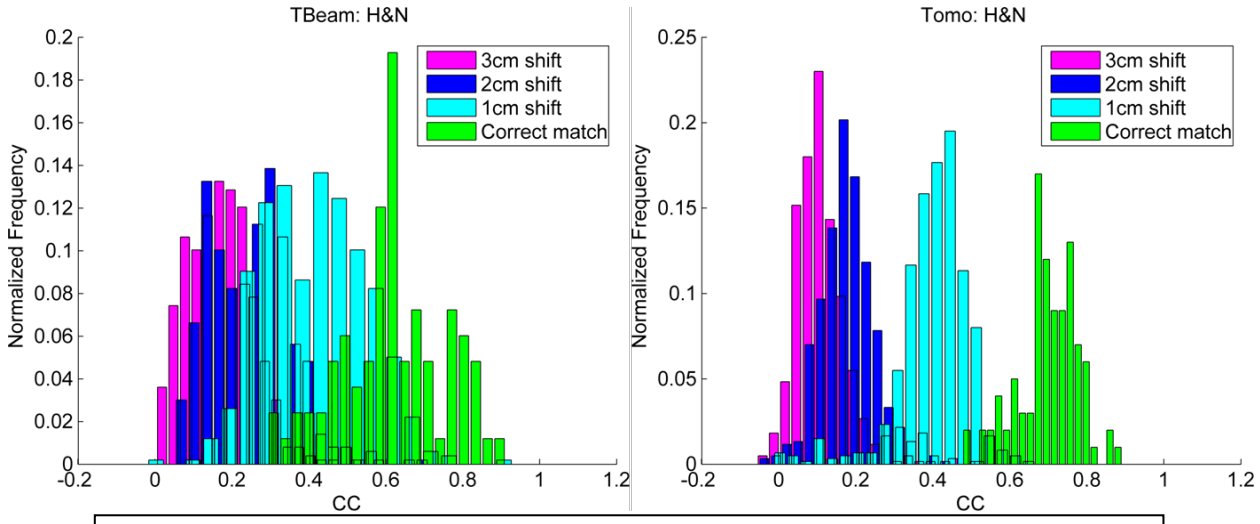


Figure 36: CC frequency histograms of H&N shifts for the TBeam and Tomo machines. (CC = cross-correlation coefficient)

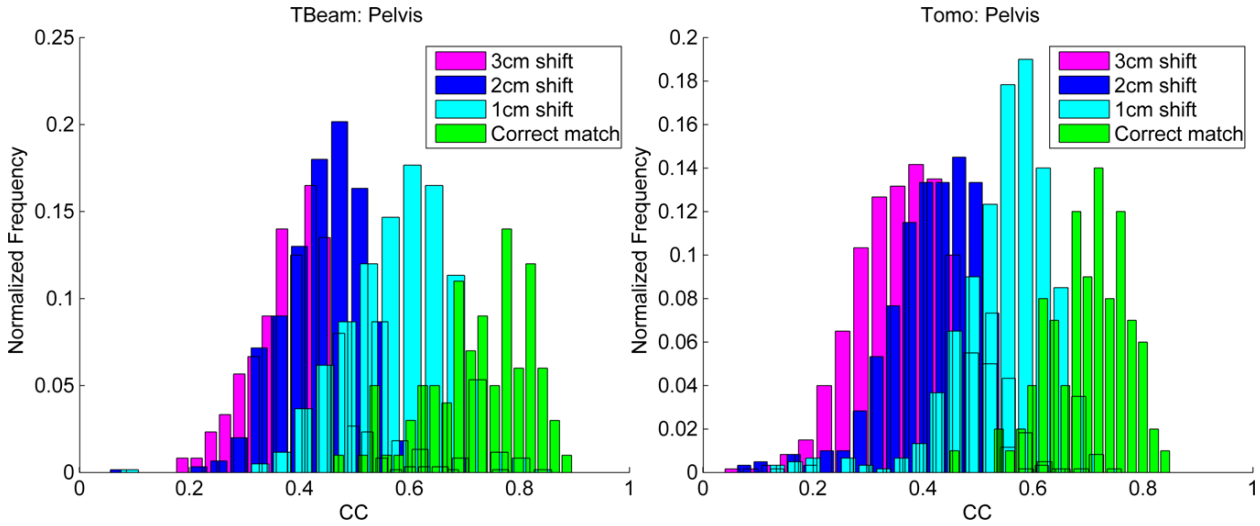


Figure 37: CC frequency histograms of pelvis shifts for the TBeam and Tomo machines. (CC = cross-correlation coefficient)

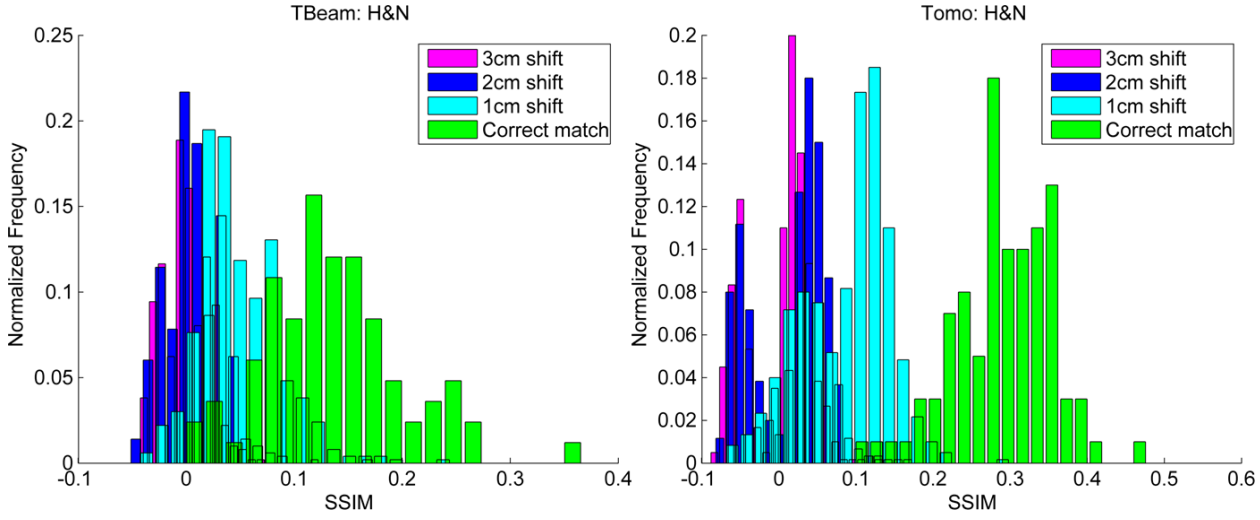


Figure 38: SSIM frequency histograms of H&N shifts for the TBeam and Tomo machines. (SSIM = structural similarity)

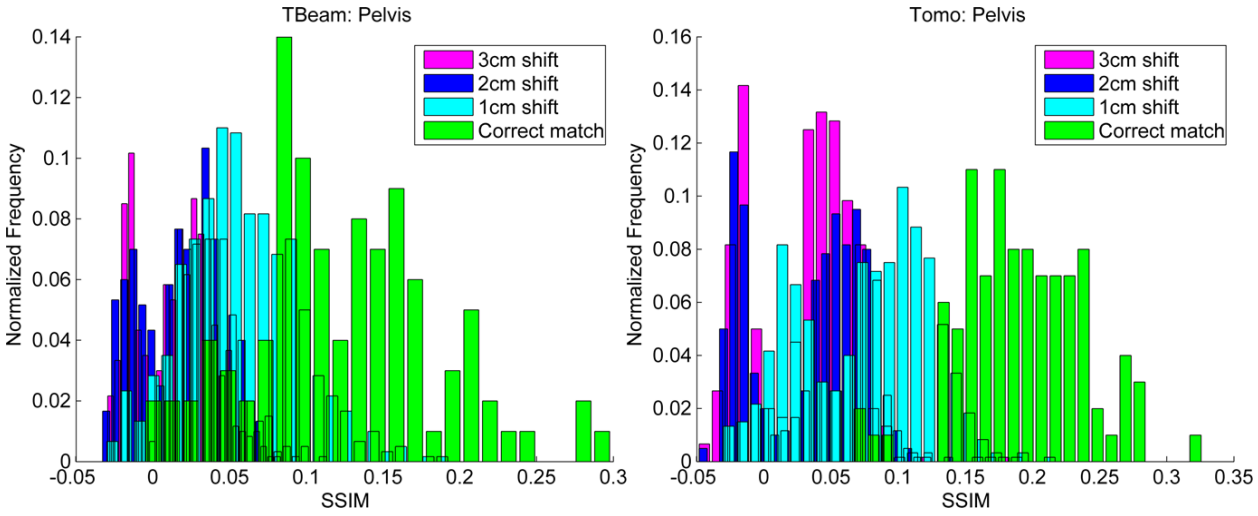


Figure 39: SSIM frequency histograms of pelvis shifts for the TBeam and Tomo machines. (SSIM = structural similarity)

Figures 40 and **41** display histograms for vertebral misalignments across both treatment machines.

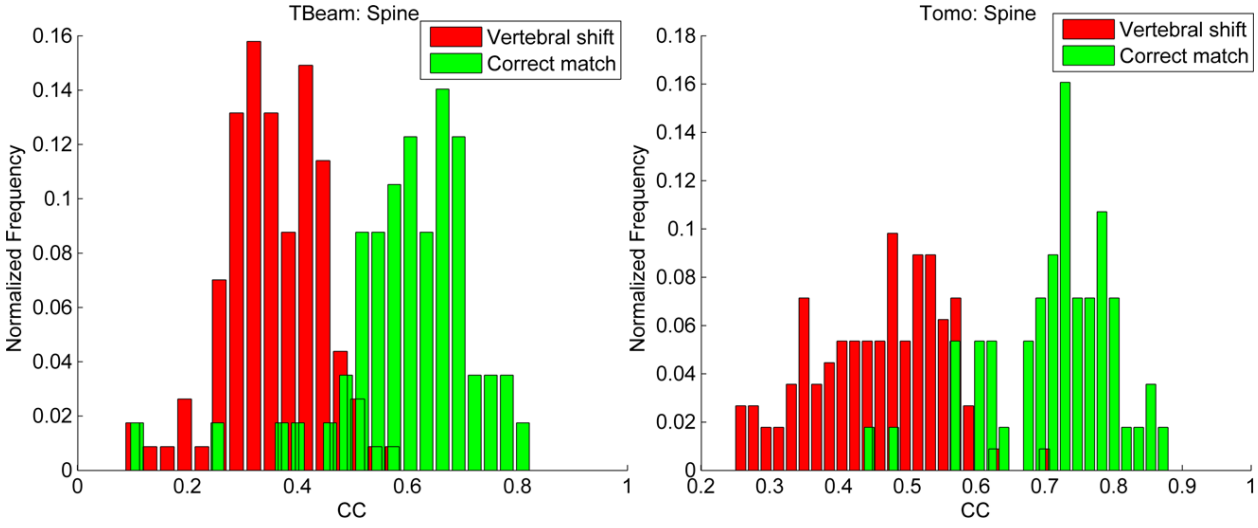


Figure 40: CC frequency histograms of vertebral shifts for the TBeam and Tomo machines. (CC = cross-correlation coefficient)

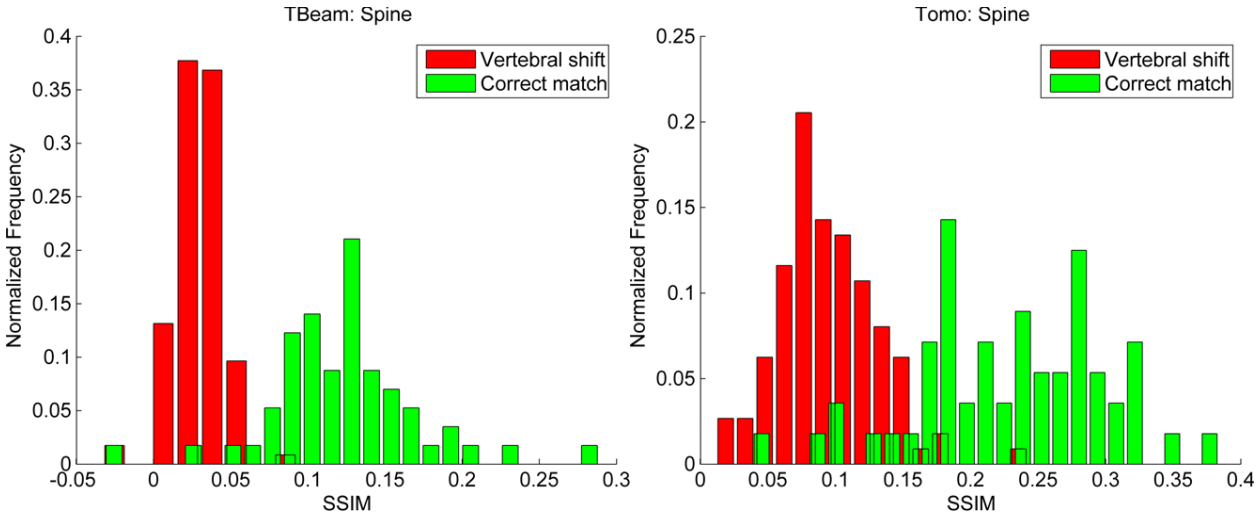


Figure 41: SSIM frequency histograms of vertebral shifts for the TBeam and Tomo machines. (SSIM = structural similarity)

Classifiers were retrained by using all three MI/CC/SSIM features, as seen in **Table 14**. Without bilateral filtering, reductions in MCE were observed for Tomo image pairs for the H&N, pelvis, and spine sites across all classification algorithms ($p<0.001$). Similar reductions were seen in spine and pelvis sites for TBeam images ($p<0.001$), and H&N had variable changes in MCE (significant reduction for KNN ($p<0.001$), significant increase for NB ($p<0.001$), and insignificant changes for LR, LDA, and QDA ($p>0.05$)). Similar to the single feature results, bilateral filtering had a generally detrimental effect on accuracy for H&N and pelvis images. For TBeam spine images, bilateral filtering resulted in a 3.8% increase in error across all algorithms ($p<0.001$). Significant improvements were observed for the KNN, LDA, and QDA algorithms in Tomo spine images ($p<0.001$). Despite variable classifier performance, LR and LDA/QDA generally produced smaller errors than the KNN and NB algorithms. Errors no greater than 0.003% and 1.1% were produced for 4cm H&N and pelvis shifts, respectively, for Tomo images; upper error bounds of 1.8% and 1.6% were achieved for the same sites for TBeam images, respectively.

MI+CC +SSIM		TOMO							TBEAM						
		H&N			Pelvis			Spine	H&N			Pelvis			Spine
		1cm	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*	*1cm*	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*
No bil	**LR**	1.6	0.04	0.01	4.7	1.1	1.0	5.5	9.7	2.9	1.6	8.0	3.3	1.6	9.0
	NB	1.4	0	0	5.1	1.3	0.9	5.0	10.2	3.9	1.7	11.9	3.9	2.4	6.4
	KNN	1.6	0.2	0	5.0	1.0	0.9	8.3	10.8	3.1	1.6	9.5	3.6	1.7	6.9
	LDA	1.6	0.3	0.2	4.3	1.6	1.3	6.4	9.3	2.9	1.8	8.5	3.0	1.5	7.5
	QDA	1.5	0	0	4.4	1.5	0.9	5.3	9.5	3.4	1.9	8.6	3.2	2.2	6.1
Bil	**LR**	2.5	0.2	0.003	7.1	1.7	0.7	4.8	10.4	3.9	1.7	9.0	3.6	1.8	11.6
	NB	2.7	0.1	0	7.6	1.6	0.5	4.9	12.7	3.9	1.5	13.2	4.3	2.3	11.6
	KNN	3.6	0.1	0	7.5	1.5	0.8	4.1	11.8	4.8	1.4	11.7	4.1	1.9	10.9
	LDA	2.7	0.1	0.1	7.4	1.6	0.6	5.0	10.6	3.7	1.4	9.3	3.3	1.4	10.0
	QDA	2.9	0.01	0	7.4	1.5	1.0	4.8	11.1	3.8	1.3	9.2	3.4	1.6	10.8

MI = mutual information; bil = bilateral filtered images; VS = vertebral shift; LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis

Table 14: Average misclassification errors after 100 trial runs using three features (CC/MI/SSIM) with and without bilateral filtering for the patient misalignment study. 4cm and 5cm results were excluded for the sake of brevity. All values are shown as a percentage. 95% confidence intervals ranged from 0.01%-0.2% and 0%-0.2% for TBeam and Tomo image pairs, respectively.

4.4: Patient alignment: including gradient-based features

Tables 15 summarizes classification results from the inclusion of the gradient-based metrics for the Tomo and TrueBeam machines.

		TOMO							TBEAM						
		H&N			Pelvis			Spine	H&N			Pelvis			Spine
		1cm	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*	*1cm*	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*
No bil	**LR**	1.7	0.2	0.001	5.2	2.0	1.4	6.6	9.9	3.3	2.0	8.6	3.3	1.6	4.4
	NB	2.1	0.4	0	7.4	1.7	1.0	6.7	13.4	4.1	2.0	15.2	3.9	2.1	3.9
	KNN	2.4	0.5	0.009	6.9	1.4	0.8	7.3	12.8	3.7	2.2	11.8	3.2	1.6	3.6
	LDA	1.3	0.4	0	5.2	1.6	0.9	4.6	9.1	3.2	2.1	8.7	3.2	1.7	5.1
	QDA	2.2	0.4	0.1	13.0	3.7	2.2	10.0	13.2	5.6	3.1	17.6	4.9	1.7	3.6
Bil	**LR**	2.3	0.3	0.01	7.8	2.1	1.8	6.8	10.6	4.2	1.6	9.4	3.3	1.6	3.1
	NB	3.0	0.6	0.01	10.0	2.3	1.2	6.5	15.2	4.5	2.1	16.3	3.9	2.4	3.3
	KNN	3.2	0.6	0.2	9.1	2.5	0.9	7.3	14.9	4.2	2.1	14.8	3.5	1.5	3.6
	LDA	2.2	0.3	0	8.3	2.3	1.0	4.2	10.3	3.6	1.9	9.6	4.1	1.6	4.3
	QDA	3.9	0.4	0.1	19.8	4.0	1.9	10.1	17.1	6.0	3.0	22.5	5.1	1.9	4.0

MI = mutual information; bil = bilateral filtered images; VS = vertebral shift; LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis

Table 15: Average misclassification errors after 100 trial runs using the gradient-based features in addition to the CC/MI/SSIM features (with and without bilateral filtering) for the patient misalignment study. 4cm and 5cm results were excluded for the sake of brevity. All values are shown as a percentage. 95% confidence intervals ranged from 0.01%-0.2% and 0%-0.2% for TBeam and Tomo image pairs, respectively.

Inclusion of the gradient-based features had a mixed effect on classifier accuracy. For 1cm Tomo shifts, H&N accuracy improved from 1.6% to 1.3% ($p<0.05$) using LDA, did not change using LR ($p>0.05$), and decreased for the NB, KNN, and QDA algorithms ($p<0.001$). Accuracy for 1cm pelvis shifts decreased across all algorithms ($p<0.001$). Vertebral shifts improved in accuracy using LDA ($p<0.001$), but decreased for all other algorithms ($p<0.001$). For TBeam images, there were no significant differences for 1cm H&N shifts using LR or LDA ($p>0.05$), but significant MCE decreases using NB, KNN, and QDA ($p<0.001$). For 1cm pelvis shifts, all algorithms except LDA ($p>0.05$) showed a decrease in accuracy ($p<0.001$). Spine shifts improved for LDA and decreased in accuracy for the other four classifiers for Tomo images ($p<0.001$), while accuracy improved across all algorithms for TBeam images ($p<0.001$). Pelvis images that underwent bilateral filtering had a decreased accuracy across both imaging modalities and all algorithms ($p<0.001$). There were few important differences between filtered H&N images including or excluding the gradient-based features. Tomo spine images decreased in accuracy (except for LDA ($p<0.001$)), while the accuracy of TBeam spine images improved drastically by an average of 7.3% ($p<0.001$).

Detailed shift results are summarized in **Tables 16** and **17**, including the sensitivity, specificity, MCC, LR+, and LR- parameters. Results are primarily shown for the MI/CC/SSIM metrics; cases when inclusion of the gradient-based features resulted in a significantly lower MCE are explicitly denoted. For H&N Tomo image pairs, there were minor differences between LDA results that included or excluded the gradient-based features. LDA also produced similar results for H&N and pelvis shifts ≥2cm for Tbeam image pairs.

TOMO		H&N			Pelvis			Spine
		1cm	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*
LR	*MCE*	1.7±0.02	0.03±0.02	0.01±7e-3	4.7±0.03	1.1±0.01	1.0±0.02	5.5±0.08
	Sens	99.5±0.01	99.9±9e-2	100±0	98.6±0.02	99.5±9e-3	99.6±0.02	97.8±0.1
	Spec	91.5±0.1	99.9±0.1	99.9±0.06	75.7±0.1	95.0±1e-5	95.1±0.05	87.7±0.1
	MCC	0.93±9e-4	0.99±7e-4	1±0	0.80±1e-3	0.95±3e-4	0.96±7e-4	0.88±2e-3
	LR+	12±0.2	9303±500	8812±633	4.1±0.02	20±2e-3	21±0.3	8.0±0.09
	LR-	0.006±1e-4	2e-4±9±-5	1e-4±0	0.02±3e-4	0.005±9e-5	0.005±2e-4	0.02±1e-3
	AUC	0.993±2e-4	0.999±2e-6	1±0	0.960±2e-4	0.992±5e-4	0.996±5e-5	0.977±8e-4
NB	*MCE*	1.4±0.02	0±0	0±0	5.1±0.03	1.3±3e-3	0.9±0	5.0±0.08
	Sens	99.2±0.05	99.9±9e-3	100±0	96.4±0.1	99.2±0.05	99.5±0.04	95.9±0.3
	Spec	96.4±0.3	99.9±0.04	100±0	89.2±0.7	96.8±0.2	97.6±0.2	93.8±0.4
	MCC	0.96±2e-3	1±0	1±0	0.85±6e-3	0.96±2e-3	0.97±1e-3	0.9±2e-3
	LR+	32±2.7	9801±274	10000±0	9.9±0.6	40±5.7	56±8.2	19±2.1
	LR-	0.009±5e-4	1e-4±9e-5	1e-4±0	0.04±9e-4	0.008±5e-4	0.005±4e-4	0.04±3e-3
	AUC	0.991±3e-4	1±0	1±0	0.953±3e-4	0.994±3e-4	0.999±2e-4	0.985±7e-4
KNN	*MCE*	1.6±0.03	0.2±0.01	0±0	5.0±0.05	1.0±0.01	0.9±0.01	7.3±0.1
	Sens	99.1±0.06	99.9±0.02	100±0	97.6±0.2	99.5±0.05	99.7±0.04	94.9±0.5
	Spec	95.7±0.3	99.0±0.1	100±0	84.6±1.3	97.0±0.2	97.1±0.3	89.5±0.8
	MCC	0.95±2e-3	0.99±5e-4	1±0	0.85±7e-3	0.97±1e-3	0.97±1e-3	0.85±3e-3
	LR+	28±2.6	521±381	10000±0	7.8±0.8	41±6.2	44±5	11±1.0
	LR-	0.009±6e-4	6e-4±2e-4	1e-4±0	0.03±2e-3	0.005±5e-4	0.003±4e-4	0.05±5e-3
	AUC	0.983±3e-5	0.999±1e-4	1±0	0.942±6e-4	0.973±1e-5	0.989±5e-4	0.965±1e-3
LDA	*MCE*	1.3±0.01	0.4±0.01	0±0	4.3±0.02	1.6±0.01	1.3±0.02	4.6±0.1
	Sens	99.7±0.04	99.8±0.02	100±0	98.4±0.1	99.3±0.07	99.6±0.06	96.2±0.3
	Spec	95.1±0.4	98.8±0.1	100±0	84.4±1.3	94.6±0.5	95.4±0.4	94.0±0.5
	MCC	0.96±2e-3	0.99±6e-4	1±0	0.87±7e-3	0.95±3e-3	0.96±2e-3	0.91±3e-3
	LR+	25±2.2	201±194	10000±0	7.6±0.7	24±2.8	28±3.2	19±1.9
	LR-	0.004±5e-4	0.002±2e-4	1e-4±0	0.02±1e-3	0.007±8e-4	0.004±6e-4	0.04±3e-3
	AUC	0.993±1e-4	0.999±7e-7	1±0	0.952±2e-4	0.992±1e-4	0.997±6e-4	0.986±5e-4
QDA	*MCE*	1.5±0.02	0±0	0±0	4.4±0.02	1.5±0.01	0.9±0.01	5.3±0.07
	Sens	99.3±0.06	100±0	100±0	98.4±0.1	99.2±0.05	99.6±0.04	95.8±0.3
	Spec	95.2±0.4	100±0	100±0	84.3±1.2	95.7±0.3	97.4±0.2	93.3±0.4
	MCC	0.96±2e-3	1±0	1±0	0.97±6e-3	0.95±2e-3	0.97±1e-3	0.89±1e-3
	LR+	25±2.7	10000±0	10000±0	7.3±0.6	26±2.0	146±195	17±1.7
	LR-	0.007±6e-4	1e-4±0	1e-4±0	0.02±1e-3	0.008±5e-4	0.004±4e-4	0.04±3e-3
	AUC	0.997±8e-4	1±0	1±0	0.959±2e-4	0.994±1e-4	0.999±4e-5	0.975±4e-4

VS = vertebral shift; LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis; MCE = misclassification error; sens = sensitivity; spec = specificity; MCC = Matthew's correlation coefficient; LR+ = positive likelihood ratio; LR- = negative likelihood ratio; AUC = area under the curve

Table 16: Average misclassification errors for misaligned Tomo image pairs after 100 trial runs. 4cm and 5cm results were excluded for the sake of brevity. Shaded cells refer to results generated from using gradient-based features and the MI/CC/SSIM metrics; results in non-shaded cells used just the MI/CC/SSIM metrics. All values are shown as percentages.

TBEAM		H&N			Pelvis			Spine
		1cm	*2cm*	*3cm*	*1cm*	*2cm*	*3cm*	*VS*
LR	*MCE*	9.7±0.04	2.9±0.03	1.6±0.03	8.0±0.04	3.3±0.02	1.6±0.03	4.4±0.06
	Sens	97.9±0.03	98.8±0.03	99.4±0.02	98.2±0.03	98.3±0.02	99.2±0.03	95.1±0.2
	Spec	45.0±0.2	86.2±0.2	92.5±0.2	54.8±0.2	87.1±0.06	93.8±0.2	95.8±0.04
	MCC	0.55±2e-3	0.88±1e-3	0.93±1e-3	0.63±2e-3	0.86±06e-4	0.94±1e-3	0.90±1e-3
	LR+	1.8±7e-3	7.2±0.09	14±0.4	2.2±0.01	7.6±0.04	16±0.4	23±0.2
	LR-	0.05±7e-4	0.01±3e-4	0.007±2e-4	0.03±5e-4	0.02±2e-4	0.008±2e-4	0.05±2e-3
	AUC	0.892±3e-4	0.985±2e-4	0.997±1e-4	0.875±2e-4	0.987±1e-4	0.997±7e-5	0.986±4e-4
NB	*MCE*	10.2±0.04	3.9±0.04	1.7±0.02	11.9±0.04	3.9±0.02	2.4±0.03	3.9±0.05
	Sens	92.3±0.2	97.1±0.1	98.6±0.06	92.2±0.2	97.6±0.1	98.2±0.07	95.6±0.2
	Spec	79.9±1.3	92.9±0.4	97.4±0.2	72.0±1.9	89.9±0.7	95.3±0.3	96.8±0.2
	MCC	0.71±0.01	0.89±5e-3	0.95±2e-3	0.64±0.02	0.89±5e-3	0.93±3e-3	0.92±1e-3
	LR+	5.1±0.3	15±1.3	47±6.4	3.7±0.2	11±0.9	24±1.8	37±5
	LR-	0.1±2e-3	0.03±1e-3	0.01±6e-4	0.1±2e-3	0.03±9e-4	0.02±7e-4	0.04±2e-3
	AUC	0.871±3e-4	0.980±2e-4	0.996±6e-5	0.833±3e-4	0.973±2e-4	0.993±1e-4	0.981±7e-4
KNN	*MCE*	10.8±0.09	3.1±0.04	1.6±0.03	9.5±0.06	3.6±0.04	1.7±0.03	3.6±0.04
	Sens	93.6±0.3	98.2±0.1	99.2±0.05	95.5±0.2	97.6±0.1	99.1±0.07	96.2±0.2
	Spec	72.5±2.1	91.6±0.6	95.7±0.3	70.6±2.2	91.9±0.5	95.0±0.4	96.9±0.2
	MCC	0.67±0.02	0.91±4e-3	0.95±2e-3	0.71±0.01	0.90±5e-3	0.95±2e-3	0.93±1e-3
	LR+	3.9±0.3	14±1.2	27.7±2.7	3.8±0.3	13±0.9	24±2.2	36±3.4
	LR-	0.09±3e-3	0.02±1e-3	0.008±5e-4	0.06±2e-3	0.03±9e-4	0.009±7e-4	0.04±2e-3
	AUC	0.841±2e-3	0.972±1e-3	0.980±4e-4	0.832±1e-3	0.960±8e-4	0.979±4e-4	0.983±3e-4
LDA	*MCE*	9.3±0.05	2.9±0.01	1.8±0.02	8.5±0.04	3.0±0.01	1.5±0.01	5.1±0.07
	Sens	96.0±0.3	99.0±0.09	99.4±0.06	96.2±0.3	98.5±0.08	99.6±0.06	94.5±0.2
	Spec	70.7±2.3	89.4±0.9	93.1±0.6	75.1±2.1	91.3±0.7	94.2±0.5	95.8±0.3
	MCC	0.71±0.02	0.91±4e-3	0.95±2e-3	0.75±0.01	0.91±4e-3	0.96±2e-3	0.90±2e-3
	LR+	3.8±0.3	11±1.0	17±1.6	4.6±0.4	13±1.0	21±2.0	26±3.0
	LR-	0.06±2e-3	0.01±9e-4	0.007±6e-4	0.05±2e-3	0.02±8e-4	0.005±6e-4	0.06±2e-3
	AUC	0.892±3e-4	0.986±1e-4	0.997±6e-5	0.869±2e-4	0.986±2e-4	0.997±6e-5	0.980±2e-4
QDA	*MCE*	9.5±0.04	3.4±0.01	1.9±0.02	8.6±0.03	3.2±0.01	2.2±9e-3	3.6±0.04
	Sens	95.7±0.3	97.6±0.08	98.2±0.06	96.3±0.3	97.8±0.07	98.5±0.06	96.5±0.2
	Spec	69.8±2.2	92.7±0.4	97.4±0.2	72.2±2.3	92.8±0.5	95.6±0.3	96.3±0.2
	MCC	0.70±0.02	0.90±4e-3	0.94±2e-3	0.74±0.01	0.91±4e-3	0.94±3e-3	0.93±1e-3
	LR+	3.7±0.3	15±1.0	44±4.0	4.2±0.4	15±1.1	26±2.1	29±1.8
	LR-	0.06±2e-3	0.03±7e-4	0.02±6e-4	0.05±2e-3	0.02±6e-4	0.02±6e-4	0.04±2e-3
	AUC	0.879±4e-4	0.984±2e-4	0.996±4e-5	0.871±3e-4	0.986±1e-4	0.997±4e-5	0.984±4e-4

VS = vertebral shift; LR = logistic regression; NB = Naïve Bayes; KNN = k-nearest neighbor; LDA = linear discriminant analysis; QDA = quadratic discriminant analysis; MCE = misclassification error; sens = sensitivity; spec = specificity; MCC = Matthew's correlation coefficient; LR+ = positive likelihood ratio; LR- = negative likelihood ratio; AUC = area under the curve

Table 17: Average misclassification errors for misaligned TBeam image pairs after 100 trial runs. 4cm and 5cm results were excluded for the sake of brevity. Shaded cells refer to results generated from using gradient-based features and the MI/CC/SSIM metrics; results in non-shaded cells used just the MI/CC/SSIM metrics. All values are shown as percentages.

ROC curves for 1-3cm shifts in H&N and pelvis sites are shown in **Figures 42-45**, and vertebral shifts are shown in **Figure 46**. All curves were generated from results from the LDA classifier, using features as described from the shaded/non-shaded cells in **Tables 16-17**.

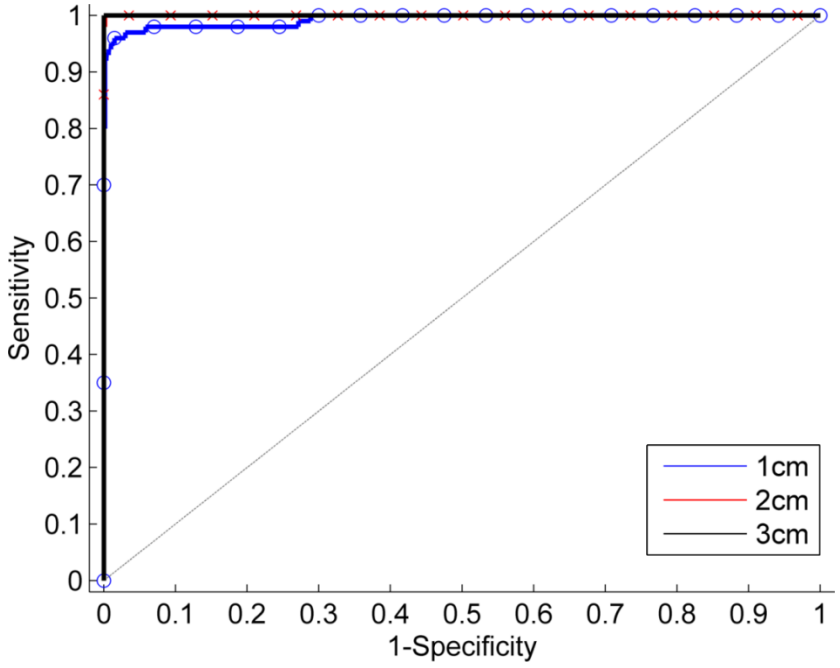


Figure 42: ROC curves for 1-3cm H&N shifts on the Tomo machine.

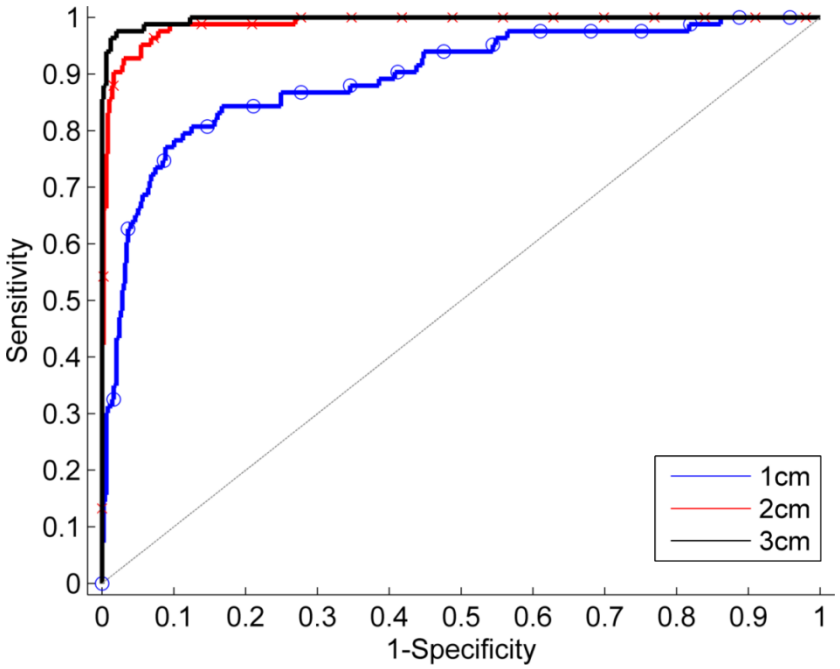


Figure 43: ROC curves for 1-3cm H&N shifts on the TBeam machine.

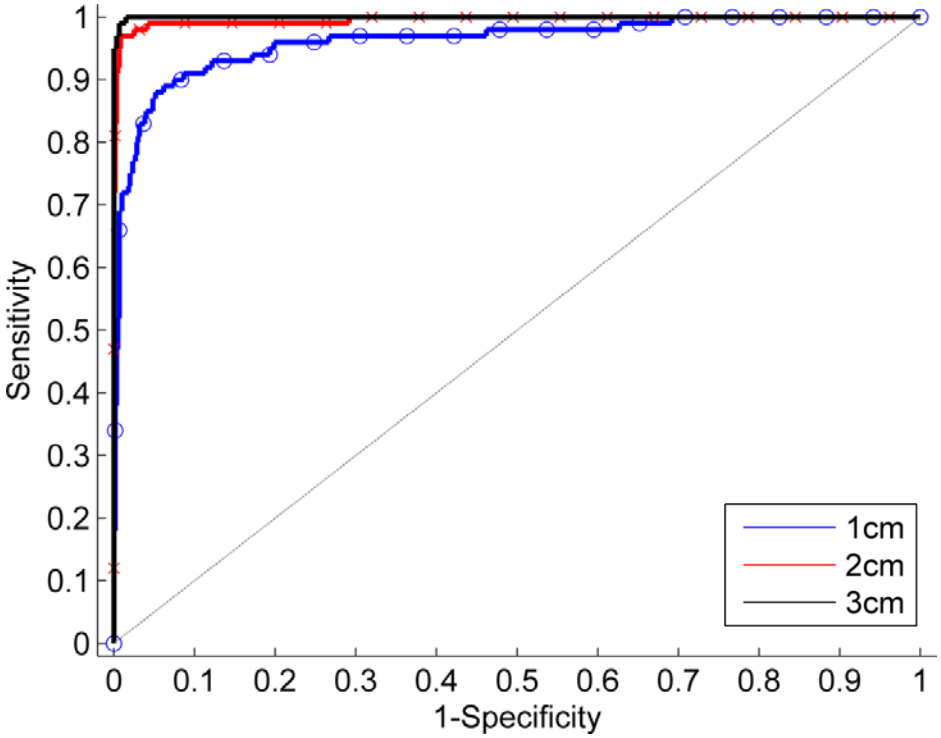


Figure 44: ROC curves for 1-3cm pelvis shifts on the Tomo machine.

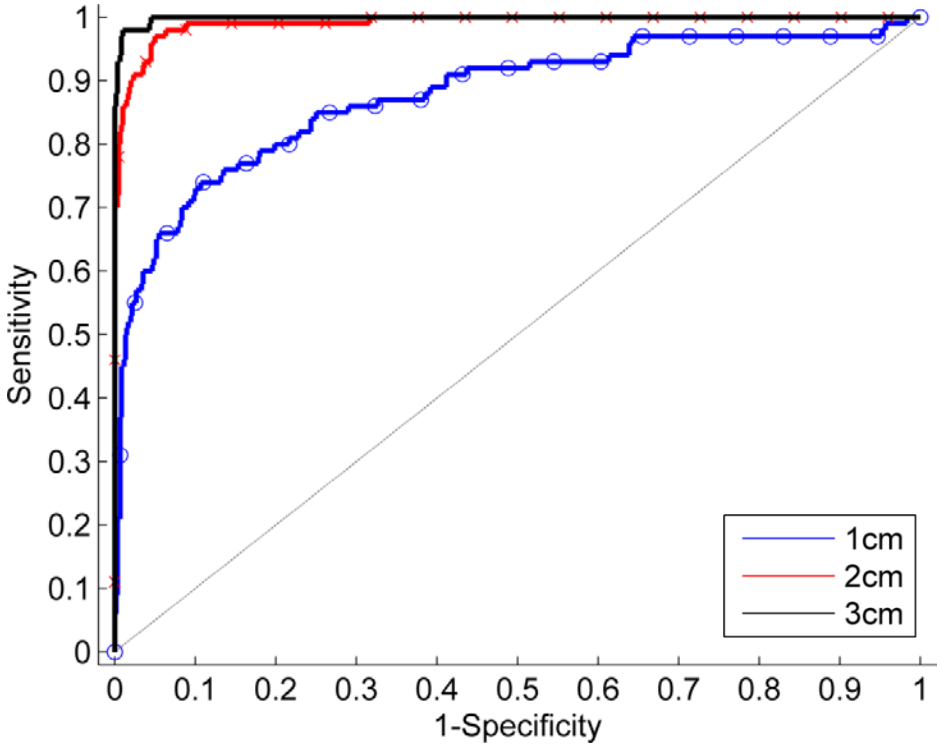


Figure 45: ROC curves for 1-3cm pelvis shifts on the TBeam machine.

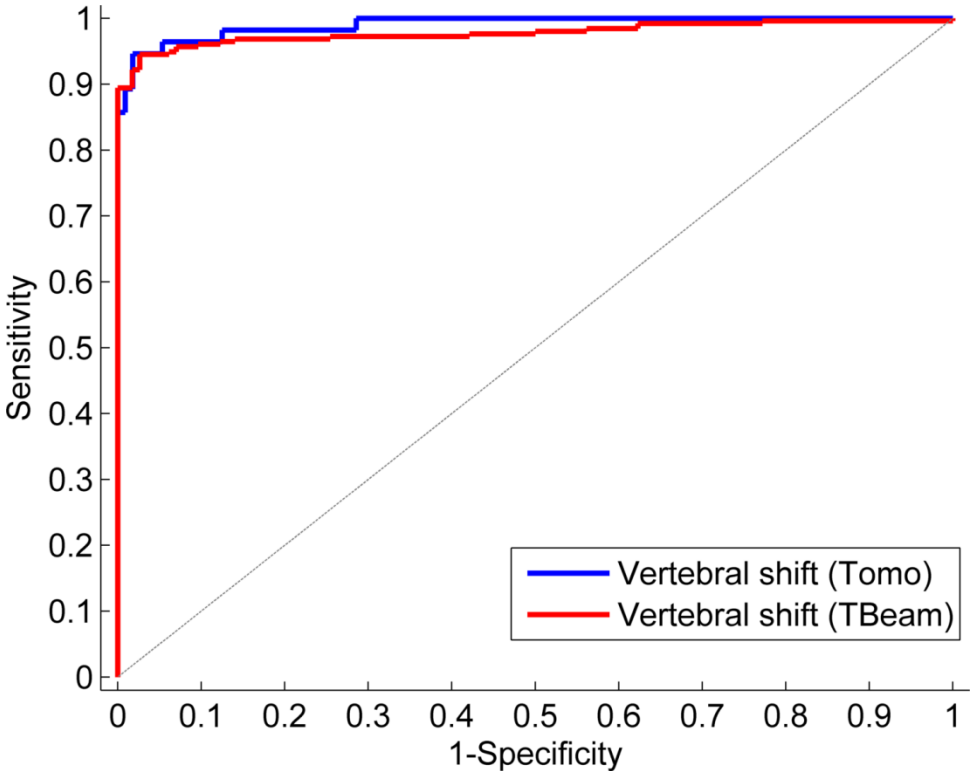


Figure 46: ROC curves for vertebral shifts on both machines.

CHAPTER 5: DISCUSSION, FUTURE STUDIES, AND CONCLUDING THOUGHTS

5.1: Discussion: workflow development

Our study involved the comparison of 3D planning kVCT images to 3D setup images acquired using two imaging modalities with kV and MV energies. As the pilot study to this work involved a 2D-3D matching (planar x-ray images compared to 2D DRRs of the 3D planning kVCT) [61], a 3D-3D image comparison was a natural extension. It is also possible to consider the case of 2D-2D matching, as many centers continue to use 2D setup imaging from integrated on-board imagers or from MV portal films. We would anticipate greater difficulties with this type of comparison, as 2D projections would require a more exact alignment in order to produce accurate RP matches. Any differences in projection angle between both 2D images could cause occlusion of regions that could degrade the resulting image similarity. Patient deformation could also cause a similar effect. The benefit of the pilot study [61] was the presence of a built-in registration system and a gradient-based correlation coefficient metric as part of the ExacTrac system [64]. A similar algorithm could be implemented for other planar image acquisitions to remove some of the aforementioned concerns with direct 2D-2D matching. There is also the potential of implementing the proposed technique with MR-based images – this could be especially useful for sites with poor soft-tissue contrast.

In selecting image pairs during the data acquisition step, we chose setup images that were closest in acquisition time to the planning kVCT. Over the course of a patient's RT, a patient could potentially undergo significant physiological changes in the treatment region (weight gain/loss, tumor shrinkage, etc.). As such, it was important to select image pairs that best reflected the same patient anatomy for error detection purposes. Although drastic changes in patient anatomy would likely necessitate reacquisition of the planning kVCT and a subsequent replan of the remaining treatment (i.e. adaptive radiotherapy), some preliminary analysis and a strategy is proposed in section 5.5 to address this issue.

To create a systematic approach for manually fusing image pairs, we devised a generalized set of guidelines based on the experiences of several therapists and oncologists in our department (see section 2.3). It is unlikely that all patients undergoing treatment would be strictly aligned using these criteria due to differences in the exact treatment site or real-time adjustments to address minor patient deformations. Natural internal organ motion due to patient breathing, deformation, or bowel movement can potentially cause large displacements of soft tissue. For example, organ motion has been found to cause prostate displacement exceeding 1cm in some cases [175]. Due to the nature of our large dataset, it was necessary to make assumptions about image matching when the registration file was unavailable (i.e. all WP matches and Tomo RP matches). We chose to perform manual image alignment over the use of any automatic registration algorithms. After conferring with several therapists at our institution, it was evident that most avoided the built-in registration software on the treatment console due to a variety of reasons, but primarily because of trusting their own judgment more than that of the registration software. Some therapists stated they used the automatic registration as a preliminary step in the fusion process, but always performed the final matching themselves.

Our experimental design included the development of workflows to address both identification and alignment errors. We assumed that both error types would be present prior to the IGRT image acquisition. For alignment errors, we looked at 1D misalignments ranging from 1cm to 5cm based on reported errors we discovered in our institution's database. There is the potential for 2D errors (i.e. misalignment in more than one anatomical direction), which would be useful to explore as part of a future study. We would expect these errors to be easier to detect than the 1D errors due to the introduction of multi-dimensional dissimilarities between the planning and setup images. In addition to mistranslations, another potential error type related to patient alignment is patient misrotation. In our experience, gross misrotations of the order of ≥10° are much less frequently observed than mistranslations. In our review of reported errors since 2009 at our clinic,

we found approximately 10 times more reported mistranslations than misrotations. As such, we decided to focus the study on mistranslations due to its larger prevalence in clinical practice. Subtle misrotations can occur frequently, particularly in conjunction with deformations. We have been studying this effect in a separate study, where image similarity comparisons are used to predict pelvic nodal coverage in high-risk prostate adenocarcinoma patients [176].

As part of the workflow development, we excluded voxels corresponding to air in all planning kVCT and setup images. We found that bowel gas, lung motion, and maxillary sinusitis introduced dissimilarities between correct patient image pairs and degraded the discriminating power of the image similarity metrics. This was confirmed by higher MCEs when running the workflow with inclusion of air voxels. Eliminating high-HU voxels (corresponding to implants composed of a high-density material) achieved the same effect. As described in 2.4, we also found a large benefit to using a patient-specific body contour as a preliminary mask. Assuming no external markers, areas outside the patient body have little importance with respect to positioning and identification error detection. We utilized a global approach by starting with the patient's entire body, then eliminating select voxels before calculating image similarity. Alternative approaches are discussed in section 5.5.

We chose a small subset of commonly-used metrics to assess image similarity. We found that the CC metric had the best overall performance, although SSIM also performed well for certain anatomical sites (**Table 13**). Combining the MI/CC/SSIM features together for classification produced superior results compared to using each individually, likely due to additional complementary information used to define the classifier decision boundary. Other metrics have been developed and successfully used in the literature for various applications requiring image similarity comparisons. One extension to cross-correlation is normalized cross-correlation, which can better account for variable exposure conditions between two images [177]. Cross-cumulative residual entropy is a multimodal measure that is defined on cumulative distributions rather than

probability densities, as is the case with MI [178]. Another multimodal measure is the sum of conditional variance – although it is also derived from joint probability distributions like MI, it is less computationally expensive [179]. Using extra measures that fundamentally assess similarity in different ways would likely add the most value as additional features for classification model development.

We chose to use a bilateral filter as one of the pre-processing steps prior to the gradient-based metric. Although Gaussian filters are commonly used in image processing for smoothing and noise reduction, the benefit of a bilateral filter is the additional intensity kernel that allows for sharp edge preservation. The next pre-processing step, the Sobel gradient operator, emphasizes these sharp edges and was the basis for image similarity assessment with the custom-designed metrics. A smoothing filter is essential prior to this step in order to avoid any large gradients as a result of noisy voxels. After both pre-processing steps, we tested various thresholds ranging from 1,000 to 10,000 intensity units to eliminate smaller gradients and found a threshold of 5,000 to be a suitable value across all treatment sites. Too low of a threshold resulted in inclusion of extra voxels that did not correlate to edges of interest, and too high of a threshold resulted in too few voxels used for image comparison. Given the large anatomical differences between different regions of the body, there are inherently a variable number of large gradient interfaces depending on the exact image FOV. Exploring the use of site-specific thresholds could be useful in developing more suitable masks. In addition, other filtering kernels could also be explored to allow for superior image filtering. The guided image filter is similar to the bilateral filter, but has been reported to have better behavior near edges and is computationally efficient in many processing applications [180].

Feature and subset selection is an essential step in developing a classification model, as the choice of features plays a major role in the modeling process and the model outcomes. Feature selection is commonly used in datasets with a large number of predictors, where it is desirable to determine a smaller subset. This introduces less variance in the feature set, allows for easier

interpretability, and lets the user better understand the domain of the problem [146]. In addition, some classifiers may not properly scale to a full-sized feature set, subsequently reducing the accuracy of algorithm performance. It is also computationally less expensive to use a smaller set of features, resulting in faster model training times. Finally, the use of too many features can result in model overfitting, or the process of inferring more structure from the training set than what actually may exist from the population [131], which can produce large and inaccurate error rates.

We used several features in our study that provided complementary pieces of information, including single-value outputs (i.e. MI/CC/SSIM) and vector outputs (i.e. gradient-based methods). From the vector outputs, we extracted several values to be used as features for classification: mean, max, third and fourth moments, and percentiles ranging from 5^{th} to 95^{th} in 5% increments. Many of these values are collinear, which has been shown to be problematic in linear or generalized linear models by inflating the variance of model coefficients, thereby producing unstable models and error estimates [181-183]. PCA was used to address this issue by creating a new set of features (i.e. principal components) that are linear combinations of the given features. The highest eigenvalues and their corresponding eigenvectors were then chosen for the final PCA-reduced gradient-based feature set. The smallest eigenvalues/eigenvectors represent a very small proportion of the dataset's variance, and therefore may be unstable and imprecise estimates for the population measures [184]. By selecting the components that capture the majority of the dataset's variance, the multicollinearity problem is addressed by eliminating components with little predictive power (i.e. smaller eigenvalues/eigenvectors) that may add noise to the discriminant equation. Increasing the percentile sampling count in extracting the gradient-based features would produce a larger number of collinear features; although PCA would address this, it is unlikely that increased sampling would result in improved model estimates in this study. An important consideration is the number of features relative to the number of observations in the dataset – it is generally recommended to have at least 5-10x data points than features [185].

One additional technique that we explored was the transformation of data prior to classification. If a given data distribution is skewed, applying a data transformation can create a more symmetric distribution that could potentially improve the performance of a classifier that relies on assumptions of linearity or normality [186]. We explored the log-transform and found that it actually had an overall detrimental effect on classification, because the use of PCA reduced the data's dimensionality and created a smaller set of features that mostly followed normal distributions. Another common data transformation is the square-root transform, used primarily for count data.

5.2: Discussion: classification and evaluation

We selected five simple classifiers that have been found to be useful in various image classification studies. We found both similarities and differences in classifier performance depending on the machine, treatment site, and study type. Discriminant analysis (primarily LDA) and LR had the best overall performance across all parameters. Using a linear model such as LDA is attractive because of its simplicity. It is easy to understand and it takes little time for model training and testing. If the assumptions of normality and equal class covariance matrices are met, the model can perform well for predictive purposes. Although the majority of real-world datasets would not be fully separable by a linear boundary, these models can still be successfully applied in practice [148]. The performance of QDA is tied to LDA by relaxing the equal class covariance matrices assumption; in practice, a linear or quadratic discriminant could be chosen depending on the outcome of a statistical test comparing these covariance matrices. LR is very useful in that it provides estimates of class posterior probabilities like discriminant analysis, but relies on less underlying assumptions about the dataset. For that reason, LR is generally considered to be a safer classifier than LDA, although both have been shown to perform similarly in practice [146, 159]. These posterior class estimates could be used as a scoring rule in addition to a binary classification,

where pre-stratified probability ranges could indicate varying levels of confidence about a correct or incorrect match.

Many alternative classification models exist and could be implemented for this study. Support vector machines (SVM) are supervised learning models that classify unknown observations by the construction of a linear boundary in a transformed version of the feature space, resulting in a nonlinear hyperplane in the original feature space [146]. SVM performs well with high-dimensional spaces and has many settings and kernels that can be optimized for a particular problem. SVMs have found utility in image classification [187] and various applications to detection and classification of medical images [188-190]. The effectiveness of SVMs is largely dependent on the kernel used and the specific tuning of model parameters, which can be a time and memory intensive process [191]. Another method that has recently been gaining popularity are random forests, which classify an unknown data point through the majority vote of a collection of decision trees [192]. Decision trees by themselves oftentimes are grown very deep, resulting in high variance and overfitting of the training dataset. Random forests average out these trees, and although the interpretability is reduced, final model performance is greatly improved [146, 192]. Random forests are non-parametric and do not require as much tuning or parameter selection as SVM would; they have found multiple uses in the literature for tasks related to classification in medical imaging [193-195]. Although model selection and its appropriate implementation is important, the choice of specific features used for model training can play an even larger role in the accuracy of a model's predictive ability [196].

We used several evaluation parameters to characterize and evaluate the strength of different classification models. With large enough sample sizes, the MCE is a useful measure for comparing model error rates. Sensitivity and specificity provide population-based measures regarding the model's ability to identify or preclude the presence of a correct/incorrect match. MCC is a single-value aggregate measure of a classifier's performance, and the likelihood ratios allow for

a probabilistic assessment of an image pair to be a correct/incorrect match. Previous studies have shown that there is no universally unbiased estimator of the variance of a k-fold cross validation estimator [197, 198]. A cross-validation scheme using bootstrap resampling was developed to generate moderately conservative confidence intervals with and without bias correction [199], which would be useful for providing more accurate confidence intervals in our study design. Given our relatively large sample size, it is unlikely that the current intervals would change significantly by implementing this bootstrap approach.

5.3: Discussion: workflow results

One of the most noticeable differences from the patient identification and misalignment study results is the lower MCE of Tomo image pairs when compared to TBeam image pairs. The registration file of therapist fusion at the time of treatment was available for TBeam images during the data acquisition process, but was inaccessible for Tomo images. To assess whether this factor could have driven this difference, we selected a subset of TBeam image pairs and manually fused them together using the same criteria used to fuse all Tomo image pairs. No significant difference was found between the MI/CC/SSIM metrics calculated from these image pairs versus the same values generated from therapist-fused image pairs. This was not an unexpected result – the process of image matching can be described as a low-dimensional optimization problem. As such, there is an increased likelihood for multiple users to converge upon the same 'solution' of image matching. Of course, this may not be the case for all patients, especially if the target area is affected by organ motion which may vary between the planning kVCT and setup image. The negligible difference between the two datasets suggests another underlying reason behind the discrepancy between the two imaging modalities.

Upon further investigation, we hypothesized that image quality was the largest contributor to the MCE differences between TBeam and Tomo image pairs. The wider collimator from CBCT

acquisition on the TBeam leads to increased scatter radiation, which in turn causes increased noise, image artifacts, and decreased contrast resolution [66, 200-203]. All of these factors can contribute to a decreased metric accuracy between correct image pairs, thereby degrading the accuracy in subsequent classification. In the frequency histograms shown in **Figures 28-30**, a larger spread can be seen on correct TBeam images, which suggests added variability between image pair matching.

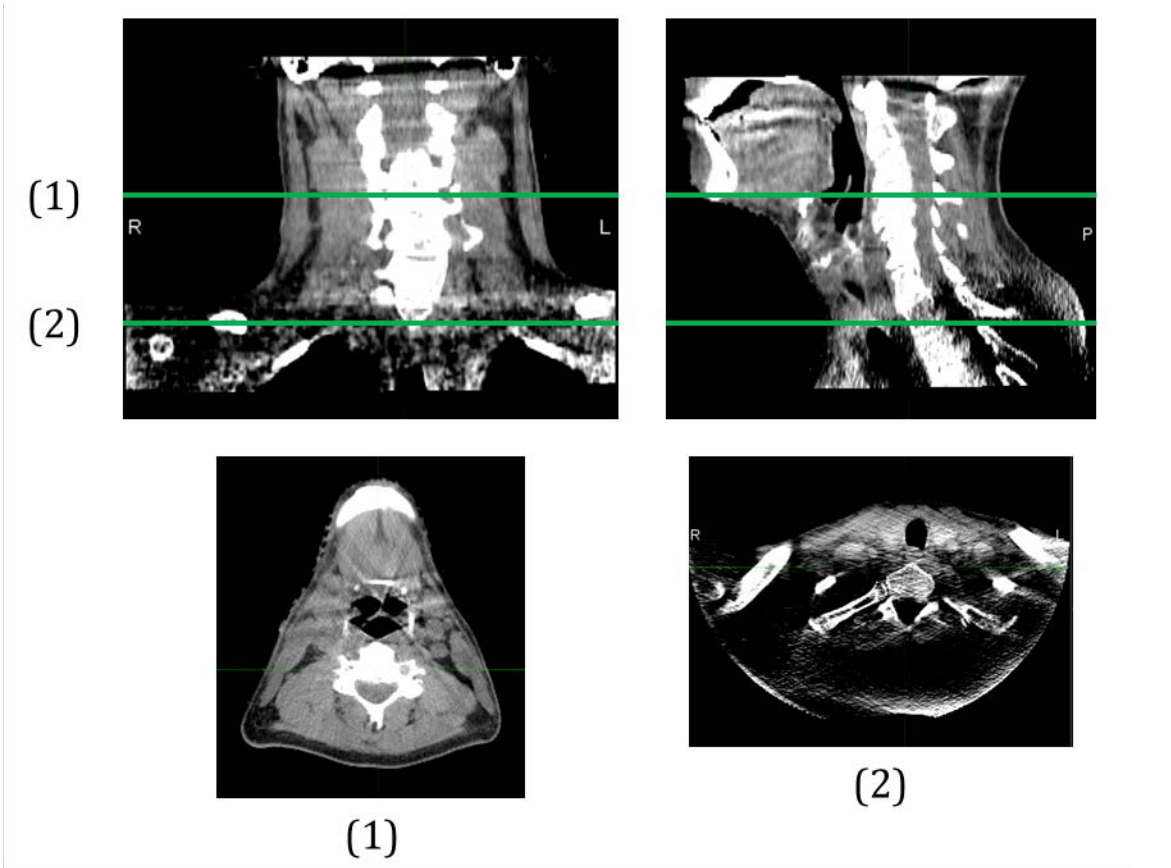


Figure 47: Example of artifacts due to the limited FOV of the CBCT scanner for H&N scans. Top row shows coronal (left) and sagittal (right) slices of a H&N CBCT. Green lines (1) and (2) represent axial slices shown in the bottom row, where reduced image quality is present in the reconstructed shoulder region when compared to the smaller FOV of the neck.

We observed that TBeam CC values for RP image pairs overlapping with the incorrect image pair histogram corresponded strongly to images containing high noise levels and severe artifacts, particularly towards the inferior slices. This is potentially due to the limited FOV in the built-in parameters of H&N scans during image acquisition; in certain field views (e.g. anterior-posterior),

the full width of a patient's shoulders may not be acquired, while the x-ray path penetrates other portions of the patient not present in the FoV, resulting in additional artifacts during image reconstruction (**Figure 47**). To further investigate this effect on model accuracy, the portion of the H&N mask corresponding to this region could be cropped prior to the image pre-processing workflow.

From the results in **Tables 7-10** and **16-17**, a large standard deviation can be seen for many of the LR+ values, especially when classification performance is strong. From equation (34), a perfect specificity would result in an infinite LR+, indicating that a positive test result (e.g. RP) is guaranteed to occur in image pairs that truly are correct. For the instances of perfect classification, we estimated the sensitivity and specificity values to be 99.99% to allow for a finite LR+, resulting in a larger spread of values. Nonetheless, high LR+ values are still useful as a general indicator of true positive results. The LR classifier has been shown to have convergence issues in cases of perfect classification resulting from data points that are linearly separable. Equation (30) shows the basic logistic regression equation, which aims to find a maximum log-likelihood fit for the data. With perfect class separation, the log-likelihood achieves a maximum of 0. The equation coefficients/weights are therefore not uniquely defined, resulting in extreme or infinite values [204]. When this occurred, we estimated the LR+/LR- parameters in a similar fashion as described above. It has been shown that adding a penalty term to the objective function can help stabilize the coefficients in cases of linear separability [205].

The rationale behind the gradient-based metrics was to provide complementary information to standard similarity metrics by meaningfully including both spatial and intensity information. This metric was designed specifically to account for same-patient misalignments (e.g. minor deformations) that could degrade the quality of a RP image pair. Our results show that although the CC/MI/SSIM metrics provided the majority of the discriminatory power between correct and incorrect image pairs, the inclusion of the custom metric features overall reduced the

error and increased the sensitivity / specificity estimates for the patient identification study. However, we observed generally better classification results for the patient alignment study by excluding the features from the gradient-based metrics, suggesting that separate workflows for error detection in patient identification and patient alignment may be preferable (see section 5.5).

Although the current system generally produces high sensitivity and specificity estimates, 1cm TBeam shifts produced relatively low specificity estimates (**Table 17**). Sensitivity and specificity are especially important values with regards to implementing this system in clinical practice. In order to minimize disruption to the clinical workflow, the system needs to have high enough a specificity to prevent excess false positives. Assuming 20-30 patients treated on a single machine, a specificity of ≥99% would produce one false positive per week on average, which is a reasonable value to permit in the clinic. Increasing a classifier's specificity comes with an inherent tradeoff of a decreased sensitivity, which prevents false negatives and allows the system to act as a robust second-check to the therapist fusion. Given that no such system is currently implemented in clinical practice, it is difficult to define a gold standard to establish a minimum sensitivity requirement; in addition, the current purpose of this system is not to replace a therapist entirely, but rather provide a robust second-check to their own judgment. As such, we feel a value of ≥85% would be a reasonable and practical sensitivity estimate for this purpose. Our workflow would perform well under these estimates for patient identification across both machines and for vertebral misalignments and gross H&N/pelvis misalignments ≥2cm, as summarized in **Table 18**. Our system is currently not robust enough for 1cm misalignments (except for Tomo H&N images) and would need further design improvements to reliably detect small shifts. Some alternative approaches are suggested in section 5.5. Our sensitivity/specificity requirements of 99%/85% are rough estimates and could change on a department-by-department basis, depending on patient throughput and structure of the treatment workflow.

		TomoTherapy	TrueBeam
Patient ID	*H&N*	**Yes** (Sp > 99%, Se = 100 %)	**Yes** (Sp ≥ 99%, Se > 96 %)
	Pelvis	**Yes** (Sp ≥ 99%, Se ≥ 97%)	**Yes** (Sp ≥ 99%, Se ≥97 %)
	Spine	**Yes** (Sp > 99%, Se = 100%)	**Yes** (Sp ≥ 99%, Se ≥ 95%)
Patient misalignment	*H&N: 1cm*	**Yes** (Sp ≥ 99%, Se ≥ 95%)	No*
	H&N: 2cm	**Yes** (Sp > 99%, Se = 100%)	**Yes** (Sp ≥ 99%, Se ≥ 86%)
	Pelvis: 1cm	No†	No‡
	Pelvis: 2cm	**Yes** (Sp ≥ 99%, Se > 96%)	**Yes** (Sp ≥ 99%, Se ≥ 85%)
	Spine: VS	**Yes** (Sp ≥ 99%, Se ≥ 91%)	**Yes** (Sp ≥ 99%, Se ≥ 90%)

VS: Vertebral shift; Se: sensitivity; Sp; specificity. Green shaded cells indicate success in achieving 99%/85% Sp/Se.
*: With 99% specificity, sensitivity = 30%; with 85% sensitivity, specificity = 81%
†: With 99% specificity, sensitivity = 73%; with 85% sensitivity, specificity = 96%
‡: With 99% specificity, sensitivity = 46%; with 85% sensitivity, specificity = 74%

5.4: Study limitations

As discussed earlier, the patient identification study included *N* RP image pairs and 2*N* WP image pairs. These images were collectively used to train and test the various classification models. One limitation of this approach is that these sample proportions of correct and incorrect image pairs are likely not representative of the true proportion of correct and incorrect patient treatments. This introduces a limited sampling bias, where the a-posteriori probability estimates from the sample size proportions are not representative of the underlying probability distributions of correct and incorrect patient treatments [206]. Although the true error rate in RT treatments is unknown [36, 45], a reasonable error rate could have been estimated and used to define the sample proportions. However, we were limited by the number of unique RP image pairs available to use in the study. Even assuming a relatively large error rate would have resulted in a disproportionate dataset with small sample sizes for the WP class, which would not be sufficient to adequately represent the distribution of values from WP image pairs. This itself would bias parameter selection during model training. In addition, a small sample size would necessitate the use of fewer features to avoid the 'curse of dimensionality' [207]. In short, smaller sample sizes combined with larger

feature spaces result in a classifier treating data as sparse, significantly decreasing the accuracy and validity of a model's output. As such, we chose to include a large number of WP image pairs in order to provide sufficient data for classifier training and testing. This concept of class imbalance is well-recognized in the literature [208], and possible solutions range from altering the misclassification cost ratio [209] to implementing various sampling techniques [210, 211].

As discussed earlier, our system had a more difficult time discriminating between correctly-aligned and 1cm-shifted patients (**Tables 16** and **17**). In the current design, discrimination between a correct and incorrect match relies on a global assessment of the anatomical features inside the patient's body. Deformations and organ motion (due to breathing, rectal filling, or other anatomical variations) could cause shifts in the local target volume relative to the overall anatomy (e.g. prostate [175]). The system would not be able to detect a correct therapist alignment to a shifted target, increasing the likelihood of a false positive. Further studies need to be performed to improve detection of smaller misalignments, and be able to distinguish whether or not they are correct (e.g. due to organ motion as described above, or internal deformation) or incorrect (e.g. therapist error).

In performing this study, we acquired several correct image pairs and used them to simulate potential identification and alignment errors. One inherent limitation in this setup is producing these errors ourselves instead of utilizing real examples of errors. We had difficulty in acquiring such data due to the low number of reported errors at our institution; in addition, there is the potential of errors passing by unnoticed or unreported. Having a large collection of these errors would allow for a more realistic classification system; in addition, it would allow for further investigation for the reasons behind such errors occurring, which could lead towards systematic changes to prevent the occurrence of such errors. A global error reporting database would allow for a pooled collection of images that could be beneficial for further research and analysis.

Another general limitation of our proposed system is that it can only target errors that can be detected from the IGRT images themselves. For example, the system assumes that no change in patient positioning occurs between the setup image acquisition and the RT treatment. In some situations, therapists may discover after patient setup that the couch needs to be moved to allow clearance for gantry rotation. If therapists do not move the patient back to the original position and fail to reacquire a new set of setup images, the system would likely be output a false positive for treatment. Another example involves patients with multiple treatment sites. Although the correct patient may be set up in the correct treatment position, selection of the incorrect beam parameters (e.g. site #1 is irradiated with the beam for site #2) would not be detected by the system. A therapist may do a correct fusion, but apply the shifts in the incorrect direction – or simply not apply the shifts at all. Many of these position-based errors could potentially be detected by a real-time camera tracking system that actively tracks a patient prior to and during RT treatment. AlignRT (Vision RT, London, UK) is a recently developed commercial system that uses optical surface imaging for tracking patient motion and has been shown to be beneficial in assessing patient setup alignment in a real-time fashion [212]. Although one major purpose of our study was to develop a technique that did not rely on additional equipment such as camera tracking, such a feature would have great value for those institutions that already have or are able to afford this equipment.

5.5: Future studies and directions

There are several ways to improve upon the current algorithmic design. Additional constraints could be placed on the training image dataset to test the algorithm's robustness, such as using patient images of the same gender, same treatment site (e.g. prostate), and more. Classification performance with a dataset comprising solely of 'edge case' patients (WP image pairs with better similarity matches, and RP image pairs with poorer similarities) should be performed.

Although it is expected that classification performance will be poorer, this would allow for additional investigation for site-specific or general improvements for the current workflow. Another design improvement could involve finding the minimum longitudinal FOV necessary on the setup image to achieve an accurate classification. This has the potential to minimize patient dose and reduce time spent in image acquisition while retaining the same discriminatory characteristics in classification. The performance of the current design should also be tested on images of additional anatomical sites, such as the abdomen. In our initial collection of spine data, we included a small subset of Tomo images (N=16) that largely contained soft tissue in the abdominal/lumbar spine regions. MCE rates for patient identification ranged from 2% to 12% across the tested classifiers and the various combinations of input features. Although additional data is needed for testing, these preliminary values suggest the need for additional design improvements in the abdominal region.

Our current workflow does not distinguish between a patient identification error and a patient alignment error. In clinical implementation, separate workflows for these error types should be implemented and run simultaneously. Site-specific workflows could be a potential large-scale improvement and future direction of this study. For practicality purposes, our goal in the present study was to construct a single workflow that could be applied to all treatment sites for a given setup image modality. However, site-specific workflows would have the potential of optimizing the algorithm's comparative performance due to inherent differences between various anatomical regions of the body. To retain automaticity, an independent algorithm could be implemented to extract the patient's treatment site from an institution's medical records and insert a given patient's image pair into the appropriate workflow. Most record-and-verify systems (such as ARIA) contain patient data in the back-end registry in tabular format. A script could be written to search for the relevant tables containing the pathology, parse the text, and search for pre-specified keywords that would assign the patient to the workflow of a certain anatomical region. Querying

the database as described can be rapidly performed and could be implemented in a real-time fashion without hindering the clinical workflow.

We hypothesized that noise and artifacts in CBCT images were large contributors to lower accuracy during image classification. Improving CBCT image quality would likely allow for better RP image matching and improve classifier performance. Some examples could include additional image processing on sinograms and projection data during reconstruction as well as improved image post-processing. Studies have been done to remove stripe/ring and shading artifacts [213, 214], reduce scatter radiation [215], and reduce metal artifacts [216] in CBCT imaging. Studies have also shown the benefits of using alternative reconstruction methods over the standard FDK reconstruction, especially for noisy or undersampled projection data [217, 218].

Specific components of the workflow could also be improved. Our current gradient-based metric includes both spatial and intensity information in each image pair comparison. Additional spatial relationships could be included by examining the output metric as a function of spatial direction or location. Several discrete clusters of similar values could suggest a true match, as this would show a similar trend in image matching as the algorithm loops through each voxel in the mask. Additional strategies involving spatial comparison could be explored as a separate comparative index, such as deformable image registration (DIR). DIR has been used to assess tumor classification using multimodality imaging by various volumetric characteristics [219]. The resultant deformable vector fields from DIR could potentially be used for image similarity assessment by finding ways to distinguish between several scenarios: correct patient and setup, correct patient but with significant deformation, correct patient but with a systematic error in setup, and a wrong patient.

The current system design is such that image similarities calculated for some unknown image pair will be classified as a correct or incorrect match based on a general collection of known image pairs. Although we achieved high accuracy with this approach, we noticed that RP values

within a given anatomical region can have an inherent spread (e.g. **Figures 28-32**). To further enhance the specificity of a correct patient identification or patient setup, additional patient-specific regularization constraints could be added. For example, a patient will undergo several setup images throughout the course of RT treatment. The analysis of the patient's first fraction would be the same as the proposed design. For subsequent fractions, the image similarity values could be compared to the initial comparison and/or to previous fractions, and visualized using a statistical control chart (**Figure 48**). A control chart is a tool that tracks the variation of a given process or behavior over time to determine if it is in a state of statistical control [220, 221]. Features from subsequent fractions can be tracked over time for each individual patient, and any deviation outside some previously-determined tolerance could indicate the need for therapist intervention.

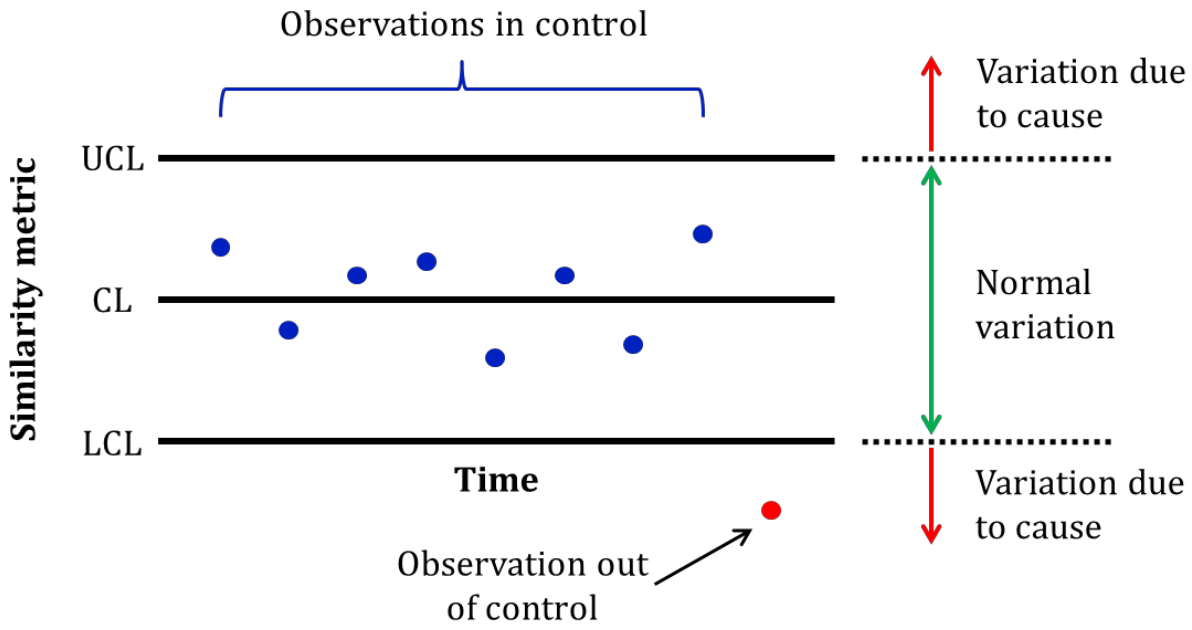


Figure 48: General schematic of a statistical control chart. Blue data points indicate image pairs of a given point with varying similarity values (y-axis) over time (x-axis). The control limit (CL) represents the average predicted value over time. The upper and lower control limits (UCL, LCL) are user-defined tolerances that allow for typical variation of the metric. The red value falling below the LCL indicates a potential error during setup, requiring further investigation before treatment.

To determine if any patient-specific changes (such as significant changes in tumor size or patient weight) would detrimentally affect the ability of the current system to detect a correct patient, we gathered 10 H&N TBeam patients that required a replan primarily because of weight loss. Image pairs corresponding to the first and last fraction prior to adaptive treatment were collected and run through the proposed workflow. The classification result of the second fraction was the same as the first fraction for all 10 patients, suggesting that clinically-significant weight loss requiring a replan would not affect the metric's ability to correctly identify the same patient. We also tested for any significant differences between similarity metrics of the image pairs between the two fractions by running three separate Wilcoxon signed-rank tests comparing the CC, MI, and SSIM features. The test resulted in p-values of 0.49, 0.23, and 0.56, respectively, indicating no significant differences. Further studies should be performed to support these findings by collecting additional image pairs between the first fraction and last fraction prior to the replan, as well as additional H&N patients as well as different sites.

Another practical regularization constraint would be only using patients currently undergoing RT treatment for model training. The WP image pairs for model training could include patients from the same site or across all treatment sites. Another possibility is to separate patients by treatment machine, although this would depend more on the specific department's structure and clinical workflow procedures.

Alternative approaches of image similarity assessment could also be explored. Instead of using the global image space, patient-specific anatomical markers could be pre-defined on the planning kVCT. For each setup image, these same areas could be automatically segmented or detected and then compared to the original planning kVCT image using various size or shape characteristics. To prevent false positives due to similar anatomical characteristics, the spatial coordinates of the markers on the setup image could be mapped to the planning kVCT, and their overlap or spatial differences could be assessed. The specific markers would need to be carefully

selected based on both the general anatomical region and the specific target area being treated, and should be reliably detected using automatic methods. Various studies have shown successful automatic segmentation of various organs (such as the liver, heart, prostate, and bladder) using 3D CT images, further expanding the potential markers that could be selected for comparative use [222-226].

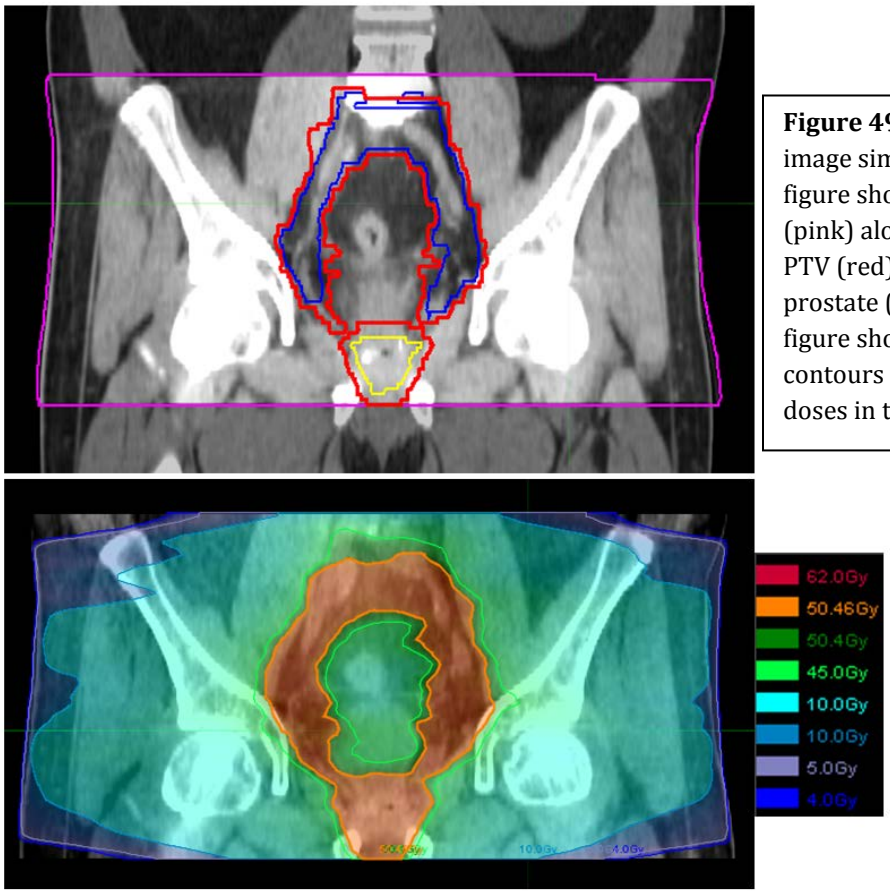


Figure 49: Alternative masks for image similarity comparison. Top figure shows a body contour (pink) along with contours of the PTV (red), CTV (blue), and prostate (yellow). The bottom figure shows various isodose contours with their respective doses in the legend.

Our workflow uses the body contour of the setup image to define an initial mask for subsequent similarity assessment. However, alternative masks could also be explored for this purpose (**Figure 49**). To focus the comparison on a region that is more relevant to the patient's treatment, an initial mask could be centered on the isocenter of the target. Along with the setup and planning kVCT images, the target contours can be exported from the planning system. A volumetric ROI could then be generated around the target volume and used for the initial mask. Another

approach is to use a pre-specified isodose line. Dose distributions of a given patient can be exported into MIM, which has the capability of creating contours from isodose lines. For both cases, CERR could be used to convert the contours into binary mask files as done in the current setup. The benefit of this approach would be the ability to detect errors that could affect the patient's dosimetric outcome for a given treatment. One potential downside is the lack of global information in the similarity comparison, which may not have enough data points to provide enough discrimination between right and wrong matches.

5.6: Concluding thoughts

Patient safety is of great importance in a complex healthcare discipline such as radiation oncology, and the RT workflow should continually be analyzed for areas of possible improvement. Failure mode and effects analysis (FMEA) is a widely-used tool, particularly in manufacturing, used to qualify potential failures through an extensive review of a given system's components and sub-system [227]. Its use has increased in the healthcare industry in recent years, and has been effectively usd for radiation oncology clinics [39, 228-230]. The benefit of FMEA is its prospective nature, seeking out vulnerabilities and hazards before any harm is delivered to the patient. These can then be prioritized by the RT staff in order to develop interventions and systematic changes for improving patient safety [231].

Our study focused specifically on the patient setup stage prior to RT treatment. The rationale behind the study was based on the idea that IGRT could have uses beyond a means for therapists to align patients. Although the use of 2D/3D image matching is a standard procedure in radiotherapy clinics, they generally rely on therapists for the final match prior to patient treatment. As mentioned earlier, we found that many therapists at our institution avoid the use of the built-in automatic registration software. From further investigation, we found that the software's performance is not consistent and may perform an incorrect registration (**Figure 6**). As such, the

element of human error is present in the majority of image fusions. IGRT could therefore be utilized in an automated fashion for the purposes of double-checking the work of the therapist. The proposed system in this dissertation could be clinically implemented as an interlock at the treatment console or as stand-alone software that operates in conjunction with the clinical system.

There are many potential benefits that this software could provide to the radiation oncology community. First and foremost, its use as an automated second-check software could help prevent any potentially significant errors related to identification and misalignment (due to therapist inattention/error during both the setup and the image registration process) from reaching the patient before RT treatment. This system relies only on the IGRT images themselves, and thus could be implemented in any institution with IGRT capabilities. As such, costs associated with expensive equipment (i.e. video tracking software) are virtually eliminated. Due to the automated nature of the system, there is also no added time to the clinical workflow (as compared to previously-published proposed safety systems where additional setup equipment was required on the patient [58, 59]). Avoiding this step also bypasses any potential for therapist error in setting up the equipment. In addition to prospective cost savings as a result of any prevented errors, there is also the potential for annual departmental savings by not requiring two therapists at the treatment console (i.e. eliminate the need of a second therapist to double-check image fusion). According to the U.S. Bureau of Labor Statistics, the median annual pay of a radiation therapist in 2012 was $77,560 [232]. As such, the potential cost savings could be immense, especially for departments with several treatment machines. Finally, the utility of this work is evident for developing or third-world nations, where larger error rates have been primarily attributed to a lack of trained staff and insufficient knowledge. Modern radiotherapy techniques and equipment, including 3D IGRT, do exist in many of these nations around the world. As of May 2015, there exist 32 TomoTherapy units in 10 developing countries, and 315 Varian CBCT units in 32 developing countries.

Many alternative approaches to developing this system have been identified in this chapter, from changing the values of specific parameters to using entirely different ways of performing similarity assessment. The current system is by no means considered complete, and it is likely that many of these alternate pathways or improvements would prove to be useful in error detection. As hardware and software technologies continue to improve, a larger set of tools will become available for potential use in further enhancing the system's design and robustness.

Many features of patient safety do not involve financial resources, but rather the implementation of safe practices. As discussed extensively throughout this dissertation, human mistakes have been and continue to be a large source of errors in both medical and non-medical work environments. Poor communication, inattention, fatigue, and improper training are some of the many reasons behind the occurrence of errors, which can have both trivial and significant repercussions. Establishing good systematic practices is key to preventing potential errors from occurring. In addition to FMEA, which prospectively assesses a systematic design for errors, reporting systems can provide information in a retrospective fashion that leads to improved safety [9]. Error reporting systems have been effectively used in the radiation therapy field to identify both systematic and specific factors that lead to error production [36, 45, 55, 233]. Understanding the factors that lead to errors is crucial for establishing changes for effective error prevention.

In a larger context, our proposed system is but a small aspect in improving the safety of the radiation oncology field. Continued research and efforts in identifying potential weaknesses in department protocols and clinical workflows is paramount to reducing any potential patient safety or risks. With the rising complexity of hardware and software technologies, it is increasingly important to establish a department culture that encourages clear and open communication without inducing any fear of reprisal for reporting errors. In the healthcare industry, the patient always comes first, and an active effort in maintaining this ideology in a systematic manner will allow for continued and improved patient safety.

REFERENCES

1. Schimmel, E.M., *The Hazards of Hospitalization.* Ann Intern Med, 1964. **60**: p. 100-10.

2. McLamb, J.T. and R.R. Huntley, *The hazards of hospitalization.* South Med J, 1967. **60**(5): p. 469-72.

3. Couch, N.P., N.L. Tilney, A.A. Rayner, et al., *The high cost of low-frequency events: the anatomy and economics of surgical mishaps.* N Engl J Med, 1981. **304**(11): p. 634-7.

4. Dubois, R.W. and R.H. Brook, *Preventable deaths: who, how often, and why?* Ann Intern Med, 1988. **109**(7): p. 582-9.

5. Friedman, M., *Iatrogenic disease: addressing a growing epidemic.* Postgrad Med, 1982. **71**(6): p. 123-5, 128-9.

6. Steel, K., P.M. Gertman, C. Crescenzi, et al., *Iatrogenic illness on a general medical service at a university hospital.* N Engl J Med, 1981. **304**(11): p. 638-42.

7. Brennan, T.A., L.L. Leape, N.M. Laird, et al., *Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I.* N Engl J Med, 1991. **324**(6): p. 370-6.

8. Leape, L.L., A.G. Lawthers, T.A. Brennan, et al., *Preventing medical injury.* QRB Qual Rev Bull, 1993. **19**(5): p. 144-9.

9. Kohn, L.T., J. Corrigan, and M.S. Donaldson. *To err is human building a safer health system*. 2000; Available from: http://site.ebrary.com/id/10038653.

10. Wilson, R.M., W.B. Runciman, R.W. Gibberd, et al., *The Quality in Australian Health Care Study.* Med J Aust, 1995. **163**(9): p. 458-71.

11. Great Britain. Department of Health., *An organisation with a memory : report of an expert group on learning from adverse events in the NHS*. 2000, London: Stationery Office. xiv, 92 p.

12. Baker, G.R., P.G. Norton, V. Flintoft, et al., *The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada.* CMAJ, 2004. **170**(11): p. 1678-86.

13. Forster, A.J., T.R. Asmis, H.D. Clark, et al., *Ottawa Hospital Patient Safety Study: incidence and timing of adverse events in patients admitted to a Canadian teaching hospital.* CMAJ, 2004. **170**(8): p. 1235-40.

14. New Zealand Ministry of Health, *Adverse Events in New Zealand Public Hospitals: Principal Findings from a National Survey.* 2001.

15. Schioler, T., H. Lipczak, B.L. Pedersen, et al., *[Incidence of adverse events in hospitals. A retrospective study of medical records].* Ugeskr Laeger, 2001. **163**(39): p. 5370-8.

16. Eisenberg, J.M., *Statement on Medical Errors: Senate Appropriations Subcommittee on Labor Health and Human Services, and Education.* 1999.

17. Runciman, W., *Iatrogenic Injury in Australia: A reported prepared by the Australian Patient Safety Foundation.* Australian Patient Safety Foundation, 2001. **24**.

18. Reason, J., *Human error: models and management.* BMJ, 2000. **320**(7237): p. 768-70.

19. Reason, J.T., *Managing the risks of organizational accidents*. 1997: Ashgate.

20. Perrow, C., *Normal Accidents: Living with High Risk Technologies*. 1984: Princeton University Press.

21. Reason, J., *The contribution of latent human failures to the breakdown of complex systems.* Philos Trans R Soc Lond B Biol Sci, 1990. **327**(1241): p. 475-84.

22. Rasmussen, J., *Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models.* Systems, Man and Cybernetics, IEEE Transactions on, 1983. **SMC-13**(3): p. 257-266.

23. Reason, J.T., *Human error*. 1990, Cambridge England ; New York: Cambridge University Press. xv, 302 p.

24. Barger, L.K., N.T. Ayas, B.E. Cade, et al., *Impact of Extended-Duration Shifts on Medical Errors, Adverse Events, and Attentional Failures.* PLoS Medicine, 2006. **3**(12): p. e487.

25. Millenson, M.L., *The Silence.* Health Affairs, 2003. **22**(2): p. 103-112.

26. Neale, G., M. Woloshynowych, and C. Vincent, *Exploring the causes of adverse events in NHS hospital practice.* Journal of the Royal Society of Medicine, 2001. **94**(7): p. 322-330.

27. Hall, E.J. and A.J. Giaccia, *Radiobiology for the Radiologist*. 2006: Lippincott Williams & Wilkins.

28. Delaney, G., S. Jacob, C. Featherstone, et al., *The role of radiotherapy in cancer treatment: estimating optimal utilization from a review of evidence-based clinical guidelines.* Cancer, 2005. **104**(6): p. 1129-37.

29. Ringborg, U., D. Bergqvist, B. Brorsson, et al., *The Swedish Council on Technology Assessment in Health Care (SBU) Systematic Overview of Radiotherapy for Cancer including a Prospective Survey of Radiotherapy Practice in Sweden 2001 - Summary and Conclusions.* Acta Oncol, 2003. **42**: p. 357-62.

30. (2007) *Comprehensive audits of radiotherapy practices: a tool for quality improvement: Quality Assurance Team for Radiation Oncology (QUATRO)*.

31. (2008) *Setting up a radiotherapy programme: clinical, medical physics, radiation protection and safety aspects.*

32. International Commission on Radiological Protection (ICRP), *Radiological protection and safety in medicine. A report of the International Commission on Radiological Protection.* Ann ICRP, 1996. **26**(2): p. 1-47.

33. Kutcher, G.J., L. Coia, M. Gillin, et al., *Comprehensive QA for radiation oncology: report of AAPM Radiation Therapy Committee Task Group 40.* Med Phys, 1994. **21**(4): p. 581-618.

34. Novotny, J., J. Izewska, A. Dutreix, et al., *A quality assurance network in Central European countries - Radiotherapy infrastructure.* Acta Oncologica, 1998. **37**(2): p. 159-165.

35. World Health Organization (1988) *Quality assurance in radiotherapy*.

36. Williams, M.V., *Improving patient safety in radiotherapy by learning from near misses, incidents and errors.* British Journal of Radiology, 2007. **80**(953): p. 297-301.

37. Holmberg, O. and B. McClean, *Preventing treatment errors in radiotherapy by identifying and evaluating near misses and actual incidents.* Journal of Radiotherapy in Practice, 2002. **3**(01): p. 13-25.

38. Holmberg, O., *Accident prevention in radiotherapy.* Biomed Imaging Interv J, 2007. **3**(2): p. e27.

39. Ford, E.C., R. Gaudette, L. Myers, et al., *Evaluation of safety in a radiation oncology setting using failure mode and effects analysis.* Int J Radiat Oncol Biol Phys, 2009. **74**(3): p. 852-8.

40. Munro, A.J., *Hidden danger, obvious opportunity: error and risk in the management of cancer.* Br J Radiol, 2007. **80**(960): p. 955-66.

41. Shafiq, J., M. Barton, D. Noble, et al., *An international review of patient safety measures in radiotherapy practice.* Radiother Oncol, 2009. **92**(1): p. 15-21.

42. Bogdanich, W., *Radiation offers new cures, and ways to do harm.* The New York Times, 2010.

43. Boadu, M. and M.M. Rehani, *Unintended exposure in radiotherapy: identification of prominent causes.* Radiother Oncol, 2009. **93**(3): p. 609-17.

44. Clark, B.G., R.J. Brown, J.L. Ploquin, et al., *The management of radiation treatment error through incident learning.* Radiother Oncol, 2010. **95**(3): p. 344-9.

45. Mutic, S., R.S. Brame, S. Oddiraju, et al., *Event (error and near-miss) reporting and learning system for process improvement in radiation oncology.* Med Phys, 2010. **37**(9): p. 5027-36.

46. Izewska, J., P. Andreo, S. Vatnitsky, et al., *The IAEA/WHO TLD postal dose quality audits for radiotherapy: a perspective of dosimetry practices at hospitals in developing countries.* Radiother Oncol, 2003. **69**(1): p. 91-7.

47. Izewska, J., S. Vatnitsky, and K.R. Shortt, *Postal dose audits for radiotherapy centers in Latin America and the Caribbean: trends in 1969-2003.* Rev Panam Salud Publica, 2006. **20**(2-3): p. 161-72.

48. Shakespeare, T.P., M.F. Back, J.J. Lu, et al., *External audit of clinical practice and medical decision making in a new Asian oncology center: results and implications for both developing and developed nations.* Int J Radiat Oncol Biol Phys, 2006. **64**(3): p. 941-7.

49. Leunens, G., J. Verstraete, W. Van den Bogaert, et al., *Human errors in data transfer during the preparation and delivery of radiation treatment affecting the final result: "garbage in, garbage out".* Radiother Oncol, 1992. **23**(4): p. 217-22.

50. World Health Organization (WHO), *Radiotherapy Risk Profile.* WHO, 2008.

51. Duffey, R. and J. Saull, *Know the Risk: Learning from errors and accidents: safety and risk in today's technology*. 2002: Elsevier Science.

52. Bogdanich, W., *Case studies: when medical radiation goes awry.* The New York Times, 2010.

53. Sorcini, B. and A. Tilikidis, *Clinical application of image-guided radiotherapy, IGRT (on the Varian OBI platform).* Cancer Radiother, 2006. **10**(5): p. 252-7.

54. Khan, F.M., *Treatment Planning in Radiation Oncology*. 2007: Lippincott Williams & Wilkins.

55. Hendee, W.R. and M.G. Herman, *Improving patient safety in radiation oncology.* Medical Physics, 2011. **38**(1): p. 78-82.

56. Zhao, W., R. Chellappa, P.J. Phillips, et al., *Face recognition: A literature survey.* Acm Computing Surveys, 2003. **35**(4): p. 399-459.

57. Jun, Z., Y. Yong, and M. Lades, *Face recognition: eigenface, elastic matching, and neural nets.* Proceedings of the IEEE, 1997. **85**(9): p. 1423-1435.

58. Lappe, C., M. Braun, S. Helfert, et al., *Computer-controlled noninvasive patient positioning in fractionated radiotherapy - A videogrammetric system for automatic patient setup, fast detection of patient motion and online correction of target point misalignment during therapy.* Cvrmed-Mrcas'97, 1997. **1205**: p. 695-704.

59. Yan, G.H., K. Mittauer, Y. Huang, et al., *Prevention of gross setup errors in radiotherapy with an efficient automatic patient safety system.* Journal of Applied Clinical Medical Physics, 2013. **14**(6): p. 322-337.

60. Santhanam, A., H. Dou, A. Kurihara, et al., *Three-dimensional Feature Recognition-based Automated Patient Treatment Mismatch Verification System for Radiation Therapy.* International Journal of Radiation Oncology • Biology • Physics, 2012. **84**(3): p. S742.

61. Lamb, J.M., N. Agazaryan, and D.A. Low, *Automated patient identification and localization error detection using 2-dimensional to 3-dimensional registration of kilovoltage x-ray setup images.* Int J Radiat Oncol Biol Phys, 2013. **87**(2): p. 390-3.

62. Agazaryan, N., S.E. Tenn, A.A. Desalles, et al., *Image-guided radiosurgery for spinal tumors: methods, accuracy and patient intrafraction motion.* Phys Med Biol, 2008. **53**(6): p. 1715-27.

63. Jin, J.Y., S. Ryu, K. Faber, et al., *2D/3D image fusion for accurate target localization and evaluation of a mask based stereotactic system in fractionated stereotactic radiotherapy of cranial lesions.* Med Phys, 2006. **33**(12): p. 4557-66.

64. Powell, M.J.D., *UOBYQA: unconstrained optimization by quadratic approximation.* Mathematical Programming, 2002. **92**(3): p. 555-582.

65. Baeza, M., *Accident prevention in day-to-day clinical radiation therapy practice.* Ann ICRP, 2012. **41**(3-4): p. 179-87.

66. Miracle, A.C. and S.K. Mukherji, *Conebeam CT of the head and neck, part 1: physical principles.* AJNR Am J Neuroradiol, 2009. **30**(6): p. 1088-95.

67. Gupta, A.K., *Diagnostic Radiology: Recent Advances and Applied Physics in Imaging*. 2013: Jp Medical Limited.

68. Gupta, R., M. Grasruck, C. Suess, et al., *Ultra-high resolution flat-panel volume CT: fundamental principles, design architecture, and system characterization.* Eur Radiol, 2006. **16**(6): p. 1191-205.

69. Khan, F.M., *The Physics of Radiation Therapy*. 2010: Lippincott Williams & Wilkins.

70. Orth, R.C., M.J. Wallace, and M.D. Kuo, *C-arm cone-beam CT: general principles and technical considerations for use in interventional radiology.* J Vasc Interv Radiol, 2008. **19**(6): p. 814-20.

71. Graham, S.A., D.J. Moseley, J.H. Siewerdsen, et al., *Compensators for dose and scatter management in cone-beam computed tomography.* Med Phys, 2007. **34**(7): p. 2691-703.

72. Siewerdsen, J.H., D.J. Moseley, B. Bakhtiar, et al., *The influence of antiscatter grids on soft-tissue detectability in cone-beam computed tomography with flat-panel detectors.* Med Phys, 2004. **31**(12): p. 3506-20.

73. Mackie, T.R., T. Holmes, S. Swerdloff, et al., *Tomotherapy: a new concept for the delivery of dynamic conformal radiotherapy.* Med Phys, 1993. **20**(6): p. 1709-19.

74. Yang, J.N., T.R. Mackie, P. Reckwerdt, et al., *An investigation of tomotherapy beam delivery.* Med Phys, 1997. **24**(3): p. 425-36.

75. Yartsev, S., T. Kron, and J. Van Dyk, *Tomotherapy as a tool in image-guided radiation therapy (IGRT): theoretical and technological aspects.* Biomed Imaging Interv J, 2007. **3**(1): p. e16.

76. Meeks, S.L., J.F. Harmon, Jr., K.M. Langen, et al., *Performance characterization of megavoltage computed tomography imaging on a helical tomotherapy unit.* Med Phys, 2005. **32**(8): p. 2673-81.

77. Greene, T.C. and X.J. Rong, *Evaluation of techniques for slice sensitivity profile measurement and analysis.* Journal of Applied Clinical Medical Physics, 2014. **15**(2): p. 281-294.

78. Levegrun, S., C. Pottgen, J. Abu Jawad, et al., *Megavoltage Computed Tomography Image Guidance With Helical Tomotherapy in Patients With Vertebral Tumors: Analysis of Factors*

Influencing Interobserver Variability. International Journal of Radiation Oncology Biology Physics, 2013. **85**(2): p. 561-569.

79. Ruchala, K.J., G.H. Olivera, E.A. Schloesser, et al., *Megavoltage CT on a tomotherapy system.* Phys Med Biol, 1999. **44**(10): p. 2597-621.

80. Deans, S.R., *The Radon Transform and Some of Its Applications*. 2007: Dover Publications.

81. Gao, H., X.S. Qi, Y. Gao, et al., *Megavoltage CT imaging quality improvement on TomoTherapy via tensor framelet.* Med Phys, 2013. **40**(8): p. 081919.

82. Deasy, J.O., A.I. Blanco, and V.H. Clark, *CERR: a computational environment for radiotherapy research.* Med Phys, 2003. **30**(5): p. 979-85.

83. Bushberg, J.T., *The Essential Physics of Medical Imaging*. 2002: Lippincott Williams & Wilkins.

84. Boas, F.E. and D. Fleischmann, *CT artifacts: causes and reduction techniques.* Imaging in Medicine, 2012. **4**(2): p. 229-240.

85. Yang, D., S.R. Chaudhari, S.M. Goddu, et al., *Deformable registration of abdominal kilovoltage treatment planning kVCT and tomotherapy daily megavoltage CT for treatment adaptation.* Med Phys, 2009. **36**(2): p. 329-38.

86. Yang, D., S. Brame, I. El Naqa, et al., *Technical note: DIRART--A software suite for deformable image registration and adaptive radiotherapy research.* Med Phys, 2011. **38**(1): p. 67-77.

87. Stigler, S.M., *Francis Galton's Account of the Invention of Correlation.* Statistical Science, 1989. **4**(2): p. 73-79.

88. Fan, Y., D.G. Shen, R.C. Gur, et al., *COMPARE: Classification of morphological patterns using adaptive regional elements.* Ieee Transactions on Medical Imaging, 2007. **26**(1): p. 93-105.

89. George, T.C., S.L. Fanning, P. Fitzgeral-Bocarsly, et al., *Quantitative measurement of nuclear translocation events using similarity analysis of multispectral cellular images obtained in flow.* Journal of Immunological Methods, 2006. **311**(1-2): p. 117-129.

90. Kaneko, S., Y. Satoh, and S. Igarashi, *Using selective correlation coefficient for robust image registration.* Pattern Recognition, 2003. **36**(5): p. 1165-1173.

91. Kim, J. and J.A. Fessler, *Intensity-based image registration using robust correlation coefficients.* Ieee Transactions on Medical Imaging, 2004. **23**(11): p. 1430-1444.

92. Zitova, B. and J. Flusser, *Image registration methods: a survey.* Image and Vision Computing, 2003. **21**(11): p. 977-1000.

93. Roche, A., G. Malandain, X. Pennec, et al., *The correlation ratio as a new similarity measure for multimodal image registration.* Medical Image Computing and Computer-Assisted Intervention - Miccai'98, 1998. **1496**: p. 1115-1124.

94. Cover, T.M. and J.A. Thomas, *Elements of Information Theory*. 2012: Wiley.

95. Latham, P.E. and Y. Roudi, *Mutual information.* Scholarpedia, 2009. **4**(1).

96. Collignon, A., F. Maes, D. Delaere, et al., *Automated multi-modality image registration based on information theory.* Information Processing in Medical Imaging, 1995. **3**: p. 263-274.

97. Maes, F., A. Collignon, D. Vandermeulen, et al., *Multimodality image registration by maximization of mutual information.* IEEE Trans Med Imaging, 1997. **16**(2): p. 187-98.

98. Pluim, J.P., J.B. Maintz, and M.A. Viergever, *Image registration by maximization of combined mutual information and gradient information.* IEEE Trans Med Imaging, 2000. **19**(8): p. 809-14.

99. Wells, W.M., 3rd, P. Viola, H. Atsumi, et al., *Multi-modal volume registration by maximization of mutual information.* Med Image Anal, 1996. **1**(1): p. 35-51.

100. Maes, F., D. Vandermeulen, and P. Suetens, *Medical image registration using mutual information.* Proceedings of the IEEE, 2003. **91**(10): p. 1699-1722.

101. Meyer, C.R., J.L. Boes, B. Kim, et al., *Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate spline warped geometric deformations.* Med Image Anal, 1997. **1**(3): p. 195-206.

102. El Maia, H., A. Hammouch, and D. Aboutajdine, *Color-texture analysis by mutual information for multispectral image classification.* 2009 Ieee Pacific Rim Conference on Communications, Computers and Signal Processing, Vols 1 and 2, 2009: p. 359-364.

103. Garcia-Laencina, P.J., J.L. Sancho-Gomez, A.R. Figueiras-Vidal, et al., *K nearest neighbours with mutual information for simultaneous classification and missing data imputation.* Neurocomputing, 2009. **72**(7-9): p. 1483-1493.

104. Kerroum, M.A., A. Hammouch, and D. Aboutajdine, *Textural feature selection by joint mutual information based on Gaussian mixture model for multispectral image classification.* Pattern Recognition Letters, 2010. **31**(10): p. 1168-1174.

105. Russakoff, D.B., C. Tomasi, T. Rohlfing, et al., *Image similarity using mutual information of regions.* Computer Vision - Eccv 2004, Pt 3, 2004. **3023**: p. 596-607.

106. Wang, Z., A.C. Bovik, H.R. Sheikh, et al., *Image quality assessment: from error visibility to structural similarity.* IEEE Trans Image Process, 2004. **13**(4): p. 600-12.

107. Chen, G.H., C.L. Yang, L.M. Po, et al., *Edge-based structural similarity for image quality assessment.* 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, Vols 1-13, 2006: p. 2181-2184.

108. Chen, G.H., C.L. Yang, and S.L. Xie, *Gradient-based structural similarity for image quality assessment.* 2006 IEEE International Conference on Image Processing, ICIP 2006, Proceedings, 2006: p. 2929-2932.

109. Li, C.F. and A.C. Bovik, *Content-partitioned structural similarity index for image quality assessment.* Signal Processing-Image Communication, 2010. **25**(7): p. 517-526.

110. Sampat, M.P., Z. Wang, S. Gupta, et al., *Complex wavelet structural similarity: a new image similarity index.* IEEE Trans Image Process, 2009. **18**(11): p. 2385-401.

111. Wang, Z., E.P. Simoncelli, and A.C. Bovik, *Multi-scale structural similarity for image quality assessment.* Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, Vols 1 and 2, 2003: p. 1398-1402.

112. Zhang, L., L. Zhang, X.Q. Mou, et al., *FSIM: A Feature Similarity Index for Image Quality Assessment.* Ieee Transactions on Image Processing, 2011. **20**(8): p. 2378-2386.

113. Low, D.A. and J.F. Dempsey, *Evaluation of the gamma dose distribution comparison method.* Medical Physics, 2003. **30**(9): p. 2455-2464.

114. Low, D.A., *Gamma dose distribution evaluation tool.* J. Phys.: Conf. Ser., 2010. **250**: p. 012071-82.

115. Tomasi, C. and R. Manduchi. *Bilateral filtering for gray and color images*. in *Computer Vision, 1998. Sixth International Conference on*. 1998.

116. Kanopoulos, N., N. Vasanthavada, and R.L. Baker, *Design of an image edge detection filter using the Sobel operator.* Solid-State Circuits, IEEE Journal of, 1988. **23**(2): p. 358-367.

117. Bellman, R.E., *Dynamic Programming*. 2003: Dover Publications.

118. Caprihan, A., G.D. Pearlson, and V.D. Calhoun, *Application of principal component analysis to distinguish patients with schizophrenia from healthy controls based on fractional anisotropy measurements.* Neuroimage, 2008. **42**(2): p. 675-682.

119. Hoffmann, H., *Kernel PCA for novelty detection.* Pattern Recognition, 2007. **40**(3): p. 863-874.

120. Kalukin, A.R., M. Van Geet, and R. Swennen, *Principal components analysis of multienergy X-ray computed tomography of mineral samples.* Ieee Transactions on Nuclear Science, 2000. **47**(5): p. 1729-1736.

121. Sinha, U. and H. Kangarloo, *Principal component analysis for content-based image retrieval.* Radiographics, 2002. **22**(5): p. 1271-1289.

122. Towey, D.J., P.G. Bain, and K.S. Nijran, *Automatic classification of 123I-FP-CIT (DaTSCAN) SPECT images.* Nucl Med Commun, 2011. **32**(8): p. 699-707.

123. Unay, D., O. Soldea, A. Ekin, et al., *Automatic Annotation of X-Ray Images: A Study on Attribute Selection.* Medical Content-Based Retrieval for Clinical Decision Support, 2010. **5853**: p. 97-109.

124. El-Dahshan, E.S.A., T. Hosny, and A.B.M. Salem, *Hybrid intelligent techniques for MRI brain images classification.* Digital Signal Processing, 2010. **20**(2): p. 433-441.

125. Lopez, M., J. Ramirez, J.M. Gorriz, et al., *Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease.* Neurocomputing, 2011. **74**(8): p. 1260-1271.

126. Lopez, M., J. Ramirez, J.M. Gorriz, et al., *Automatic tool for Alzheimer's disease diagnosis using PCA and Bayesian classification rules.* Electronics Letters, 2009. **45**(8): p. 389-390.

127. Ramirez, J., J.M. Gorriz, F. Segovia, et al., *Computer aided diagnosis system for the Alzheimer's disease based on least squares and random forest SPECT image classification.* Neuroscience Letters, 2010. **472**(2): p. 99-103.

128. Ryan, E.A. and M.J. Farquharson, *Breast tissue classification using x-ray scattering measurements and multivariate data analysis.* Physics in Medicine and Biology, 2007. **52**(22): p. 6679-6696.

129. Zhang, Y.D., Z.C. Dong, L.N. Wu, et al., *A hybrid method for MRI brain image classification.* Expert Systems with Applications, 2011. **38**(8): p. 10049-10053.

130. Turk, M., *A random walk through Eigenspace.* Ieice Transactions on Information and Systems, 2001. **E84d**(12): p. 1586-1595.

131. Michie, D., D.J. Spiegelhalter, C.C. Taylor, et al., eds. *Machine learning, neural and statistical classification*. 1994, Ellis Horwood.

132. Wing, S., D. Richardson, D. Armstrong, et al., *A reevaluation of cancer incidence near the Three Mile Island nuclear plant: the collision of evidence and assumptions.* Environ Health Perspect, 1997. **105**(1): p. 52-7.

133. Altman, N.S., *An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression.* American Statistician, 1992. **46**(3): p. 175-185.

134. Wu, X.D., V. Kumar, J.R. Quinlan, et al., *Top 10 algorithms in data mining.* Knowledge and Information Systems, 2008. **14**(1): p. 1-37.

135. Chen, X.W., X.B. Zhou, and S.T.C. Wong, *Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy.* Ieee Transactions on Biomedical Engineering, 2006. **53**(4): p. 762-766.

136. Karssemeijer, N., *Automated classification of parenchymal patterns in mammograms.* Physics in Medicine and Biology, 1998. **43**(2): p. 365-378.

137. Christodoulou, C.I., C.S. Pattichis, M. Pantziaris, et al., *Texture-based classification of atherosclerotic carotid plaques.* Ieee Transactions on Medical Imaging, 2003. **22**(7): p. 902-912.

138. Fisher, R.A., *THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS.* Annals of Eugenics, 1936. **7**(2): p. 179-188.

139. Friedman, J.H., *Regularized Discriminant-Analysis.* Journal of the American Statistical Association, 1989. **84**(405): p. 165-175.

140. Chan, H.P., D.T. Wei, M.A. Helvie, et al., *Computer-Aided Classification of Mammographic Masses and Normal Tissue - Linear Discriminant-Analysis in Texture Feature Space.* Physics in Medicine and Biology, 1995. **40**(5): p. 857-876.

141. Krafft, C., S.B. Sobottka, K.D. Geiger, et al., *Classification of malignant gliomas by infrared spectroscopic imaging and linear discriminant analysis.* Analytical and Bioanalytical Chemistry, 2007. **387**(5): p. 1669-1677.

142. Cox, D.D. and R.L. Savoy, *Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex.* Neuroimage, 2003. **19**(2): p. 261-270.

143. Etemad, K. and R. Chellappa, *Discriminant analysis for recognition of human face images.* Journal of the Optical Society of America a-Optics Image Science and Vision, 1997. **14**(8): p. 1724-1733.

144. Preul, M.C., Z. Caramanos, D.L. Collins, et al., *Accurate, noninvasive diagnosis of human brain tumors by using proton magnetic resonance spectroscopy.* Nature Medicine, 1996. **2**(3): p. 323-325.

145. Zhao, W., R. Chellappa, and A. Krishnaswamy, *Discriminant analysis of principal components for face recognition.* Automatic Face and Gesture Recognition - Third Ieee International Conference Proceedings, 1998: p. 336-341.

146. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. 2009: Springer.

147. Zhao, W., R. Chellappa, and N. Nandhakumar, *Empirical performance analysis of linear discriminant classifiers.* 1998 Ieee Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings, 1998: p. 164-169.

148. Narsky, I. and F.C. Porter, *Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning*. 2013: Wiley.

149. Hand, D.J. and K.M. Yu, *Idiot's Bayes - Not so stupid after all?* International Statistical Review, 2001. **69**(3): p. 385-398.

150. Domingos, P. and M. Pazzani, *On the optimality of the simple Bayesian classifier under zero-one loss.* Machine Learning, 1997. **29**(2-3): p. 103-130.

151. Klement, W., S. Wilk, W. Michalowski, et al., *Predicting the need for CT imaging in children with minor head injury using an ensemble of Naive Bayes classifiers.* Artificial Intelligence in Medicine, 2012. **54**(3): p. 163-170.

152. Prasad, M.N., A. Sowmya, and I. Koch, *Feature subset selection using ICA for classifying emphysema in HRCT images.* Proceedings of the 17th International Conference on Pattern Recognition, Vol 4, 2004: p. 515-518.

153. Xu, Y., M. Sonka, G. McLennan, et al., *MDCT-based 3-D texture classification of emphysema and early smoking related lung pathologies.* Ieee Transactions on Medical Imaging, 2006. **25**(4): p. 464-475.

154. Marshall, L.F., S.B. Marshall, M.R. Klauber, et al., *A New Classification of Head-Injury Based on Computerized-Tomography.* Journal of Neurosurgery, 1991. **75**: p. S14-S20.

155. Kurt, I., M. Ture, and A.T. Kurum, *Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease.* Expert Systems with Applications, 2008. **34**(1): p. 366-374.

156. Langer, D.L., T.H. van der Kwast, A.J. Evans, et al., *Prostate Cancer Detection With Multi-parametric MRI: Logistic Regression Analysis of Quantitative T2, Diffusion-Weighted Imaging, and Dynamic Contrast-Enhanced MRI.* Journal of Magnetic Resonance Imaging, 2009. **30**(2): p. 327-334.

157. Bewick, V., L. Cheek, and J. Ball, *Statistics review 14: Logistic regression.* Critical Care, 2005. **9**(1): p. 112-118.

158. Hosmer, D.W. and S. Lemeshow, *Applied Logistic Regression*. 2004: Wiley.

159. Pohar, M., M. Blas, and S. Turk, *Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study* Metodoloski zvezki, 2004. **1**(1): p. 143-161.

160. Fienberg, S.E., *The Analysis of Cross-Classified Categorical Data*. 2007: Springer.

161. McLachlan, G., *Discriminant Analysis and Statistical Pattern Recognition*. 2004: Wiley.

162. Breiman, L. and P. Spector, *Submodel Selection and Evaluation in Regression - the X-Random Case.* International Statistical Review, 1992. **60**(3): p. 291-319.

163. Kohavi, R., *A study of cross-validation and bootstrap for accuracy estimation and model selection*, in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*. 1995, Morgan Kaufmann Publishers Inc.: Montreal, Quebec, Canada. p. 1137-1143.

164. Metz, C.E., *Basic Principles of Roc Analysis.* Seminars in Nuclear Medicine, 1978. **8**(4): p. 283-298.

165. Hanley, J.A. and B.J. Mcneil, *The Meaning and Use of the Area under a Receiver Operating Characteristic (Roc) Curve.* Radiology, 1982. **143**(1): p. 29-36.

166. Matthews, B.W., *Comparison of Predicted and Observed Secondary Structure of T4 Phage Lysozyme.* Biochimica Et Biophysica Acta, 1975. **405**(2): p. 442-451.

167. Powers, D.M.W., *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation.* J Mach Learn Tech, 2011. **2**(1): p. 37-63.

168. Baldi, P., S. Brunak, Y. Chauvin, et al., *Assessing the accuracy of prediction algorithms for classification: an overview.* Bioinformatics, 2000. **16**(5): p. 412-424.

169. Gallagher, E.J., *The problem with sensitivity and specificity ...* Annals of Emergency Medicine, 2003. **42**(2): p. 298-303.

170. Akobeng, A.K., *Understanding diagnostic tests 1: sensitivity, specificity and predictive values.* Acta Paediatrica, 2007. **96**(3): p. 338-341.

171. Akobeng, A.K., *Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice.* Acta Paediatrica, 2007. **96**(4): p. 487-491.

172. Snedecor, G.W. and W.G. Cochran, *Statistical Methods: By George W. Snedecor and William G. Cochran*. 1989: Iowa State University Press.

173. Lilliefors, H.W., *On Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown.* Journal of the American Statistical Association, 1967. **62**(318): p. 399-&.

174. Wand, P. and C. Jones, *Kernel Smoothing*. 1994: Taylor & Francis.

175. Langen, K.M. and D.T.L. Jones, *Organ motion and its management.* International Journal of Radiation Oncology Biology Physics, 2001. **50**(1): p. 265-278.

176. Jani, S., A. Kishan, D. O'Connell, et al., *Prediction of Pelvic Nodal Coverage Using Mutual Information Between Cone-Beam and Planning kVCTs.* Medical Physics, 2014. **41**(6): p. 198-198.

177. Lewis, J.P., *Fast normalized cross-correlation.* Vision Interface, 1995: p. 120-123.

178. Wang, F., B.C. Vemuri, M. Rao, et al., *A new & robust information theoretic measure and its application to image alignment.* Information Processing in Medical Imaging, Proceedings, 2003. **2732**: p. 388-400.

179. Pickering, M.R., A.A. Muhit, J.M. Searvell, et al., *A New Multi-Modal Similarity Measure for Fast Gradient-Based 2D-3D Image Registration.* 2009 Annual International Conference of the Ieee Engineering in Medicine and Biology Society, Vols 1-20, 2009: p. 5821-5824.

180. He, K.M., J.A. Sun, and X.O. Tang, *Guided Image Filtering.* Computer Vision-Eccv 2010, Pt I, 2010. **6311**: p. 1-14.

181. Wold, S., A. Ruhe, H. Wold, et al., *The Collinearity Problem in Linear-Regression - the Partial Least-Squares (Pls) Approach to Generalized Inverses.* Siam Journal on Scientific and Statistical Computing, 1984. **5**(3): p. 735-743.

182. Naes, T. and B.H. Mevik, *Understanding the collinearity problem in regression and discriminant analysis.* Journal of Chemometrics, 2001. **15**(4): p. 413-426.

183. Weisberg, S., *Applied Linear Regression*. 2013: Wiley.

184. Mardia, K.V., J.T. Kent, and J.M. Bibby, *Multivariate analysis*. 1979, London: Academic Press.

185. Kutner, M.H., *Applied Linear Statistical Models*. 2005: McGraw-Hill Irwin.

186. McDonald, J.H. and U.o. Delaware, *Handbook of Biological Statistics*. 2009: Sparky House Publishing.

187. Chapelle, O., P. Haffner, and V.N. Vapnik, *Support vector machines for histogram-based image classification.* Ieee Transactions on Neural Networks, 1999. **10**(5): p. 1055-1064.

188. Fung, G. and J. Stoeckel, *SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information.* Knowledge and Information Systems, 2007. **11**(2): p. 243-258.

189. Gangeh, M., L. Sørensen, S. Shaker, et al., *A Texton-Based Approach for the Classification of Lung Parenchyma in CT Images*, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, T. Jiang, et al., Editors. 2010, Springer Berlin Heidelberg. p. 595-602.

190. Wolz, R., V. Julkunen, J. Koikkalainen, et al., *Multi-Method Analysis of MRI Images in Early Diagnostics of Alzheimer's Disease.* Plos One, 2011. **6**(10).

191. Santos, F., P. Guyomarc'h, and J. Bruzek, *Statistical sex determination from craniometrics: Comparison of linear discriminant analysis, logistic regression, and support vector machines.* Forensic Science International, 2014. **245**.

192. Breiman, L., *Random forests.* Machine Learning, 2001. **45**(1): p. 5-32.

193. Gray, K.R., P. Aljabar, R.A. Heckemann, et al., *Random forest-based similarity measures for multi-modal classification of Alzheimer's disease.* Neuroimage, 2013. **65**: p. 167-175.

194. Ko, B.C., S.H. Kim, and J.Y. Nam, *X-ray Image Classification Using Random Forests with Local Wavelet-Based CS-Local Binary Patterns.* Journal of Digital Imaging, 2011. **24**(6): p. 1141-1151.

195. Montillo, A., J. Shotton, J. Winn, et al., *Entangled Decision Forests and Their Application for Semantic Segmentation of CT Images.* Information Processing in Medical Imaging, 2011. **6801**: p. 184-196.

196. Liu, H., J. Li, and L. Wong, *A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns.* Genome Inform, 2002. **13**: p. 51-60.

197. Jiang, B., X.G. Zhang, and T.X. Cai, *Estimating the confidence interval for prediction errors of support vector machine classifiers.* Journal of Machine Learning Research, 2008. **9**: p. 521-540.

198. Nadeau, C. and Y. Bengio, *Inference for the generalization error.* Machine Learning, 2003. **52**(3): p. 239-281.

199. Jiang, W.Y., S. Varma, and R. Simon, *Calculating confidence intervals for prediction error in microarray classification using resampling.* Statistical Applications in Genetics and Molecular Biology, 2008. **7**(1).

200. Dong, X., M. Petrongolo, T. Niu, et al., *Low-dose and scatter-free cone-beam CT imaging using a stationary beam blocker in a single scan: phantom studies.* Comput Math Methods Med, 2013. **2013**: p. 637614.

201. Schulze, R., U. Heil, D. Groβ, et al., *Artefacts in CBCT: a review.* Dentomaxillofacial Radiology, 2011. **40**(5): p. 265-273.

202. Létourneau, D., R. Wong, D. Moseley, et al., *Online planning and delivery technique for radiotherapy of spinal metastases using cone-beam CT: Image quality and system performance.* International Journal of Radiation Oncology*Biology*Physics, 2007. **67**(4): p. 1229-1237.

203. Jaffray, D.A., J.H. Siewerdsen, J.W. Wong, et al., *Flat-panel cone-beam computed tomography for image-guided radiation therapy.* International Journal of Radiation Oncology*Biology*Physics, 2002. **53**(5): p. 1337-1349.

204. Zhu, J. and T. Hastie, *Classification of gene microarrays by penalized logistic regression.* Biostatistics, 2004. **5**(3): p. 427-443.

205. Park, M.Y. and T. Hastie, *L1-regularization path algorithm for generalized linear models.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2007. **69**(4): p. 659-677.

206. Panzeri, S., R. Senatore, M.A. Montemurro, et al., *Correcting for the sampling bias problem in spike train information measures.* Journal of Neurophysiology, 2007. **98**(3): p. 1064-1072.

207. Richards, J.A., *Remote Sensing Digital Image Analysis: An Introduction*. 2012: Springer.

208. Chawla, N.V., *Data Mining for Imbalanced Datasets: An Overview.* Data Mining and Knowledge Discovery Handbook, Second Edition, 2010: p. 875-886.

209. Ciraco, M., M. Rogalewski, and G. Weiss, *Improving classifier utility by altering the misclassification cost ratio*, in *Proceedings of the 1st international workshop on Utility-based data mining*. 2005, ACM: Chicago, Illinois. p. 46-52.

210. Batista, G.E.A.P.A., R.C. Prati, and M.C. Monard, *A study of the behavior of several methods for balancing machine learning training data.* SIGKDD Explor. Newsl., 2004. **6**(1): p. 20-29.

211. Chawla, N.V., K.W. Bowyer, L.O. Hall, et al., *SMOTE: Synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research, 2002. **16**: p. 321-357.

212. Krengli, M., S. Gaiano, E. Mones, et al., *Reproducibility of patient setup by surface image registration system in conformal radiotherapy of prostate cancer.* Radiation Oncology, 2009. **4**.

213. Munch, B., P. Trtik, F. Marone, et al., *Stripe and ring artifact removal with combined wavelet - Fourier filtering.* Optics Express, 2009. **17**(10): p. 8567-8591.

214. Niu, T., M. Sun, J. Star-Lack, et al., *Shading correction for on-board cone-beam CT in radiation therapy using planning MDCT images.* Med Phys, 2010. **37**(10): p. 5395-406.

215. Jin, J.Y., L. Ren, Q.A. Liu, et al., *Combining scatter reduction and correction to improve image quality in cone-beam computed tomography (CBCT).* Medical Physics, 2010. **37**(11): p. 5634-5644.

216. Zhang, Y.B., L.F. Zhang, R. Zhu, et al., *Reducing metal artifacts in cone-beam CT images by preprocessing projection data.* International Journal of Radiation Oncology Biology Physics, 2007. **67**(3): p. 924-932.

217. Choi, K., J. Wang, L. Zhu, et al., *Compressed sensing based cone-beam computed tomography reconstruction with a first-order method.* Medical Physics, 2010. **37**(9): p. 5113-5125.

218. Jia, X., B. Dong, Y.F. Lou, et al., *GPU-based iterative cone-beam CT reconstruction using tight frame regularization.* Physics in Medicine and Biology, 2011. **56**(13): p. 3787-3807.

219. Brock, K.K., L.A. Dawson, M.B. Sharpe, et al., *Feasibility of a novel deformable image registration technique to facilitate classification, targeting, and monitoring of tumor and normal tissue.* International Journal of Radiation Oncology Biology Physics, 2006. **64**(4): p. 1245-1254.

220. Burr, I.W., *Statistical Quality Control Methods*. 1976: Taylor & Francis.

221. Russo, S.L., M.E. Camargo, and J.P. Fabris, *Applications of Control Charts Arima for Autocorrelated Data*. Practical Concepts of Quality Control. 2012.

222. Zheng, Y.F., A. Barbu, B. Georgescu, et al., *Four-Chamber Heart Modeling and Automatic Segmentation for 3-D Cardiac CT Volumes Using Marginal Space Learning and Steerable Features.* Ieee Transactions on Medical Imaging, 2008. **27**(11): p. 1668-1681.

223. Zheng, Y.F., A. Barbu, B. Georgescu, et al., *Fast automatic heart chamber segmentation from 3D CT data using marginal space learning and steerable features.* 2007 Ieee 11th International Conference on Computer Vision, Vols 1-6, 2007: p. 762-769.

224. Gao, L.M., D.G. Heath, B.S. Kuszyk, et al., *Automatic liver segmentation technique for three-dimensional visualisation of CT data.* Radiology, 1996. **201**(2): p. 359-364.

225. Pan, S.Y. and B.M. Dawant, *Automatic 3D segmentation of the liver from abdominal CT images: a level-set approach.* Medical Imaging: 2001: Image Processing, Pts 1-3, 2001. **2**(27): p. 128-138.

226. Costa, M.J., H. Delingette, S. Novellas, et al., *Automatic segmentation of bladder and prostate using coupled 3D deformable models.* Medical Image Computing and Computer-Assisted Intervention - Miccai 2007, Pt 1, Proceedings, 2007. **4791**: p. 252-260.

227. Stamatis, D.H., *Failure Mode and Effect Analysis: FMEA from Theory to Execution*. 2003: ASQ Quality Press.

228. Ciocca, M., M.-C. Cantone, I. Veronese, et al., *Application of Failure Mode and Effects Analysis to Intraoperative Radiation Therapy Using Mobile Electron Linear Accelerators.* International Journal of Radiation Oncology*Biology*Physics, 2012. **82**(2): p. e305-e311.

229. Denny, D.S., D.K. Allen, N. Worthington, et al., *The Use of Failure Mode and Effect Analysis in a Radiation Oncology Setting: The Cancer Treatment Centers of America Experience.* Journal for Healthcare Quality, 2014. **36**(1): p. 18-28.

230. Perks, J.R., S. Stanic, R.L. Stern, et al., *Failure Mode and Effect Analysis for Delivery of Lung Stereotactic Body Radiation Therapy.* International Journal of Radiation Oncology*Biology*Physics, 2012. **83**(4): p. 1324-1329.

231. Terezakis, S.A., P. Pronovost, K. Harris, et al., *Safety strategies in an academic radiation oncology department and recommendations for action.* Jt Comm J Qual Patient Saf, 2011. **37**(7): p. 291-9.

232. Bureau of Labor Statistics (2014) *Occupational Outlook Handbook 2014-15 Edition, Radiation Therapists*.

233. Kalapurakal, J.A., A. Zafirovski, J. Smith, et al., *A Comprehensive Quality Assurance Program for Personnel and Procedures in Radiation Oncology: Value of Voluntary Error Reporting and Checklists.* International Journal of Radiation Oncology Biology Physics, 2013. **86**(2): p. 241-248.