Title

Consolidating the Safety of Cone-Beam Computed Tomography Guided Radiotherapy through a Deep Learning-Based Patient Setup Error Detection System.

Author

Luximon, Dishane Chand

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Consolidating the Safety of Cone-Beam Computed Tomography Guided Radiotherapy through a Deep Learning-Based Patient Setup Error Detection System.

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Physics and Biology in Medicine

by

Dishane Chand Luximon

2024

ABSTRACT OF THE DISSERTATION

Consolidating the Safety of Cone-Beam Computed Tomography Guided Radiotherapy through a Deep Learning-Based Patient Setup Error Detection System.

by

Dishane Chand Luximon

Doctor of Philosophy in Physics and Biology in Medicine

University of California, Los Angeles, 2024

Professor James Michael Lamb, Chair

Although safety procedures are in place within the radiation therapy (RT) workflow, incidents are still occurring due to human errors. To enhance patient safety, it is critical to identify and consolidate the vulnerabilities present within the various processes performed prior and during treatment. One such domain is within the Cone Beam Computed Tomography (CBCT)-guided RT workflow where there is currently no built-in system to check for errors in the registration of the simulation Computed Tomography (simCT) to the setup CBCT performed during the patient positioning step prior to radiation beam delivery. This lack of safeguards poses a risk to the patient as an incorrect registration may go undetected, leading to a compromised patient setup and treatment. The overall objective of the proposed work is to develop a deep learning-based error detection algorithm (EDA) which can serve as a secondary safety check to the radiation therapist while also helping to consolidate the robustness of CBCT-guided radiotherapy treatments.

Additionally, this work explores the feasibility of a fully-unsupervised anomaly detection framework (ADF), based on a CBCT inpainting technique using a variational autoencoder, which would highlight anomalies for human review during regular physics quality assurance chart checks.

Initial results show that the EDA has a strong error-catching ability with areas under the receiver operating characteristic (ROC) curves of at least 99.2% when tested on simulated translational errors. When assessed against expert observers in a qualitative assessment of patient setup registrations, the EDA's predictions achieved statistically significant correlations to the observer scores. Additionally, during a retrospective error search on a multi-institutional dataset of 17,612 registrations, the EDA successfully flagged the three known patient-setup incidents, and additionally identified four previously unreported incidents, proving its effectiveness on real-life cases. Those results validated the clinical utility of the EDA for bulk image reviews and highlighted the reliability and safety of CBCT-guided RT, with an absolute gross patient misalignment error rate of 0.04% ± 0.02% per delivered fraction. The ADF also demonstrated promising error detection ability when applied to a test dataset containing real patient setup incidents and simulated translational errors, with an area under the ROC curve of 98.1%.

The results described in this work validate the clinical utility and strong error-catching ability of both the EDA and the ADF when applied to real-world cases. We demonstrated that EDA and ADF can facilitate bulk image reviews, which can be useful for incident learning and can also expedite regular quality assurance chart checks performed by physicists or physicians. Additionally, if applied in real-time, EDA can consolidate the safety of CBCT-guided radiotherapy by serving as a secondary safety check to the therapist, thereby minimizing the risk of gross patient setup errors. Whether used prospectively or retrospectively, we believe that the

proposed tools can add substantial value to the safety aspect of radiotherapy treatments, especially in low-middle income communities where the lack of workforce and safeguards often translates into a higher risk of treatment incidents.

The dissertation of Dishane Chand Luximon is approved.

Matthew Sherman Brown

Minsong Cao

Daniel Abraham Low

John Paul Neylon

Timothy A. Ritter

James Michael Lamb, Committee Chair

University of California, Los Angeles

2024

DEDICATION

To Mama, Papa, Aji and Aja.

Thank you for showing me the power of resilience.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

tr_{90}	Threshold Resulting in Sensitivity ≥90%
tr_{99}	Threshold Resulting in Sensitivity ≥99%
2D	Two Dimensional
3D	Three Dimensional
AAPM	American Association of Physicists in Medicine
ACR	American College of Radiology
ADF	Anomaly Detection Framework
AI	Artificial Intelligence
ARL	Anatomical Region Labeling
ASTRO	American Society for Radiation Oncology
AUC	Area under Curve
BCE	Binary Cross-entropy
CBCT	Cone Beam Computed Tomography
CNN	Convolutional Neural Network
CT	Computed Tomography
CTV	Clinical Target Volume
D95	Minimum dose delivered to 95% of the target
DB	Dense Block
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DQR	DICOM Query and Retrieval
EBRT	External Beam Radiation Therapy
EDA	Error Detection Algorithm
EMR	Electronic Medical Records
EX	Extremity
FMEA	Failure Modes and Effects Analysis
fn	False Negative
FOV	Field-of-view
fp	False Positive
GMS	Gradient Magnitude Similarity
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
gtCBCT	Ground Truth Cone Beam Computed Tomography
GTV	Gross Tumor Volume
Gy	Gray
HN	Head & Neck

HU	Hounsfield Unit
IGRT	Image-Guided Radiotherapy
IMRT	Intensity-modulated Radiation Therapy
IRB	Institutional Review Board
KL	Kullback–Leibler
kV	Kilovoltage
LINAC	Linear Accelerator
MCC	Matthews Correlation Coefficient
MI	Mutual Information
MPPG	Medical Physics Practice Guidelines
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
MV	Megavoltage
OAR	Organ at Risk
OVBM	Off-by-one Vertebral Body Misalignment
PCA	Principal Component Analysis
PL	Pelvis
PTV	Planning Target Volume
QA	Quality Assurance
R&V	Record and verify
RAM	Random Access Memory
reconCBCT	Reconstructed Cone Beam Computed Tomography
REG	Registration
ROC	Receiver Operating Characteristic
RO-ILS	Radiation Oncology Incident Learning System
RT	Radiation Therapy
SA	Specific Aim
SBRT	Stereotactic Body Radiation Therapy
SCS	Spinal Canal Segmentation
simCT	Simulation Computed Tomography
SRS	Stereotactic Radiosurgery
SSIM	Structural Similarity Index Measure
SVM	Support Vector Machine
TA	Thoracic-abdominal
TB	Transition Block
TG	Task Group
tn	True Negative
tp	True Positive
UCLA	University of California, Los Angeles
VAE	Variational Autoencoder

VCU	Virginia Commonwealth University
VRAM	Video Random-Access Memory

Acknowledgements

Firstly, I would like to thank my advisor, Dr. James Lamb for his guidance and mentorship throughout my Ph.D. journey. You have not only taught me how to think and act with integrity as a researcher, but also showed me what it takes to be a supportive and visionary leader. I am certain that the values I have learnt from you will follow me throughout my career and life. To the rest of my dissertation committee – Dr. Brown, Dr. Cao, Dr. Low, Dr. Neylon, and Dr. Ritter, thank you for your expertise, constant motivation, and invaluable support. I feel very honored to have such a great committee with so much knowledge and experience.

I would also like to thank the Physics & Biology in Medicine (PBM) faculty for all the knowledge they instilled in me during my graduate studies. To the program directors, Dr. Michael McNitt-Gray and Dr. Magnus Dahlbom, I will be forever grateful to you for believing in me and accepting me into the PBM family – and thank you for creating such a wonderful program environment for us, students. Big thanks to Reth and Alondra, for all of your support.

Thanks to the whole ROSA-ML research group – Rachel, John, Philip, Yasin, Justin. This work would not have been possible without your contributions, friendship, and support. Thanks to all PBM students for all the wonderful times we had together. Special mentions to Pav, Louise, John, Jonathan, and Peter, my graduate years would not have been the same without you all.

None of my accomplishments would have been possible without the unwavering support from my family. To my parents and grandparents, I will always be grateful for all the sacrifices you have made in your life so that your children and grandchildren could get the opportunities you were never able to get. Thank you for all the values you have taught me, which made me into the

person I am today. To my siblings, Dion, Mishla and Mahita, thank you for all of your support and guidance, you three are my role models and driving force in life. And finally, to Mohini, who has been by my side throughout this journey. Thank you for your patience, love, and support, and for being my main source of positivity and inspiration at times I needed it the most.

Chapter 2 is a version of Luximon DC, Ritter T, Fields E, Neylon J, Petragallo R, Abdulkadir Y, Charters J, Low DA, Lamb JM. Development and interinstitutional validation of an automatic vertebral-body misalignment error detector for cone-beam CT-guided radiotherapy. Medical Physics. 2022 Oct;49(10):6410-23.

Chapter 3 is a version of Luximon DC, Neylon J, Lamb JM. Feasibility of a deep-learning based anatomical region labeling tool for Cone-Beam Computed Tomography scans in radiotherapy. Physics and Imaging in Radiation Oncology. 2023 Jan 1;25:100427.

Chapter 4 is a version of Neylon J, Luximon DC, Ritter T, Lamb JM. Proof-of-concept study of artificial intelligence-assisted review of CBCT image guidance. Journal of Applied Clinical Medical Physics. 2023 Sep;24(9):e14016.

Chapter 5 is a version of Luximon DC, Neylon J, Ritter T, Agazaryan N, Hegde JV, Steinberg ML, Low DA, Lamb JM. Results of an Artificial Intelligence–Based Image Review System to Detect Patient Misalignment Errors in a Multi-institutional Database of Cone Beam Computed Tomography–Guided Radiation Therapy. International Journal of Radiation Oncology* Biology* Physics. 2024 Mar 12.

Chapter 6 is a version of Luximon DC, Petragallo R, Neylon J, Ritter T, Lamb JM. An Unsupervised Anomaly Detection Framework For CBCT Setup Images And Registrations In Image-Guided Radiotherapy Using A Variational Auto-Encoder. (Manuscript in Preparation)

The research reported in this dissertation was supported by the Agency for Healthcare Research and Quality (AHRQ) under award number 1R01HS026486.

Vita

EDUCATION

M.S. Physics & Biology in Medicine, University of California, Los Angeles 2021
B.A. Physics, Connecticut College 2019

AWARDS

Moses A. Greenfield Award (University of California, Los Angeles) 2023
Graduate Council Diversity Fellowship (University of California, Los Angeles) 2021

PEER-REVIEWED PUBLICATIONS

1. Petragallo R, **Luximon DC**, Neylon J, Bardach NS, Ritter T, Lamb JM. Clinical physicists' perceptions of weekly chart checks and the potential role for automated image review assessed by structured interviews. Journal of Applied Clinical Medical Physics. 2024 Apr 22:e14313.

2. **Luximon DC**, Neylon J, Ritter T, Agazaryan N, Hegde JV, Steinberg ML, Low DA, Lamb JM. Results of an Artificial Intelligence–Based Image Review System to Detect Patient Misalignment Errors in a Multi-institutional Database of Cone Beam Computed Tomography–Guided Radiation Therapy. International Journal of Radiation Oncology* Biology* Physics. 2024 Mar 12.

3. Charters JA, **Luximon D**, Petragallo R, Neylon J, Low DA, Lamb JM. Automated detection of vertebral body misalignments in orthogonal kV and MV guided radiotherapy: application to a comprehensive retrospective dataset. Biomedical Physics & Engineering Express. 2024 Feb 21.

4. **Luximon DC**, Neylon J, Lamb JM. Feasibility of a deep-learning based anatomical region labeling tool for Cone-Beam Computed Tomography scans in radiotherapy. Physics and Imaging in Radiation Oncology. 2023 Jan 1;25:100427.

5. Abdulkadir Y, **Luximon D**, Morris E, Chow P, Kishan AU, Mikaeilian A, Lamb JM. Human factors in the clinical implementation of deep learning-based automated contouring of pelvic organs at risk for MRI-guided radiotherapy. Medical physics. 2023 Oct;50(10):5969-77.

6. Neylon J, **Luximon DC**, Ritter T, Lamb JM. Proof-of-concept study of artificial intelligence-assisted review of CBCT image guidance. Journal of Applied Clinical Medical Physics. 2023 Sep;24(9):e14016.

7. **Luximon DC**, Ritter T, Fields E, Neylon J, Petragallo R, Abdulkadir Y, Charters J, Low DA, Lamb JM. Development and interinstitutional validation of an automatic vertebral-body misalignment error detector for cone-beam CT-guided radiotherapy. Medical Physics. 2022 Oct;49(10):6410-23.

8. **Luximon DC**, Abdulkadir Y, Chow PE, Morris ED, Lamb JM. Machine-assisted interpolation algorithm for semi-automated segmentation of highly deformable organs. Medical physics. 2022 Jan;49(1):41-51.

SELECTED CONFERENCE PRESENTATIONS

1. **Luximon DC**, Neylon J, Ritter T, Agazaryan N, Hegde JV, Steinberg ML, Low DA, & Lamb JM. Results of an AI-Based Image Review System to Detect Patient Setup Errors in a Multi-Institutional Database of CBCTGuided Radiotherapy Treatments. AAPM 65th Annual Meeting: Houston TX, July 2023.

2. **Luximon DC**, Neylon J, & Lamb JM. Feasibility of a deep-learning based anatomical region labeling tool for Cone-Beam Computed Tomography scans in radiotherapy. AAPM 65th Annual Meeting: Houston TX, July 2023.

3. **Luximon DC**, Neylon J, Ritter T, & Lamb J.M. Consolidating the safety of CBCT-guided radiotherapy using an AI-based patient setup error detection system. Global Health Catalyst Summit, Philadelphia PA, May 2023.

4. **Luximon DC**, Ritter T, Fields E, Neylon J, Petragallo R, Abdulkadir Y, Charters J, Low DA, & Lamb JM. Development and Inter-Institutional Validation of an Automatic Vertebral-Body Misalignment Error Detector for Cone-Beam CT Guided Radiotherapy. AAPM 64th Annual Meeting, Washington D.C., July 2022.

5. **Luximon, DC**, Ritter T, Fields E, Neylon J, Petragallo R, Abdulkadir Y, Charters J, Low DA, & Lamb JM. Automatic Detection of Vertebral Body Misalignment Errors in Cone-Beam CT-guided Radiotherapy Using a Convolutional Neural Network. Conference on Machine Intelligence in Medical Imaging, September 2021 (conducted remotely due to the Covid-19 pandemic).

6. **Luximon DC**, Ritter T, Fields E, Neylon J, Petragallo R, Abdulkadir Y, Charters J, Low DA, & Lamb JM. Automatic Detection of Vertebral Body Misalignment Errors in Cone-Beam CT-guided Radiotherapy Using a Convolutional Neural Network. AAPM 63rd Annual Meeting, July 2021 (conducted remotely due to the Covid-19 pandemic).

7. **Luximon DC**, Abdulkadir Y, Chow PE, Morris ED, & Lamb JM. Machine-assisted interpolation algorithm for semi-automated segmentation of highly deformable organs. AAPM 63rd Annual Meeting, July 2021 (conducted remotely due to the Covid-19 pandemic).

Chapter 1: Introduction

1.1 External Beam Radiation Therapy Workflow

Radiation therapy, or radiotherapy, is a medical treatment whereby high-energy radiation is used to damage the DNA of abnormal cells (commonly cancerous cells), thereby inducing their death. [1] In cancer treatments, RT can be used for both curative purposes, alongside other treatments such as surgery or chemotherapy, or for palliative purposes to alleviate symptoms and improve the quality of life of patients with advanced stages of cancer. According to the World Health Organization, more than half of cancer patients receive radiation therapy treatment at some stage of their treatment course. [2] Based on a 2016 Surveillance, Epidemiology, and End-Results database linked to U.S. Census data, there were an estimated 3.05 million 5-year cancer survivors treated with radiation in 2016, accounting for 29% of all cancer survivors. [3]

There are currently different types of radiation therapy, including external beam radiation therapy (EBRT) and internal radiation therapy (such as brachytherapy). EBRT is a non-invasive treatment which involves the use of high energy radiation to damage and kill diseased cells (e.g. cancerous cells). This treatment type is continuously evolving, allowing more precise and highly conformal radiation delivery to the treatment target, while minimizing excess radiation dose to healthy cells. This has been shown to improve both tumor control and treatment outcomes, while minimizing the side effects caused by ionizing radiation. [4-5]

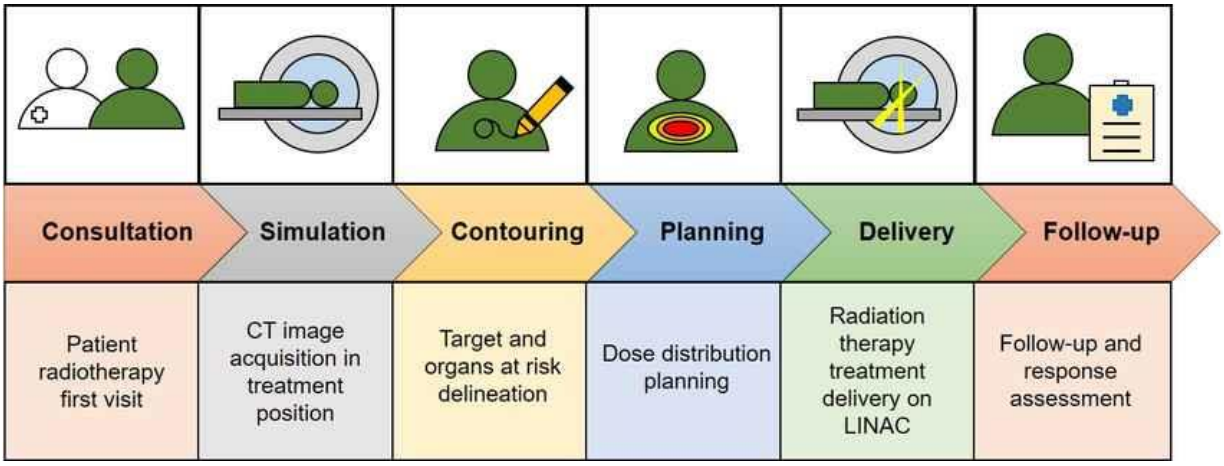


Figure 1.1: Overview of the external beam radiation therapy workflow from Marvaso et al. [120]

Figure 1.1 above shows a summary of the EBRT workflow. Following a consultation with the radiation oncologist to discuss the role, risks, and side effects of EBRT, the patient will undergo the simulation phase of the treatment process. This simulation phase involves the acquisition of a three-dimensional (3D) computed tomography (CT) scan of the patient, which will be used to plan the EBRT treatment. During the CT simulation, radiation therapists will also delineate the target region on the patient's body, marking pertinent locations on the skin—typically utilizing permanent tattoos—to obtain a frame of reference for patient setup prior to the treatment delivery. Additionally, immobilization devices, such as custom-molded masks or body cushions, may be used to securely stabilize specific body parts of the patient.

The 3D CT dataset, referred to as the simulation CT (simCT), will then be transmitted to the dosimetry team for the treatment planning phase. During this stage, the radiation oncologists will collaborate closely with dosimetrists and medical physicists to plan the patient’s treatment. Dosimetrists will first outline non-cancerous organs neighboring the treatment region, known as organs at risk (OARs), that may be affected by the EBRT treatment. The radiation oncologists

will then contour the pertinent treatment region(s) and provide the dosimetrists and medical physicists with precise treatment prescriptions, specifying the dose tolerances for both the target region and the OARs. Based on those prescribed doses, the medical physicists and/or dosimetrists will tailor the treatment plan by optimizing the radiation dose delivery to the treatment region while minimizing the dose to the OARs.

Upon approval of the treatment plan by the radiation oncologist, the patient will then proceed to the treatment delivery phase. Prior to the radiation dose delivery, the radiation therapists will be in charge of positioning and immobilizing the patient at the treatment machine, such that the patient is positioned exactly as in the CT simulation phase. Following this positioning step, the prescribed radiation dose will be delivered to the patient. Throughout the treatment course, the radiation oncologist will diligently monitor the patient's progress through routine weekly assessments and progress notes, while medical physicists will conduct comprehensive weekly chart checks to ensure that the patient has been receiving the intended treatment.

1.2 Patient Setup in Cone Beam CT-guided Radiation Therapy

In order to facilitate the patient positioning step prior to radiation dose delivery, image guidance is used in modern EBRT treatments. During this process, verification imaging scans such as two-dimensional (2D) X-rays, 3D CT scans, or 3D Magnetic Resonance Imaging (MRI) scans are acquired while the patient is on the treatment table to visualize the target area and surrounding anatomy on the treatment day. The verification scan is then registered, or aligned, with the planning scan to correct any discrepancy between the planned and actual positions of the target regions and OARs. In modern radiotherapy linear accelerator (LINAC) systems, the

registration between the planning scan and the setup scan is either done automatically using dedicated software integrated in the image verification platform, or manually by the radiation therapists. In either case, the registration is manually verified by the radiation therapists, and depending on clinical procedures in place, may require the radiation oncologist and/or physicist to approve the registration prior to dose delivery. Once verified and approved, the final couch shifts are applied remotely to place the patient in the required treatment position, and the radiation dose is delivered to the target volume, as per the planned treatment.

One of the most common setup verification imaging systems used in EBRT is the cone beam computed tomography (CBCT). In CBCT, a cone-shaped kilovoltage (kV) X-ray beam is used to acquire images of the patient's anatomy from various angles around the patient body. This cone-shaped beam diverges outward from the X-ray source and intersects with a two-dimensional flat-panel detector that captures the transmitted X-rays after they go through the patient's body. The acquired X-ray images are then processed using computer algorithms to reconstruct a 3D image of the patient's anatomy, similar to the conventional CT scanning, where multiple 2D X-ray projections are processed to create cross-sectional images.

An advantage of the CBCT imaging system is that it provides detailed information about the patient’s internal structures, and hence allows radiation therapists to precisely locate the target volume and accurately align the patient. [6] Additionally, the acquisition time of a CBCT ranges from 30 seconds to 2 minutes (depending on the imaging protocols and anatomy to be imaged), making this imaging system very convenient for daily setup verification during the treatment course. [7-8] In modern EBRT units, the kV source and flat-panel detector are included within the treatment machine to facilitate and expedite the image acquisition and setup verification process while the patient is on the treatment couch.

In CBCT-guided radiotherapy treatments, the setup CBCT scan is registered to the simCT scan to properly align the patient prior to beam delivery. However, one pitfall of this current setup verification system is that it does not include any tool to assess the quality of the registration being performed and alert the radiation therapist in case of gross registration errors. Hence, if the setup CBCT and the simCT are incorrectly registered, the treatment couch may be shifted to a wrong position, and the radiation dose may be delivered to the incorrect structure, deviating from the planned treatment. To this day, unless a person at the treatment console visually identifies the registration error and corrects it, there is no secondary barrier to such incident occurring and hence poses a risk to the patient.

1.3 Patient Setup Incidents in Radiotherapy

Studies have shown that incidents regarding incorrect dose delivery are still prevalent even with quality assurance protocols and technologies in place, such as image-guidance. [9-11] Between 2014 and March 2023, a total of 3,730 therapeutic radiation incidents were reported to the Radiation Oncology Incident Learning System (RO-ILS) Portal, with 18.4% of those being identified as having severe or moderate severity scores. [9]

According to the American Association of Physicists in Medicine (AAPM) Task Group 100, the patient positioning step within the EBRT workflow is a high severity and high-risk failure mode, ranking in the top 20% most hazardous steps in the entire external radiotherapy workflow following a failure modes and effects analysis (FMEA). [12] With the widespread adoption of stereotactic radiosurgery and stereotactic body radiation therapy (SRS/SBRT) and the introduction of ultrahigh-dose-rate treatments such as FLASH, proper patient setups become even more critical due to their high-dose per fraction. In a 2017 study of RO-ILS SRS/SBRT

events, Hoopes et al. have found that one of the most common event types was the incorrect shift and alignment of the patient. [13]

McGurk et al. further reinforced this finding when they discovered, in a 2023 study involving four institutions in the United States, a patient who was wrongly aligned for one of five of their multilevel spine SBRT fractions. [14] Additionally, Ezzell et al. have shown that in a cohort of 336 critical events submitted to RO-ILS between 2014 and 2016, 34 such errors occurred due to the wrong shift performed at the treatment table, with 29 of them reaching the patient. [15] This report also highlighted a T12-L5 spine case where the automatic registration of the CBCT was incorrect by 3 cm in the superior–inferior direction for the first two fractions of a five-fraction treatment. The error was only captured on the third fraction when the therapists realized that something was wrong and called the physicist for a review. Although the outcome of this treatment is unknown, this incident demonstrates the risks involved when relying solely on human perception to catch errors.

In the thoracic region particularly, there is a higher risk of these types of errors occurring due to the similarity between adjacent the vertebral bodies, which are often used as landmark during the registration process in image-guided radiotherapy (IGRT). This region is also prone to motion artifacts, which can complicate the registration process. Shah et al. have shown, for example, that the anatomical variations and anomalies in the thoracic vertebra and surrounding regions can cause improper labeling of vertebral bodies and contribute to wrong-level spine surgery. [16] Furthermore, between 2015 and 2017, the French Nuclear Safety Authority reported 40 events related to vertebral body misalignments between the planning and setup scans. [17]

1.4 Error prevention in Radiotherapy

The American College of Radiology (ACR)-American Society for Radiation Oncology (ASTRO) Practice Parameter for Image-Guided Radiotherapy, the AAPM Task Group (TG)-275 and the Medical Physics Practice Guidelines (MPPG) 11.a all provide recommendations on how to mitigate alignment-based failure modes and promote incident learning as a way to reduce future events. [18-20] However, those reports note that many components of the current safety checks, including those involving patient setups, are heavily human-reliant and are therefore prone to be overlooked due to the fast-paced working conditions in the clinic. For instance, McGurk et al. have shown through a Human Factor Analysis and Classification System that 95.2% of 189 reported SBRT safeguard failures occurred due to human errors. [14]

As recommended by TG-275 and MPPG 11.a, the use of automation during these safety checks can act as a safety barrier and help in the analysis of bulk data for efficient incident learning by identifying error pathways that may not be easily detected by a human reviewer.

1.5 Using Automation for Patient Setup Error Mitigation in EBRT

With the constantly evolving technologies in EBRT, comes the need for error-mitigating systems that can reduce the risk of setup errors and make EBRT safer for the patient. Although some have been trying to solve this problem with real-time monitoring systems using camera tracking,[21-23] others have proposed the use of automated processes to detect setup errors by analyzing IGRT images acquired before beam delivery. [24] As no additional equipment is required in the latter solution beyond what is used for IGRT, it is more cost-effective and potentially accessible to a larger number of facilities.

Jani et al. have developed an automated system for the detection of patient identification and setup errors in EBRT using setup kilovoltage CT images and simCT images. [25] Their work made use of image similarity metrics as features, which were applied to a linear discriminant analysis for the error classification. Although this classical machine learning method produced acceptable results in classifying wrong-vertebral-body errors, it was limited by the feature selection, which relied on human observation for pattern recognition.

Deep learning, on the other hand, can automatically determine and extract high-level features from raw data, which allows it to obtain patterns undiscernible by human observation. [26] Convolutional neural networks [27] (CNN) have previously been used for image classification problems and this has huge potential in the field of medical imaging. [28-30] CNNs have allowed the transition from conventional machine learning relying on observable and user-defined features to trainable deep neural networks using large-scale image data. Using its ability to determine and extract high-level features from raw data, a CNN model is able obtain patterns undiscernible by human observation and has shown remarkable performance in many computer vision tasks. In the medical field, several deep learning methods have been proposed for disease or tissue characterization, diagnosis, and prognosis. [31-32] Additionally, deep learning techniques have been found to be promising for automated detection of patient setup misalignments for both 2D x-ray-guided radiotherapy and orthogonal kV-Megavoltage (MV)-guided radiotherapy. [33-34]

1.6 Overview and Specific Aims

The overall objective of this dissertation is to consolidate the safety of CBCT-guided radiotherapy by developing a fully automatic deep-learning based setup error detection system which can not only reduce incidents but can also aid in incident learning by retrospectively analyzing registrations from past treatments, and supplement physicists in regular physics chart checks. The following specific aims (SA) outline the approach to meeting this objective:

Specific Aim 1: Develop a fully-automatic deep-learning based gross setup error detection algorithm for CBCT-guided radiotherapy using fully-supervised learning.

Specific Aim 2: Conduct a retrospective study on registrations performed during CBCT-guided radiotherapy to catch and analyze reported and unreported setup errors.

Specific Aim 3: Develop a 3D and fully-unsupervised error detection framework for anomaly detection in CBCT-guided radiotherapy.

Chapter 2 and **Chapter 3** address SA1, developing and assessing a fully-supervised deep learning-based gross patient setup error detection algorithm (EDA) for CBCT-guided radiotherapy using simulated errors at two different institutions. In **Chapter 4**, the EDA described in **Chapter 2** and **Chapter 3** is clinically implemented for daily quantitative analysis of CBCT-guided patient setup registrations and is validated against independent expert observers. **Chapter 5** addresses SA2, applying the AI-based EDA pipeline described in **Chapter 4** to perform a fully-automated retrospective error search on clinical CBCT-guided RT image databases and determining an absolute gross patient misalignment error rate. In **Chapter 6**, a fully-unsupervised method based on a Variational Autoencoder (VAE) is developed to detect anomalies in CBCT-guided RT images, providing a solution to SA3.

Chapter 2: Development and inter-institutional validation of an automatic patient setup misalignment error detector for Cone-Beam CT-guided Radiotherapy

2.1 Introduction

The technologies behind External Beam Radiation Therapy (EBRT) are continuously evolving to enhance treatment planning and beam delivery. The use of Image Guided Radiotherapy (IGRT), for example, has allowed for more precise and highly conformal beam delivery and treatment planning. [35] While these technologies promise to reduce setup uncertainties, they also bring more complexities to the EBRT processes, which may increase the risk of incidents in the absence of safeguards. [36-37] Lack of experience, inadequate procedures, inattention, and miscommunications between therapists may result in setup and treatment errors. [38-40]

Hence, with these new evolving technologies comes the need for error-mitigating systems that can reduce the risk of setup errors and make EBRT safer for the patient. While some have been trying to solve this problem with real-time monitoring systems using camera tracking [21-23], others have proposed the use of automated processes to detect setup errors by analyzing IGRT images acquired before beam delivery [24]. As no additional equipment is required in the latter solution beyond what is used for IGRT, it is more cost-effective and potentially accessible to a larger number of facilities.

Jani et al. have developed an automated system for the detection of patient identification and setup errors in EBRT using setup kilovoltage CT images and simCT images. [25] Their work made use of image similarity metrics as features, which were applied to a linear discriminant

analysis for the error classification. While this classical machine learning method produced acceptable results in classifying patient setup misalignment errors, it was limited by feature selection, which relied on human observation for pattern recognition. Deep learning, on the other hand, can automatically determine and extract high-level features from raw data, which allows it to obtain patterns undiscernible by human observation. [26] Convolutional neural networks [27] (CNN) have previously been used for image classification problems and this has huge potential in the field of medical imaging. [28-30] Several deep learning methods have been proposed for disease or tissue characterization, diagnosis, and prognosis. [31-32] However, to this day, the use of deep learning has yet to be applied to CBCT-guided radiotherapy setup error detection.

The long-term goal of this project is to develop a fully-automated error detection system that can act as a real-time secondary barrier to prevent gross patient misalignments from occurring in the clinic. Additionally, by analyzing all the treatment scans performed within a user-defined time and flag possible anomalies, this tool could potentially aid and supplement regular chart checks performed by medical physicists for quality assurance of CBCT-guided EBRT. However, for successful clinical implementation, it is essential to have a tool that minimally disrupts the clinical workflow due to false positives. Hence, in the development of our tool, a large focus was placed on the model's ability to catch gross patient misalignments with a threshold value that leads to less than 1% of false positives, which can be deemed acceptable in comparison to other false-positive interrupts and interlocks in the clinical workflow.

Inter-institutional validation is also key in assessing a deep learning model's generalizability power on a variety of patients, registration practices, image quality, and scanning protocols. In this study, which included patient data from two different institutions, the performance of the tool on cross-institutional data was investigated. Such experiment could be

helpful in determining the generalizability and performance of the tool on multi-institutional data.

2.2 Materials and Methods

2.2.1 Methods Overview

The type of misalignment may vary by treatment site based on the landmarks used during the registration of the simCT to the setup CBCT. For example, in the thoracic and abdominal regions where the vertebral bodies are often used as landmarks, off-by-one vertebral body misalignments may be a potential source of error. For all regions, the risk of smaller 5-10 mm misalignments may be present due to tumor size variability and anatomy changes. Some anatomy changes that may occur in between the simulation and treatment day include loss of weight, brain swelling for head & neck treatments, lung, stomach and bowel shape variations for thoracic-abdominal cases, as well as bladder and bowel size and composition variations for pelvic cases. As different features can be used to catch errors in each of these anatomical regions, one independent error-check pipeline was constructed for each of the three anatomical regions specified above. An anatomy region labeling model, as described in **Chapter 3**, was also introduced into our algorithm such that the patient scan could be automatically sent to the appropriate error-check pipeline for image analysis and error classification.

2.2.2 Dataset Acquisition

Under an IRB approved protocol, simCTs and CBCTs were collected from 580 patients undergoing radiotherapy treatment at the University of California, Los Angeles Medical Center (Institution #1) between 2018 and 2022. The treatments at Institution #1 had been performed on

three TrueBeam and one NovalisTx linear accelerator treatment machines (Varian Medical Systems, California, United States). From those 580 patients, 3,316 clinically aligned planning CT- CBCT pairs were obtained and used in our work. Additionally, 100 patient datasets were collected from patients undergoing radiotherapy treatment at the Virginia Commonwealth University Medical Center (Institution #2) between 2017 and 2019. The patients at Institution #2 had been treated on Varian Trilogy and TrueBeam linear accelerator treatment machines and all patients were treated in the thoracic-abdominal regions. For each CBCT acquired from the two facilities, a Registration (REG) file in the DICOM format was extracted to obtain the clinically applied alignment. Additionally, the RT Structure file for each planning CT was collected.

The patient data from Institution #1 was collected using an in-house DICOM query and retrieval (DQR) application programming interface using the pynetdicom* Python package. Our custom DQR software allowed automatic retrieval of patient data from the ARIA image management system (Varian Medical Systems) based on user-defined date ranges, plan names, and image types. This tool was built to fully automate our data acquisition protocol, thereby allowing the possibility of a fully-automated error detection pipeline. The open-access script for this DQR application is available on the following website: https://rosaml.net/.

Based on the treatment site, three general anatomy regions were used to classify the patient datasets: head & neck (HN), thoracic/abdomen (TA), and pelvis (PL). For each anatomical region, specific error types were simulated. For the TA region, off-by one vertebral body misalignments were simulated as it is a well-known error type and risk to the patient. For the other regions, including HN and PL, the risk of smaller, variable 5-10 mm misalignments

* https://pydicom.github.io/pynetdicom/stable/#

may occur due to tumor growth or shrinkage, anatomy changes, and patient pose changes. However, depending on several factors, such as tumor location and size, the 5-10 mm misalignments may lead to variable treatment impact. In order to have an overall balance between clinical significance of the misalignment and a practical implementation, our definition of an error required a 10 mm misalignment for those regions. Hence 10mm misalignments in randomly chosen directions were simulated for HN and PL cases. For each region, the dataset was subsequently randomly split into a training, validation and test set using the anonymized unique patient identifiers, as shown in Table 2.1.

Table 2.1: Description of the dataset used to train and test the deep-learning models in the error detection algorithm. The dataset, composed of images from two radiotherapy sites (Institution1/Institution2), was partitioned into a training, validation, and testing set for each treatment site based on unique patient identifiers.

Treatment Region	Simulated Error Type	Dataset Partition	Number of Patients	CBCT Image Pairs		
				Total	Aligned	Misaligned
Thoracic/Abdominal	OVBM	Training (Institution1/Institution2)	374 (304/70)	1887 (1677/210)	1139 (1069/70)	748 (608/140)
		Validation (Institution1/Institution2)	39 (29/10)	186 (156/30)	108 (98/10)	78 (58/20)
		Testing (Institution1/Institution2)	67 (47/20)	303 (243/60)	169 (149/20)	134 (94/40)
Head & Neck	10 mm shift	Training (Institution1 only)	60	912	456	456
		Validation (Institution1 only)	10	76	38	38
		Testing (Institution1 only)	30	354	177	177
Pelvis	10 mm shift	Training (Institution1 only)	60	1600	800	800
		Validation (Institution1 only)	10	262	131	131
		Testing (Institution1 only)	30	796	398	398

OVBM: Off-by-one Vertebral Body Misalignment.

2.2.3 Image Pre-processing

The REG files, both aligned and misaligned, were consequently used to match the coordinates of the CBCT volume with those of the planning CT volume. To ensure uniformity over the whole dataset, all volume pairs were resampled using a 1x1x1.5 mm^3 grid.

The couch position from the planning CT is very rarely aligned to the couch from the CBCT due to differences in material and structure. Hence, the positional and structure differences in the images are of trivial importance in our error detection system and can even be misleading in the detection of wrongly aligned patients. In order to remove the couch from the images, the body contours found in the structure files were used to clean-up both the CT and CBCT volumes such that the couch and other irrelevant regions outside of the body were assigned voxel values equivalent to the Hounsfield unit (HU) of air (-1000 HU).

The eventual goal of this project is to develop a tool that could run in real-time simultaneously with treatment delivery processes, and hence, run time and memory footprint were primary considerations. Therefore, instead of using the entire 3D image volumes as inputs to the deep learning model, orthogonal 2D slices were used, as shown in Figure 2.1. By extracting one slice in each orthogonal plane, the memory requirement of our model was considerably minimized, while the important features of the patients' anatomy were kept and used by our model to analyze the patient alignment. Selection of an appropriate origin for the coordinate axes was important to assure that the relevant image features were present, as described in sections 2.2.4-2.2.5.

In the clinic, for thoracic and abdominal cases, the spine is often used as a marker during the registration and patient alignment step. An automated process was therefore used to select axial, sagittal, and coronal planes that intersected at the approximate center of the vertebral bodies at a location midway through the image in the cranio-caudal direction. This particular point was chosen, rather than the treatment plan isocenter, as it offers more details about the vertebral location in the detection of off-by-one vertebral body misalignments.

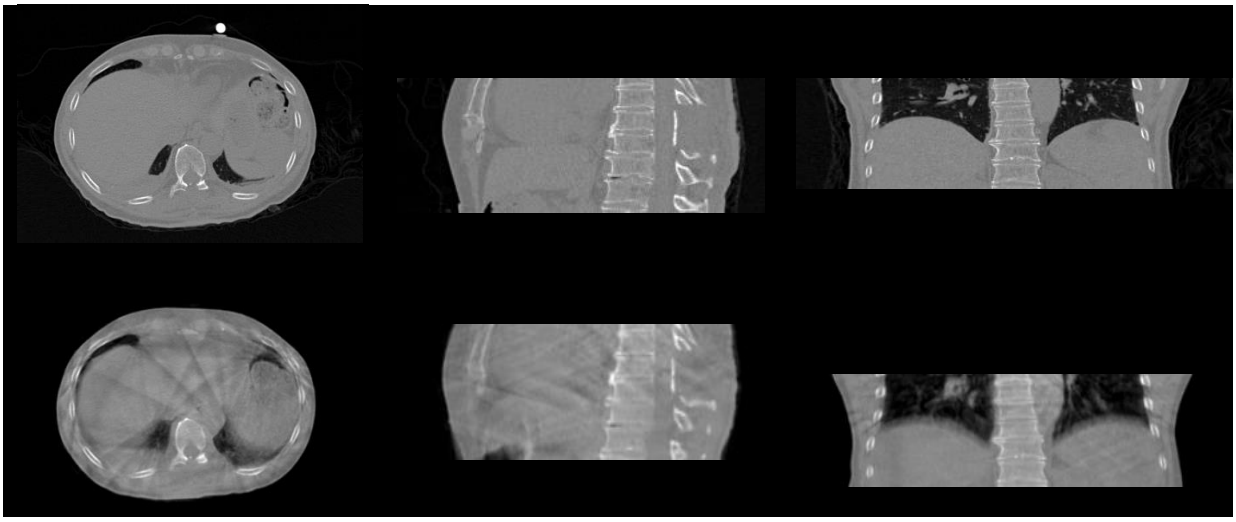
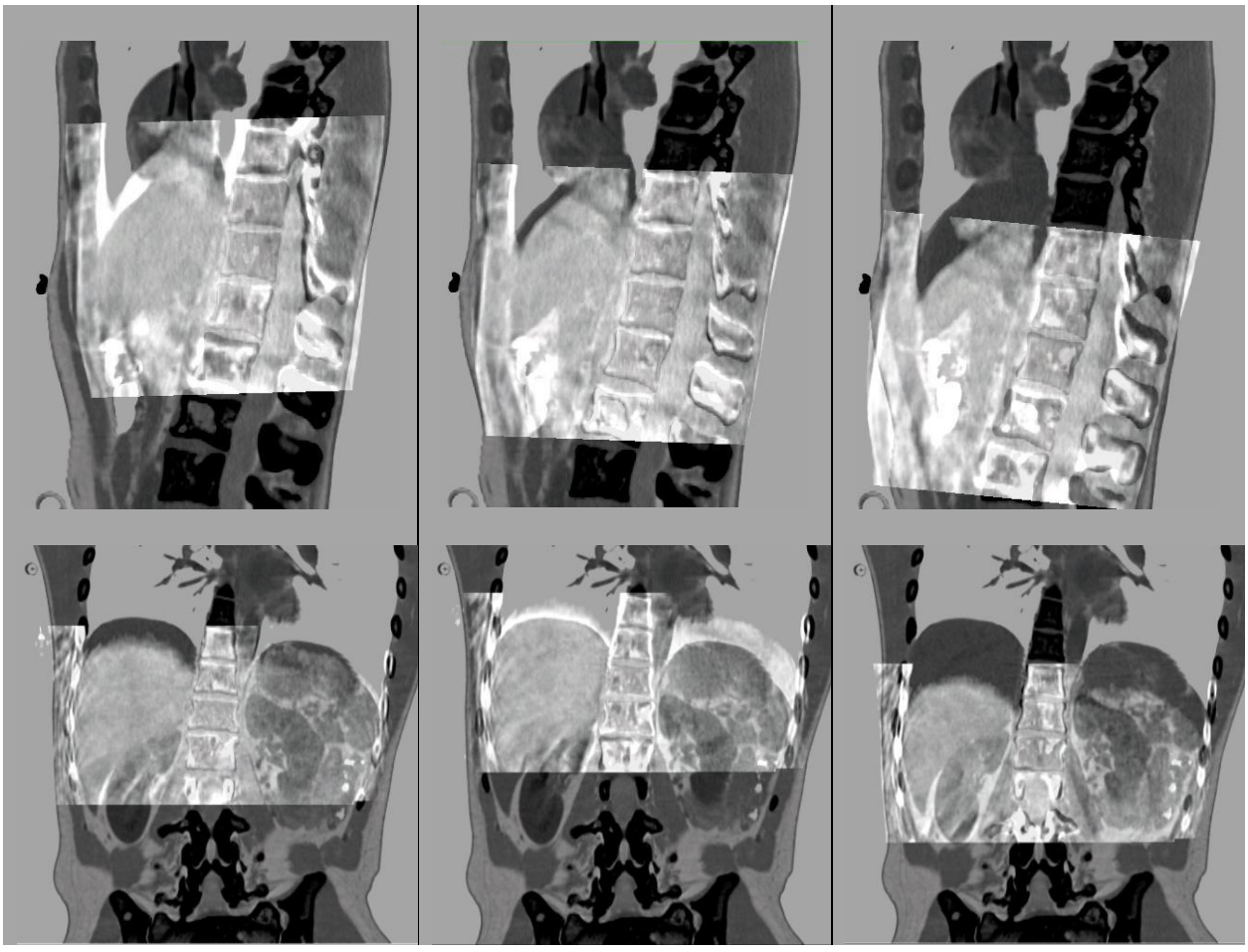


Figure 2.1: Orthogonal 2D slices extracted from the planning CT (top row) and its corresponding CBCT (bottom row).

2.2.4 The Thoracic-Abdominal (TA) Model

Algorithm training and test data consisted of planning CTs and setup CBCTs from the 480 Thoracic-Abdominal datasets. The clinically applied registration was used to derive true-negative (no error) data. The setup and planning images were then be misaligned by one vertebral body in both the superior and inferior directions, simulating the most likely

misalignment scenarios, as shown on Figure 2.2. In the misalignment generation process, the planning CT-CBCT pair from the earliest treatment fraction of each patient, together with its corresponding clinically applied REG file, were selected and imported into MIM (MIM Software Inc, Ohio, United States). For the 480 selected pairs, off-by-one vertebral body misalignments were simulated on MIM by manually shifting the CBCT by one vertebral body in the cephalad-caudad direction with respect to the planning CT.



(a) (b) (c)

Figure 2.2: Image fusions to demonstrate the manually-generated off-by-one vertebral body misalignments. In column (a), the CBCT was up-shifted by one vertebral body with respect to the

planning CT. Column (b) shows the correct clinical alignment, and column (c) shows a misalignment where the CBCT was down-shifted by one vertebral body with respect to the planning CT.

For thoracic and abdominal cases, the spine is often used as a marker during the registration and patient alignment step. An automated process was therefore used to select axial, sagittal, and coronal planes that intersected at the approximate center of the vertebral bodies at a location midway through the image in the cranio-caudal direction. This particular point was chosen, rather than the treatment plan isocenter, as it offers more details about the vertebral location in the detection of off-by-one vertebral body misalignments.

A binary mask of the patient body was first extracted from the CBCT using a thresholding method. A morphological dilation followed by an erosion were applied on the binary mask to fill any gaps after the thresholding operation. The dilation and erosion operations used 20 x 20 pixel^2 and 5 x 5 pixel^2 rectangular structuring element, respectively. The axial slice index was then extracted by locating the middle slice of the mask containing the patient body on the CBCT, denoted as X_{Ax}.

Using the axial slice index obtained in the previous step, the corresponding axial slice images were extracted from both the CT scan and the CBCT scan. The vertebral body location in the coronal and sagittal planes, denoted as (X_{Cor}, X_{Sag}) was then obtained by applying a constant 10-pixels translation in the anterior direction from the central point of the spinal canal. The spinal canal location was derived from the spinal canal structure in the treatment plan if existing, or from a dedicated UNet-based [23] spinal canal segmentation algorithm (see Appendix A.1) if the plan did not contain a spinal canal contour. This vertebral body's coordinates (X_{Cor}, X_{Sag},

X_{Ax}) were then used as the coordinate origin for the coronal and sagittal 2D slices extracted from the planning CT scan and the CBCT scan.

Following the orthogonal slice extraction, the 2D images in the coronal and sagittal planes were cropped to reduce the empty regions around the patient body, and hence minimize the number of unnecessary computations in our error detection model. The coronal and sagittal slices were cropped to 400 x 110 and 280 x 110 images about the center of the CBCT image, which was found using the binary mask of the patient body. The axial slice images were down-sampled using a linear interpolation method to obtain 256 x 256 arrays. By repeating our experiment on the original 512 x 512 axial images, we found that the downsizing step performed did not have any adverse effect on the accuracy of our error detection model.

After the orthogonal slice extraction, the 2D arrays from the planning CT and CBCT were then concatenated with respect to their plane to obtain one 256 x 256 x 2 axial array, one 400 x 110 x 2 coronal array, and one 280 x 110 x 2 sagittal array. For each of the orthogonal arrays, the first channel is the planning CT image, and the second channel is the respective CBCT image.

The three slice pairs obtained were then be inputted to the TA model which returned a probability of vertebral misalignment. The TA model, based on the Dense-Net architecture [41], consisted of three branches which processed the three orthogonal images separately before merging into a final densely connected layer, as shown in Figure 2.3. In the clinical setting, the manual image registration process is often performed using all three orthogonal views. However, coronal and sagittal planes may be more sensitive to cranio-caudal mis-registrations such as one vertebral body displacements. Therefore, more convolutional filters were placed in the coronal

and sagittal branches such that the model extracts a higher number of features from these two planes, as compared to the axial branch. Hence, this resulted in the model placing higher weights on the coronal and sagittal plane during the off-by-one vertebral body misalignment detection.

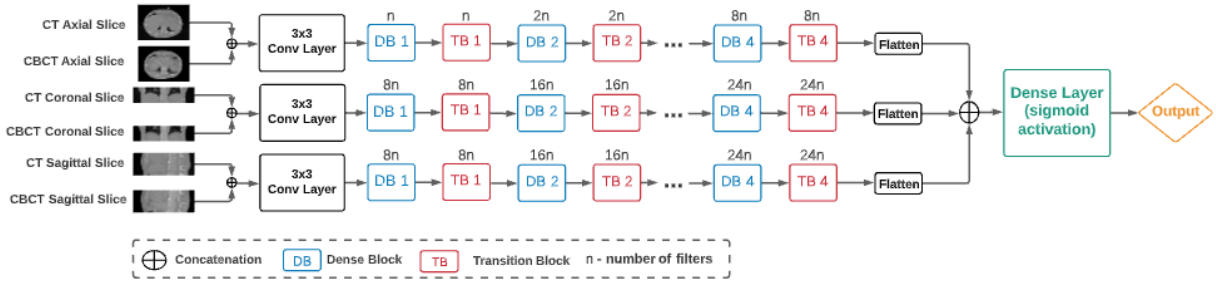


Figure 2.3: Depiction of the network architecture for the TA model (n = 4). The Dense Block consists of two densely connected layers connected in a feed-forward mode (each composed of two convolutional layers, two batch normalization layers, two activation layers, and one dropout layer) and the Transition Block of three layers (batch normalization layer, convolutional layer, and max pooling layer).

2.2.5 The Head & Neck (HN) Model

The 1,342 HN CBCT-CT pairs from 100 patients were used for the HN model training and testing. Similarly to the TA model, the clinically applied registration was used as the true-negative data. For each treatment scan, a 10 mm misalignment in a randomly chosen direction was automatically generated to produce the true-positive (misaligned) data. For each of the aligned and misaligned 3D image pairs, 2D slice pairs were automatically extracted in each anatomical plane about a point within the cervical vertebrae (neck treatments) or at the CBCT isocenter (head scans) and inputted to the HN model.

The HN model had a similar architecture to the TA architecture. However, equal number of convolutional filters were placed in each branch of the network as the misalignment can occur in all three orthogonal planes and the model may benefit from the features extracted from all three branches equally. However, 10mm misalignments for HN treatments is quite large as compared to the standard 3-5 mm planning target volume (PTV) margin used in these cases. [42] Hence, 5 mm misalignments in random directions were also generated to evaluate the model performance.

2.2.6 The Pelvis (PL) Model

The 2,658 pelvis datasets from 100 patients were used for the PL model training and testing. Similarly to the HN model experiment, the clinically applied registration was used as the true-negative data and a 10 mm misalignment in a randomly chosen direction was automatically generated to produce the true-positive data. 2D slice pairs were extracted in each anatomical plane about the treatment isocenter. The registration performed for pelvic cases is highly reliant on the tumor position and fiducial markers around it. As these features may help in the detection of misalignments, the treatment isocenter was chosen as the point where the 2D slice pairs were extracted from. The PL model's architecture used the same one as the HN model's as the pelvic misalignments may be in any direction.

2.2.7 Model Training Configurations

In our experiments, we refrained from using pre-trained classification networks which are trained on natural images, such as ResNet50 [43], and fine-tune the model using our dataset. As natural image classification tasks are essentially very different from medical image classification

tasks in terms of image characteristics, dataset sizes and number of classes, transfer learning using powerful pre-trained network has shown to offer little benefit in the medical imaging domain as compared to training the network from scratch using medical images. [44]

The proposed patient setup error detection models were implemented using Tensorflow 2.2 with Keras backend. The models were trained using Adam Optimizer [45] with a starting learning rate of $5x10^{-5}$. During training, the models were evaluated after each epoch using their respective validation set, and the learning rate was reduced by a factor of 0.75 if the validation loss did not improve for 15 consecutive epochs. All models were trained until the validation AUC did not improve for 50 consecutive epochs, or for a maximum of 200 epochs.

2.2.8 Loss Function and Evaluation Metrics

During the model training, the binary cross-entropy (BCE) loss was used as the loss function, as shown in Equation 1. BCE has been shown to be an effective loss function for binary classification problems. [46]

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} -(y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i)) \quad [1]$$

Where y is the ground truth label, and p is the predicted probability of misalignment.

The receiver operating characteristic (ROC) curve was used to assess the performance of our models in classifying the registrations from our validation dataset. [47] The areas under each ROC curve (AUC) were used to quantify the performance of our models.

While the principal target of our proposed algorithm is to catch misalignment errors, due to the rarity of the event in the clinic, it is crucial to minimize unnecessary disruption in the clinical workflow due to false positives. Based on our analysis on the patient load at institution #1, which approximates to 300-350 treatments per week, it was deduced that a specificity of ≥ 99% would be equivalent to about one false positive per treatment machine per week, which was deemed acceptable in comparison to other false-positive interrupts and interlocks in the clinical workflow. While this false-positive rate may vary from institution to institution based on the patient load, we believe that the chosen specificity is reasonable enough to limit clinical disruptions at most facilities. Hence, in our evaluation, a base threshold value yielding a specificity of at least 99% was chosen. From the binary results obtained, the true positive (tp), false positive (fp), false negative (fn), and true negative (tn) counts were obtained. These were then used to calculate the sensitivity, F-1 score, and Matthews correlation coefficient (MCC).

The F-1 score combines both the precision and recall of a binary classifier and is shown in Equation 2 below. [48]

$$F\text{-}1 = \frac{2tp}{2tp + fp + fn} \quad [2]$$

MCC, shown in Equation 3, is another metric used to quantify the performance of a binary classifier, and has been shown to be a balanced measure in the case of class imbalances. [49-50] MCC can take a value between -1 and +1, where +1 means perfect positive correlation between prediction and ground truth, and -1 means perfect negative correlation.

$$MCC = \frac{(tp * tn) - (fn * fp)}{\sqrt{(tp + fn) * (tn + fp) * (tp + fp) * (tn + fn)}} \quad [3]$$

Additionally, for the TA model only, the mean model prediction probability was calculated for varying caudal-cranial misalignment magnitudes between the planning CT and CBCT. For each image pair in the test set, the CBCT was automatically misaligned in the caudal-cranial direction by ±10 mm, ±20 mm, and ±40 mm with respect to the planning CT. These image pairs were then inputted to the best performing model to obtain the misalignment prediction probabilities. Provided that the human thoracic vertebral body is on average 20 mm in length [51], this test can add value to the clinical utility of our algorithm by validating its potential at catching misalignment errors which are off by less than one vertebral body, and also misalignment errors which are greater than one vertebral body in magnitude.

2.3 Results and Analysis

After evaluating the models on their respective test dataset containing the simulated errors, target thresholds (tr_{90} and tr_{99}) were found using an ROC analysis. For each target threshold, the sensitivity and specificity were obtained and reported in Table 2.2. Additionally, the F-1 score and MCC were calculated to evaluate each model using the tr_{99} threshold. Those results are reported in Table 2.3.

Table 2.2: Description of the performance and target thresholds of each model using a receiver operating characteristic (ROC) analysis.

Model	Error Type	AUC	tr_{90}	Specificity @ tr_{90}	Sensitivity @ tr_{90}	tr_{99}	Specificity @ tr_{99}	Sensitivity @ tr_{99}
Thoracic-Abdominal	OVBM	99.4%	0.812	98.8%	95.0%	0.001	84.6%	99.2%
Head & Neck	10 mm shift	99.6%	0.990	98.9%	90.0%	0.050	98.9%	100.0%
Pelvis	10 mm shift	99.2%	0.920	95.0%	97.0%	0.360	90.5%	100.0%

AUC: Area under the ROC curve; OVBM: off-by-one vertebral body misalignment; tr_{90}: threshold resulting in sensitivity ≥90%; tr_{99}: threshold resulting in sensitivity ≥99%

Table 2.3: Classification results of the three models on their respective test dataset(s) using a threshold that yields at least 99% specificity.

Model	Error type	F-1 Score	MCC
Thoracic-Abdominal	OVBM	0.97	0.95
Head & Neck	10 mm shift	0.99	0.98
	5 mm shift	0.85	0.76
Pelvis	10 mm shift	0.94	0.88

OVBM: off-by-one vertebral body misalignment; MCC: Matthews correlation coefficient.

Additionally, Figure 2.4 below shows the mean model prediction probability obtained from the TA model as a function of caudal-cranial distances between the planning CT and CBCT.

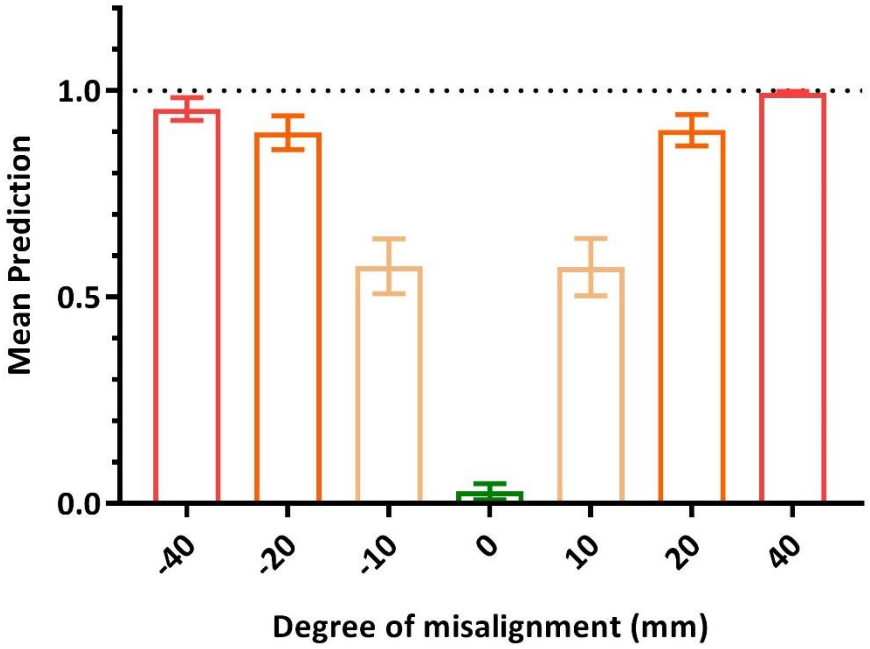


Figure 2.4: Column bars to represent the mean misalignment prediction of the TA model on its test dataset as a function of caudal-cranial misalignment distances. The error bars represent the 95% confidence interval of the mean value.

2.4 Discussion

In this work, a deep-learning based patient setup misalignment error detection algorithm (EDA) for cone beam CT-guided radiotherapy was presented. All three models within EDA were shown to achieve high accuracy in detecting their respective error mode, with areas under the receiver characteristic curve of 99.6%, 99.7%, and 99.2% for the HN, TA, and PL models respectively.

Using an Nvidia Quadro P1000 4GB Graphics Processing Unit (GPU) (Nvidia Corporation, Santa Clara, CA, USA) system with 16 GB RAM, our algorithm takes an average of 6.8 seconds to pre-process the input images, run through the EDA and output a probability of misalignment. If the system were implemented as a third-party system independent of the clinical

record and verify (R&V) system, the images would have to be retrieved from the treatment machine or R&V system, thereby increasing the runtime. In our implementation in a Varian environment, retrieval of CBCT and alignment (REG file) from the ARIA servers using our DQR software required an additional 58 seconds, on average. Ideally, the proposed algorithm would be incorporated into the R&V system, obviating this data transfer contribution to the runtime. While run-time optimization could further decrease the execution time of the algorithm, we believe that it can be clinically implemented as-is.

As compared to a similar work that uses non-deep learning (non-DL) techniques [25] to find patient setup errors in CBCT-guided radiotherapy treatments, the EDA model resulted in higher sensitivities (0.99 vs 0.73 for HN, 0.96 vs 0.90 for TA, and 0.89 vs 0.74 for PL) for a fixed specificity of 99% for all regions. Additionally, our model was validated on a larger test set composed of unseen data from two different institutions as compared to the non-DL techniques that were validated using a 10-fold cross-validation method for a training-testing dataset composed of 240 patients from a single institution.

The TA model also obtained significantly higher mean prediction scores for 10mm, 20mm, and 40mm caudal-cranial misalignments as compared to the correct clinical alignments. This demonstrates the potential of the TA model in detecting misalignments smaller and larger in magnitude than 1 vertebral body, in addition to the off-by-one vertebral body misalignments, which validates the appreciable value that the model can add to the clinical workflow and to the patients' safety.

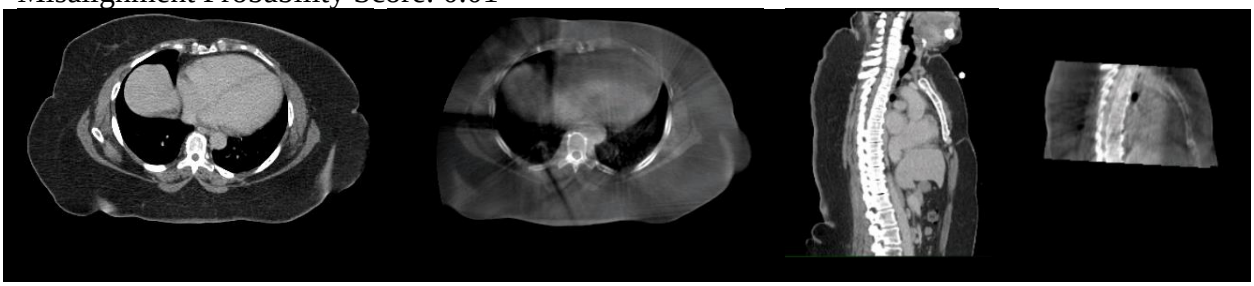
Upon analysis of the false-positive cases, EDA was seen to be sensitive to cases where the CBCT clinical setup instructions indicated the prioritization of the soft tissue alignment over

the bony alignment. For those cases, misalignments may be present at the bony anatomy, as shown in Figure 2.5 (Case 1), which we believe likely triggered the mis-classification. Additionally, considerable streak artifacts were often observed on the CBCT image of the false-positive cases, which may have affected the model output. Of the seven false-negative cases in the TA region, four had a limited field of view, where part of the patient anatomy was not captured on the CBCT as shown in Figure 2.5 (Case 2). The other three cases showed considerable streak artifacts on the CBCT (see Figure 2.5, Case 2 and 3), which could be due to beam hardening effects, photon starvation, or exponential edge gradient effects. [52] The image properties discussed above may have contributed to the wrong classification of those few cases; however, further tests on a larger and more diverse dataset are required to verify the exact causes of failure. Future work could include a dedicated model that flags lower quality scans such that the results of EDA can be interpreted accordingly. Alternatively, attention gates [53] could be incorporated in EDA such that the model focusses on targeted regions instead of irrelevant regions which may contain artifacts.

However, patient data from two institutions may not be enough to capture the variability in scanning protocol, image quality, registration techniques and error modes across all treatment facilities and treatment machines. Hence, this work calls for the importance and need for more data across multiple facilities, such that the generalizability power of the EDA could be improved, and the system could benefit a wider range of facilities. Further work in this direction should include a determination of a minimum diversity of cross-institutional data that would lead to an expectation of similar model performance on data from an unseen institution.

Case 1: False Positive
Misalignment Probability Score: 0.99

Case 2: False Negative
Misalignment Probability Score: 0.01



Case 3: False Negative
Misalignment Probability Score: 0.42

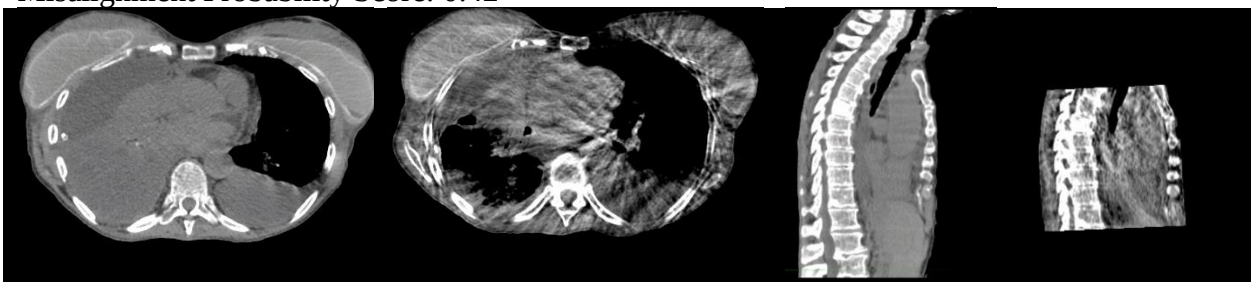


Figure 2.5: Three examples of mis-classification by the TA model using a ≥99% specificity threshold. For each case, the planning CT slice is shown to the left of the corresponding CBCT slice. Case 1 shows a correct clinically performed registration where the soft tissue alignment was prioritized over the bony alignment (the contours of the planning target volumes are shown to demonstrate the misalignment present at the vertebral body). Case 2 shows an example where part of the patient body was not present on the CBCT axial scan, in addition to considerable streak artifacts. In Case 3, substantial streak artifacts were observed on the CBCT scan.

Another limitation of this study is the use of 2D orthogonal slices as input to the EDA, instead of the whole 3D volumes. While the 2D images lead to faster computation time, the number of features captured by the model is limited to the selected slices. A 3D model could capture many more useful features from the entire scans, which could further improve the detection of misalignment errors. With the current systems available in the clinic, the 3D model is currently deemed impractical due to its memory requirements. However, with the rise in computation technologies and easier access to high-end GPUs, the 3D EDA could be a more effective and practical approach in the future, as compared to the 2D EDA.

The EDA algorithm also focuses on specific types of translational error that could occur in CBCT-guided radiotherapy. Other subtle errors occurring at the soft tissue level during the registration of the CBCT to the planning CT can possibly lead to sub-optimal treatments and must be avoided. The algorithm is not currently optimized to catch these soft-tissue misalignments.

Even though patient setup misalignment errors occur very rarely in the clinic, they can have serious consequences to the patient if not detected prior to treatment. With its strong error-catching ability, the EDA could prove to be useful as a fully-automated online secondary safety check to the therapist, minimizing the risk of wrong registration during patient alignment. It can be deemed even more useful in facilities that have a shortage of radiation therapy technologists, which is often seen in underserved communities and developing countries. [54-56] As compared to real-time monitoring systems using surface imaging [21-23], the software requires minimal external hardware (standalone computer plus interface hardware and software) and a low up-front cost for clinical implementation. Hence, the software can be of particular interest to facilities that lack resources for additional equipment for patient safety.

Additionally, the EDA could be used as an aid or supplement to image review performed by medical physicists as part of weekly chart check quality assurance (QA) of external beam radiotherapy treatments. Although physicians are responsible for approving image guidance results, medical physicists commonly spot-check image alignments on a daily or weekly basis, which has a high-risk priority number in the radiation therapy workflow [19] and is time-consuming [57]. The tool offers the possibility of automatically analyzing all of the scans of patients being treated within a particular time frame, and flagging all of the possible anomalies detected through a time-stamped report. This way, the physicist can effectively review the handful of treatments which have been flagged as highest probability of treatment error, instead of randomly choosing plans or going through all the treatment plans. Hence, our tool can not only make treatment QA more time-effective, but it can also make it more robust to incident detection and improve incident learning.

2.5 Conclusion

Patient setup misalignment represents a rare but serious error in image-guided radiotherapy. An automatic deep-learning based misalignment error detection algorithm was proposed which can flag potential cases of patient setup misalignment in the registration of the planning CT to the CBCT. The results have shown that our algorithm has sufficient sensitivity and specificity for routine clinical use. Algorithm robustness was validated by applying it to inter-institutional data. This algorithm can be used as an online safety-check during CBCT-based guided radiotherapy and can also facilitate the quality assurance of external beam radiotherapy treatments by aiding medical physicists during regular physics chart checks.

Chapter 3: Feasibility of a deep-learning based anatomical region labeling tool for Cone-Beam Computed Tomography scans in Radiotherapy.

3.1 Introduction

Cone-beam computed tomography (CBCT) is commonly used for radiotherapy image guidance because it facilitates accurate and precise positioning and alignment of the patient. In real-time adaptive radiotherapy, the CBCT may also be used to adapt the treatment plan based on the new target location and size, and the position of organs at risk (OARs). In this case, the delineation of the target(s) and OARs may be required on the CBCT scan prior to the plan adjustment [58]. With the rise of machine learning and deep learning techniques in the field of medical image analysis, many algorithms are being developed to automate and expedite this delineation process [59-62]. However, these algorithms are typically anatomical region-specific and assume the presence of the organs-of-interest irrespective of the body region input to the algorithm.

The recognition of the global body region may be useful as a pre-processing step for these tools, such that they are applied to body regions within their domain. However, this step is often neglected due to the assumption that the anatomy information is present on the Digital Imaging and Communications in Medicine (DICOM) headers. While a ‘Body Part Examined’ tag is indeed present in the DICOM headers of the planning Computed Tomography (CT), it has been shown that this information is not very reliable, with a mis-labeling rate of 15.3% [63]. Furthermore, these pre-defined labels are driven by the acquisition protocol. Due to the variability and differences among the patients' anatomies, an imaging protocol for a different body region may be used by the clinical personnel in order to obtain better image quality. While

the header can be adjusted following the CT acquisition, this is not commonly done in the clinic, which may lead to a wrong body region label [64]. Additionally, this 'Body Part Examined' tag may be completely absent in the CBCT DICOM headers, as is the case at our institution, highlighting the need for an automatic region-labeling algorithm to recognize the global patient anatomy and treatment region.

Several algorithms have recently been proposed for the classification of anatomical regions in CT and MRI scans [65-68]. Among those, Ouyang et al. [68] achieved the highest classification accuracy of 97.3% on their test dataset composed of 663 CT scans. These previous studies showcase the potential of deep learning techniques on such region labeling problem. Nevertheless, if these techniques are used as a pre-processing step for other clinical tools, which have their intrinsic error rate, it is imperative to minimize the pre-processing error rate as much as possible to improve the reliability of the labeling tool and reduce the overall algorithm's failure rate. Hence, it is vital to continuously identify and address limitations of such region labeling tools.

One common characteristic and limitation of the previous studies is that they have all been developed and tested on CT and MR images, which typically have improved image quality as compared to pre-treatment CBCT images [69-70]. Hence, classifying CBCT images may become a challenge as fewer useful features and more artifacts may be present on the CBCT scan for accurate region labeling. CBCT scans may also have a small field-of-view (FOV), which is usually restricted to the treatment region only, making a consecutive body part recognition algorithm as in [68] more complicated.

To address this current limitation, we propose a CNN-based anatomical region labeling (ARL) tool which can classify a CBCT scan into four global regions, namely head & neck (HN), thoracic-abdominal (TA), pelvis (PL) and extremity (EX) using a single coronal slice from the CBCT volume. This tool will subsequently be plugged into the EDA, such that scans of patients undergoing CBCT-guided radiotherapy can be automatically directed to their respective error detection model based on the anatomy present on the scan.

3.2 Materials and Methods

3.2.1 Dataset Acquisition

Under an IRB approved protocol (UID 18-001430), CBCTs were collected from 631 patients undergoing radiotherapy treatment at the University of California, Los Angeles Medical Center (UCLA) between January 2017 and April 2022. The dataset collection was performed using an in-house DICOM query and retrieval (DQR) application programming interface using the pynetdicom Python package. The treatments at UCLA had been performed on three TrueBeams and one NovalisTx linear accelerator treatment machines (Varian Medical Systems, California, United States). CBCT scans were acquired using the on-board imager of each machine. For each CBCT, the corresponding planning CT, REG file and RTStruct file were also collected and used during the image pre-processing step in our implementation.

A visual inspection of the treatment isocenter was performed to sort the CBCT scans into four different global regions: head & neck (HN), thoracic-abdominal (TA), pelvis (PL), and extremity (EX). The C7 vertebral body was used as a limit to the HN region such that the CBCT

scan only contained these two body parts, as shown in Figure 3.1. However, in the clinical setting, it is possible to have neck scans containing part of the thorax. For the first part of our experiment, which included model training and testing, these scans with substantial overlapping regions were withdrawn from our dataset to maintain the distinction between each category. Following the triage, 3802 CBCT scans from 596 patients remained, as described in Table 3.1. The limits of the TA region were the T1 vertebra and the L2 vertebra, avoiding the neck and pelvis regions. For the PL scans, the L3 and S2 vertebra were used as markers, avoiding the abdominal region and area below the pubic symphysis. Scans of the arms, legs and extremity of the shoulder were placed in the EX dataset.

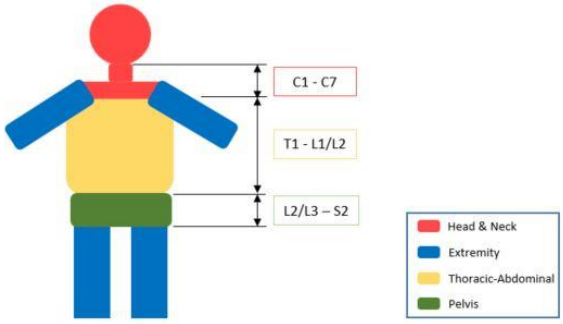


Figure 3.1: Depiction of the strategy used to sort the CBCT scans based on the position of the treatment isocenter and through visual inspection of the scan. For the neck, thoracic, abdominal and pelvis regions, the vertebral bodies were used as reference, as shown above.

Table 3.1: Description of the dataset partitioned into the training, validation, and testing sets for each global region.

		Number of patients	Number of CBCT scans	Number of coronal slices*
Head & Neck	Training	92	674	2,022
	Validation	15	62	186
	Testing	46	321	321
Thoracic-Abdominal	Training	148	483	1,449
	Validation	24	85	255
	Testing	72	310	310
Pelvis	Training	94	1,095	3,285
	Validation	16	173	519
	Testing	45	359	359
Extremity	Training	27	118	354
	Validation	5	22	66
	Testing	12	100	100
Total	Training	361	2,370	7,110
	Validation	60	342	1,026
	Testing	175	1,090	1,090

**For each scan in the training and validation sets, three coronal slices were extracted and used as an augmentation method during model training. During model testing, only one slice was used per scan.*

Each of the four datasets was then separately and randomly split into a training, validation and test set using a 60:10:30 ratio. As scans from multiple treatment fractions were used in our study, the dataset split was performed based on the patients' unique anonymized identifiers to avoid having scans from the same patient overlapping across the training, validation, and test sets.

3.2.2 Image Pre-processing

The pixel spacing of the CBCT scans ranged from 0.51-1.17 mm and the slice thickness from 1-2.5 mm. The CBCT scans were resampled based on their corresponding planning CT to produce uniform images with a voxel spacing of 1x1x1.5 mm^3. In our pipeline, this resampling,

and volume matching was performed using the REG file present with the CT-CBCT pair. However, the CBCT resampling can be made independently from the planning CT and REG file for another application of the ARL. Furthermore, the treatment couch and immobilization devices were removed from the CBCT image using the body contour present in the RTStruct file. In the event that a body contour is not present in the RTStruct file, a thresholding method, including a morphological dilation followed by erosion, was used to extract the body contour from the CBCT. The dilation and erosion operations used 20×20 and 5×5-pixel^2 rectangular structuring element, respectively.

A coronal slice was then extracted from each CBCT present in our dataset and each image was labeled using their corresponding global region. The primary coronal slice was extracted by locating the CBCT slice with the highest mean Hounsfield Unit (HU). This slice-selection method was chosen such that the coronal slice would cover the whole extent of the patient scan while containing considerable bony structures (higher HU) which can be useful features in the recognition of the anatomical region.

For training purposes, two additional slices were extracted from the CBCT scans in the training and validation datasets, with each slice being 10 pixels away from the primary coronal slice location; one being 10 pixels in the anterior direction and the other being 10 pixels in the posterior direction. The extraction of these two extra slices were performed as an augmentation method during the model training due to the inter-patient variability in anatomy which can be present on the primary coronal slice. The slices were then cropped about the center of the patient body to reduce empty spaces around the body and obtain 150x400 pixel^2 images, as shown in Figure 3.2, and used as input to our ARL model.

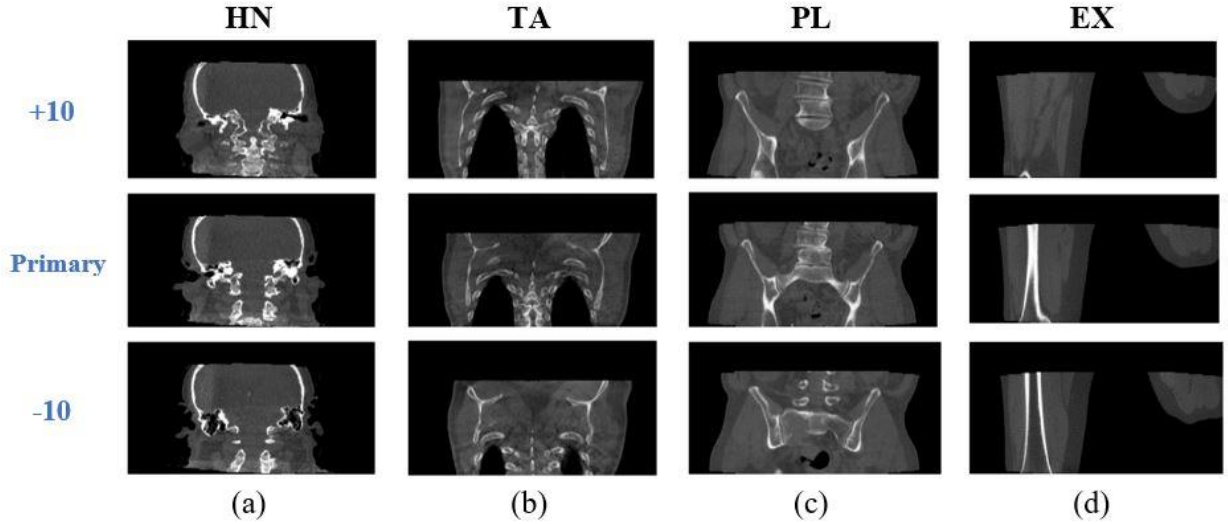


Figure 3.2: Twelve coronal slices which were used as input to the ARL model during algorithm training. Each column (a-d) shows the three slices extracted from four different CBCT scans, one from each anatomical region. The first row represents the slices extracted 10 pixels away from the primary coronal slice location in the anterior direction. The second row show the slices which are extracted at the primary coronal slice location, and the third row represents the slices extracted 10 pixels away from the primary coronal slice location in the posterior direction. HN: Head & Neck, TA: Thoracic-abdominal, PL: Pelvis, EX: Extremity.

3.2.3 Anatomical Region Labeling (ARL) Model

The ARL model used the Dense-Net architecture as shown in Figure 3.3. The ARL model makes use of densely contracting paths to capture contextual information from the CBCT coronal image before outputting a probability of occurrence for each of the four classes. The Dense Block in our architecture constitutes of two densely connected layers, each comprising of seven layers. The two densely connected layers in the Dense Block were connected to each other in a feed-forward mode to maximize feature reuse, which been shown to be computationally efficient, hence allowing a deeper network [41].

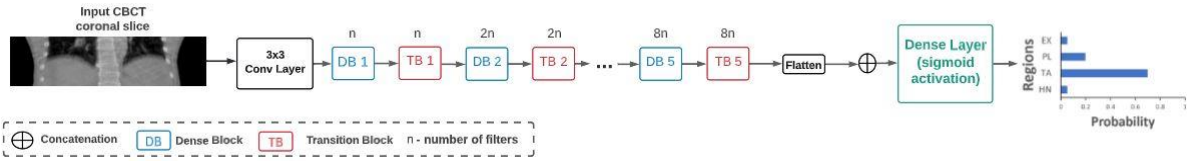


Figure 3.3: (a) Depiction of the network architecture used in the proposed Anatomical Region Labeling (ARL) model (n = 4). The Dense Block consists of 2 densely connected layers (each composed of 2 convolutional layers, 2 batch normalization layers, 2 activation layers, and 1 dropout layer) and the Transition Block of 3 layers (batch normalization layer, convolutional layer, and max pooling layer).

3.2.4 Model Training Configuration

The ARL model was implemented using Tensorflow 2.2 with Keras backend. Model training was performed using the Adam Optimizer [45], with a starting learning rate of $2x10^{-5}$. During training, the model was evaluated on the validation dataset after each epoch, and a learning rate reducer (0.75) was applied if the validation loss did not decrease for 15 consecutive epochs. To avoid overfitting the model on the training dataset, an early stopping method was applied such that training would stop if the validation accuracy did not improve for 50 consecutive epochs, or for a maximum of 400 training epochs. For comparison purposes, we also trained a Support Vector Machine (SVM) [71] and saved the parameters which produced the highest accuracy on the validation set.

3.2.5 Evaluation Metrics and Quality Control

After the trained ARL model and SVM was applied to the test dataset, the true positive (tp), false positive (fp), false negative (fn), and true negative (tn) counts were obtained for each anatomical region. Subsequently, the four metrics shown in Equations 4-7 were used to evaluate and compare the performance of our models.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad [4]$$

$$F\text{-}1\ Score = \frac{2tp}{2tp + fp + fn} \quad [5]$$

$$Precision = \frac{tp}{tp + fp} \quad [6]$$

$$Recall = \frac{tp}{tp + fn} \quad [7]$$

To obtain visual explanations of the model's prediction, the Gradient-weighted Class Activation Mapping (Grad-CAM) [72] was implemented. This Grad-CAM method uses the gradients from the final convolutional layer from the ARL model to produce a heat map

describing the regions which contributed the most to the activation of the predicted anatomical region.

3.2.6 Clinical Implementation and Validation

Using our in-house DQR system to interface with the ARIA system (Varian Medical Systems, Palo Alto, CA), the ARL was implemented at our clinic to automatically classify incoming CBCT data on a daily basis for 22 consecutive days between August and September 2022 as part of a pilot process for an automated weekly chart check image analysis described in **Chapter 4**. For validation purposes, the predictions for the first 100 unique patients were compared to a human perspective. Without any information about the predictions of the ARL, each of the 100 unique scans was visually analyzed and labeled by a human observer to obtain the ground truth label.

However, in contrast to the dataset used during model training and testing this validation dataset did not exclude scans containing overlapping regions, such as neck and thorax, or abdomen and pelvis. Hence, the ground truth labels were obtained by identifying the dominant region (i.e. the region encompassing the majority of the CBCT scan). Furthermore, the other less pronounced region(s), if present, was noted as a 'less-pronounced region'. For example, for a neck treatment scan containing mostly the neck and part of the thorax, the region would be labeled as HN, with the 'less-pronounced region(s)' being TA. Following the human annotations,

the predictions from the ARL were compared with their respective ground truth labels and the model performance was evaluated.

3.3 Results and Analysis

3.3.1 Model Training and Evaluation

During the algorithm training, the ARL model achieved convergence after 49 epochs with training and validation accuracies of 99.8% and 99.3%, respectively. Following the testing phase, the ARL model resulted in 9 misclassifications out of the 1,090 test cases, for an overall accuracy of 99.2%. Selected true-positives and misclassifications are shown in Figure 3.4 and Figure 3.5, respectively.

For the SVM, a polynomial kernel was found to produce the best fit, with training and validation accuracies of 96.0%. Following testing, the SVM obtained an overall accuracy of 91.5%. Using Student paired t-tests, results from the ARL model and SVM were found to be statistically significant (p-value < 0.0001). The detailed results obtained from the ARL model and SVM are reported in Table 3.2.

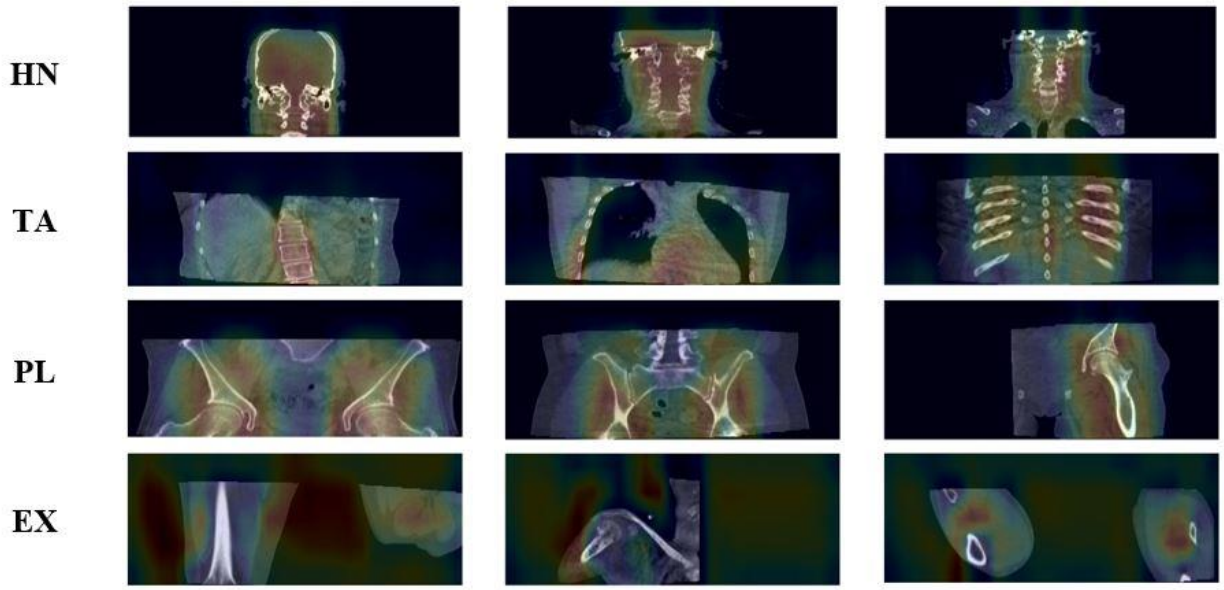
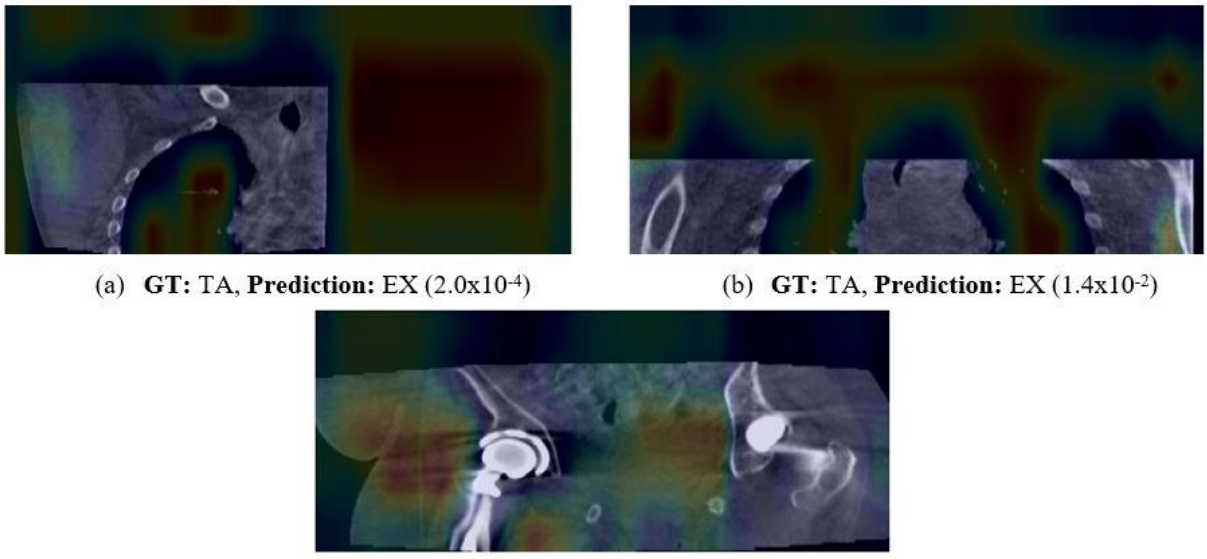


Figure 3.4: 12 selected CBCT slices (from unique patients) which were inputted to the ARL model and resulted in true-positives. The Grad-CAM activation heat map is overlaid on the CBCT image to display the regions which had the greatest weight in the prediction. The red areas mean the region contributed more to the prediction. HN: Head & Neck, TA: Thoracic-abdominal, PL: Pelvis, EX: Extremity.

Table 3.2: Performance of the Anatomical Region Labeling (ARL) model and the Support Vector Machine (SVM) on the 1,090 test cases. The results are shown for each three global regions separately. Bold texts represent the better result between the two models.

	HN		**TA**		**PL**		**EX**	
	ARL	SVM	ARL	SVM	ARL	SVM	ARL	SVM
Accuracy	**99.9%**	97.9%	**99.4%**	92.8%	**99.6%**	95.3%	**99.4%**	97.0%
F-1 Score	**99.8%**	96.4%	**98.9%**	86.2%	**99.4%**	93.1%	**97.1%**	85.3%
Precision	**100.0%**	96.3%	**99.0%**	94.3%	**100.0%**	90.3%	**94.3%**	76.8%
Recall	**99.7%**	96.6%	**98.7%**	79.4%	**98.9%**	96.1%	**100.0%**	96.0%

HN: Head & Neck, TA: Thoracic-abdominal, PL: Pelvis, EX: Extremity, ARL: anatomical region labeling model, SVM: support vector machine.



(a) **GT:** TA, **Prediction:** EX (2.0×10^{-4})

(b) **GT:** TA, **Prediction:** EX (1.4×10^{-2})

(c) **GT:** PL, **Prediction:** EX (3.5×10^{-3})

Figure 3.5: Coronal slices of three selected misclassified cases, with their corresponding activation map overlaid on top. The red area signify higher weight in the model decision for the predicted area. The output probability of the model class prediction is also shown for each case. GT: Ground Truth; HN: Head & Neck, TA: Thoracic-abdominal, PL: Pelvis, EX: Extremity.

3.3.2 Validation of the proof-of-concept implementation

During 22 consecutive treatment days between August and September 2022, 798 patient scans were processed and classified by the ARL algorithm. The validation dataset was composed of the first 100 unique patient scans, which were labeled by a human observer, and described in Table 3.3.

Table 3.3: Distribution of the 100 patient scans used in the clinical validation. The labeling of the dominant and less-pronounced region was performed by a human observer for each scan in the dataset, independent from the ARL prediction.

		Less-pronounced region					
		None	HN	TA	PL	EX	Total
Dominant Region	HN	33	-	8	-	-	41
	TA	21	4	-	2	1	28
	PL	16	-	11	-	2	29
	EX	1	-	-	1	-	2

HN: Head & Neck, TA: Thoracic-abdominal, PL: Pelvis, EX: Extremity.

The ARL model prediction for each of the 100 cases was compared to its respective ground truth label (dominant region), and the results of this validation study are reported in Table 3.4. Out of the 100 individual cases, two cases had an ARL model prediction-ground truth mismatch. However, it was found that each of these two cases had overlapping regions present on the CBCT scan, and the ARL model prediction matched with the referenced less-pronounced regions, as shown in Figure 3.6.

Table 3.4: Performance of the Anatomical Region Labeling (ARL) model on the 100 cases used for clinical validation. The results are shown for each three global regions separately.

	HN	**TA**	**PL**	**EX**
Accuracy	99.0%	99.0%	99.0%	99.0%
F-1 Score	98.8%	98.2%	98.3%	66.7%
Precision	97.6%	100.0%	96.7%	100.0%
Recall	100.0%	96.4%	100.0%	50.0%

HN: Head & Neck, TA: Thoracic-abdominal, PL: Pelvis, EX: Extremity.

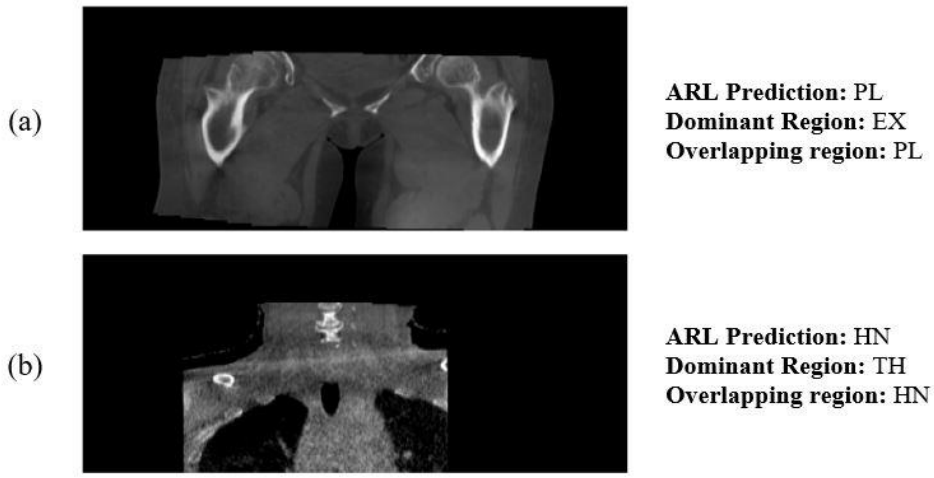


Figure 3.6: Coronal slices of the two misclassified cases in the clinical validation, with the ARL prediction and human annotations (dominant region and overlapping region) reported. HN: Head & Neck, TA: Thoracic-abdominal, PL: Pelvis, EX: Extremity.

3.4 Discussion

The ARL model presented in this study has shown high classification ability for each of the four global regions (HN, TA, PL and EX), with accuracies of 99.9%, 99.4%, 99.6%, and 99.4% respectively, outperforming the SVM model in all four regions. As compared to other CNN-based anatomy recognition algorithms developed by Roth et al. and Ouyang et al., which achieved the highest reported accuracies in the literature (94.1% and 97.3%, respectively) [66, 68], our ARL resulted in a better performance with an overall accuracy of 99.2%. However, a direct comparison between those methods is not the primary aim of this study as different imaging modalities, number of classes, and imaging planes have been used in each method. Nevertheless, the high classification accuracy produced by the ARL model demonstrates the feasibility of applying such deep learning tool to pre-treatment CBCT scans to identify the global anatomical region.

Figure 3.4 shows the input coronal slices of 12 true-positive cases with the Grad-CAM activation heat map of the ARL model overlaid on the CBCT slice. It can be observed that the regions which activated the model are in the vicinity of the craniovertebral junction for HN cases, the spine, abdominal organs and the ribs for the TA cases, and the pelvic bones for PL cases. As for the extremity cases, the model was activated by the empty regions around the patient anatomy. While this may not be the most logical and robust way of identifying extremity cases to a human observer, this feature is a characteristic of most extremity scans. However, it must be noted that the amount of extremity cases in the training dataset was limited, which may be the source of the decrease in performance for EX classification.

Out of the 1,090 scans, 9 scans were wrongly classified by the ARL model. Figure 3.5 illustrates some misclassified cases, with their corresponding activation maps overlaid on top. It can be observed from Figure 3.5(a) and 3.5(b) that the limited FOV resulted in a wrong classification of the thorax as an extremity due to the empty spaces around the patient. On Figure 3.5(c), the presence of metal artifacts may have been the cause of the misclassification as shown by the heatmap. A potential solution would be to use activation gates [53] within the ARL model such that it focuses on targeted regions instead of irrelevant regions on the image.

Nevertheless, our proof-of-concept implementation and validation study have shown that the ARL model predictions correlate with the human observer annotations, with accuracies of 99.0% for all four global regions. Out of the 100 cases, two cases had an ARL model prediction-dominant region mismatch, as shown in Figure 3.6. However, it can be observed that the ARL model prediction was still consistent with the overlapping region present on the scan in both cases. The results of this validation study hence reinforces the relevance and ability of the ARL tool to label CBCT images from daily treatments.

To be more robust to the entire patient population, the current algorithm could be further refined to accommodate for outlier cases, such as extremity treatment scans. However, these types of CBCT scans are seen more sporadically in the clinical setting due to the rare occurrence of soft tissue sarcomas [73], leading to too few cases for optimal model training or refinement. Furthermore, the ARL model was trained and tested on only a single institution's data. To validate and improve the generalizability of the ARL model on other facilities' datasets, a multi-institutional study needs to be performed, which will be part of our future studies.

Another limitation of the current ARL model is that it uses a single 2D coronal slice which contains limited anatomical information as compared to the whole 3D image. A 3D Dense-Net [74] may potentially improve the performance of the ARL model by obtaining more useful features as compared to the current 2D model [75]. However, training a 3D CNN is computationally expensive and the inference time of the tool will be higher with our current system. With the increased availability of high-performance Graphics Processing Units, this 3D method may be feasible in the future.

3.5 Conclusion

In this work, a CNN-based Anatomical Region Labeling (ARL) model was developed to classify pre-treatment CBCT scans into four regions, namely head & neck, thoracic-abdominal, pelvis, and extremity. Our results have shown strong agreement between the model predictions and human annotations for all four regions, confirming the strong performance of the model. The ARL model may be employed in the clinical setting as a pre-processing step for radiotherapy tools which have been developed for pre-treatment CBCTs containing specific anatomical regions, such as auto-segmentation algorithms, patient setup error detection algorithms, and

radiomics tools for early treatment response assessment. Furthermore, the tool may be used as a quality assurance check by comparing the model's prediction to the treatment site to avoid wrong-site radiotherapy treatment.

3.6 Open-Source Code Access

The python scripts for the DQR and ARL algorithm are available on the following website: https://github.com/dcluximon/ARL_repo

Chapter 4: Proof-of-Concept Study of Artificial Intelligence-Assisted Review of CBCT Image Guidance

4.1 Introduction

As technology and techniques for radiotherapy treatment have evolved, the prevalence of image guidance has increased. Since image-guidance is one of the final steps in the radiotherapy workflow prior to initiating beam delivery, any error during pre-treatment imaging and positioning can severely impact treatment. [76-77] As such, multiple levels of quality assurance have been implemented which require human observers to review pre-treatment image registrations. Therapists review all registrations at the console during treatment, physicians review and approve all image guidance daily, and physicists typically perform some level of review during weekly chart checks. With increased use of image-guidance, the workload for each of these reviewers increases in kind, carrying a danger of inducing fatigue in the reviewer and allowing errors to pass by undetected. [78] Quantifying registration performance has historically been a difficult task. [79-80] Intensity-based metrics (such as correlation coefficient) and feature-based metrics (such as mutual information) have many limitations, including sensitivity to intensity range differences and artifacts. In addition, most on-board tools to assist in pre-treatment image registration rely on these metrics, so they offer little additional benefit as a secondary safety layer.

Recent years have seen a significant effort to apply artificial intelligence (AI) and deep learning techniques in radiotherapy with the aim to improve overall quality and efficiency by utilizing the abundance of prior data to standardize and optimize steps in the radiotherapy workflow. Recent publications detail the efforts in the areas of automatic segmentation,

treatment planning, treatment optimization, patient-specific QA, treatment log analysis, and plan adaptation. [81-82]

There remains a significant gap in the radiotherapy workflow where AI and deep learning have not yet been applied – image review. As discussed in the previous paragraph, most current applications are built on the treatment planning data to include dose distributions, structure sets, and dose volume constraints. McNutt et al expanded the scope and discussed the application of big data for QA purposes, particularly for identifying anomalies. [81] However, patient treatments continue producing data throughout the entire course of treatment, and more so now than ever with the increased reliance on image-guided radiotherapy (IGRT) techniques.

In 2020, the American Association of Physicists in Medicine published the findings of their Task Group 275, on effective strategies for physics plan and chart review. [19] In this report, they recommend that software vendors develop methods to automate chart reviews, and 'highlight items that are difficult to check and review.' Additionally, AAPM's recently published practice guidelines for plan and chart review emphasizes the safe application of computer-aided programs by 'calling special attention to missed or mismatched items' but should not fully replace a thorough and robust chart review. [20]

With most of the field transitioned to electronic medical records, tools have been developed to automate routine chart checks – comparing logistical data within the database to identify and highlight discrepancies. [20, 83-84] To the authors' knowledge, there is no current tool available on the market to provide an automated, independent evaluation of the pre-treatment image-guided patient alignment and anatomy-of-the-day.

In this work, we examine the implementation of a deep learning algorithm as a decision-support tool for image review in weekly physics chart checks. Algorithms were previously developed for the detection of IGRT setup errors such as misalignments of 1-2 cm or more, alignments to the incorrect vertebral body, and anatomic misidentifications using IGRT images. [24-25, 85] These algorithms were originally developed to detect rare but serious gross errors and return a misalignment score based on similarity of the aligned IGRT image with the planning CT, accounting for applied IGRT shifts, as described in **Chapter 2**. An appropriately high threshold value of the misalignment score detects gross errors with high sensitivity and specificity. Recently, we hypothesized that an intermediate threshold could be used to differentiate perfectly aligned cases needing minimal human review, from imperfectly aligned cases that, while not gross errors, require human review, clinical judgement, and potentially remediating actions. In this manuscript we evaluate a working prototype of such a system. We discuss considerations of clinical implementation, including strategy, validation, impact, effectiveness, and efficiency.

4.2 Materials and Methods

4.2.1 Clinical Implementation of EDA for Physics Chart Reviews

In accordance with the recommendation in the AAPM practice guidelines [82] the intent of integrating automated software-based supervision into the radiotherapy QA process was not to supplant weekly physics chart reviews, but to supplement it and aid in identifying cases which may need closer inspection. As such, our implementation strategy was to highlight a subset of cases each day for in-depth investigation.

The proof-of-concept implementation was developed with Python scripting and utilized a DICOM networking protocol to query and retrieve data from the clinical record and verify (R&V) system. The prototype system was developed to interface with the ARIA R&V system (Varian Medical Systems, Palo Alto, CA) using the pynetdicom Python package. Results were compiled into daily reports and aggregated into an interactive dashboard. The workflow is illustrated in Figure 4.1, and the nine modular components can be described as follows: (1) Query the clinical database for a list of daily treatments. (2) Query the clinical database for a list of daily cone-beam CT (CBCT) acquisitions. (3) Cross-reference lists to identify patients for analysis. (4) Retrieve relevant DICOM Registrations (REGs) for identified patients. (5) Inspect REGs, and retrieve referenced RTPlans. (6) Inspect RTPlans, and retrieve associated RTStructs, planning CT images, and CBCT images. (7) Run AI-based misalignment model on each dataset. (8) Compile predictions into a daily report. (9) Archive intermediate results, logs, and remove temporary files.

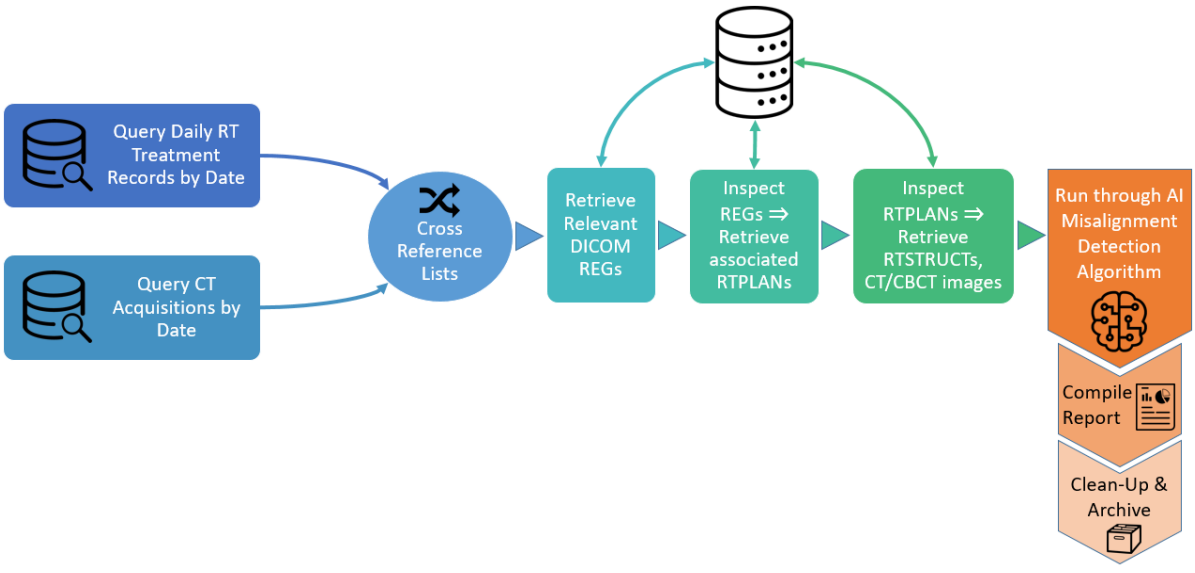


Figure 4.1. Workflow illustration of the automated image retrieval and analysis pipeline.

A web-based dashboard was implemented to facilitate access to the daily reports, where results are sorted by prediction value and highlight the most likely misalignments from the previous days' treatments. Additional features were developed to facilitate a superficial review, but with the stipulation that full in-depth review should still be performed using the R&V system's built-in tools.

These features included single-slice CT-CBCT fusions along each axis to quickly identify if there was an issue with the data retrieval or preprocessing. The user may also scroll quickly through the patient's past treatments to compare the registrations or look for trends. Lastly, summary statistics are compiled and charted, allowing breakdown by machine, date, treatment site, prediction value etc. to help explore overall data trends in the clinic.

The AI-algorithm scored the likelihood of a misalignment between 0 and 1: 0 being most unlikely and 1 being most likely for a misalignment. An initial threshold was implemented at 0.1 with the intention of flagging 5-10% of cases for further investigation.

4.2.2 Clinical Validation Study of the AI-based Image Review Tool

A clinical validation study was performed to assess the suitability of the chosen threshold and the ability of the algorithm to identify setup images that might, in the clinical judgement of a medical physicist, require further investigation. In an IRB approved study, the proof-of-concept implementation was run retrospectively on data from treatments between February 7th 2022 and May 10th 2022. Treatments were performed on 4 machines: 1 NovalisTx, 1 NovalisSTx, and 2

TrueBeams (Varian Medical Systems, Palo Alto, CA). A total of 1357 treatment cases comprising registered (post-shift) CBCT images, were analyzed from 197 unique patients.

From this data, 100 cases were selected for expert review based on the following criteria: 50 cases having the highest misalignment scores (lowest model prediction – 0.495), and 50 randomly selected cases from the rest of the population. Each case was from a unique patient to avoid image-score correlations. The list of cases was then randomly permutated. Three independent observers, board certified medical physicists, manually reviewed each of the 100 cases and scored them from 1 to 4. Review criteria were based on clinical action levels, and were scored based on overall alignment, taking into account the reviewer's estimation of whether a given target should be aligned to soft tissue or bone. The numerical review score was defined as follows: 1 showing a perfect match of target and surrounding anatomy; 2 showing some deviation, but clinically acceptable; 3 showing enough deviation to require further investigation, and 4 showing significant deviation that would preclude treatment until investigation is resolved.

The mean values of the reviewer's score were correlated with the AI prediction. Algorithm performance at discriminating cases with mean overall score greater than 2 (i.e., action level requiring investigation) was quantified using a receiver operating characteristic (ROC) curve. As a random selection process was applied to the stratified patient dataset (50 highest scoring cases, then 50 randomly chosen from the remaining 147 cases), the weight of each sample was calculated by taking the inverse of the sample proportion from each of the two strata. The weight of a sample found in the lower 147 cases was 147/50 = 2.97. The samples found in the 50 highest scoring cases were assigned a weight of 1 as the whole population was used in the analysis. Using these weights, the weighted sensitivity and specificity were calculated and used to build a weighted ROC curve. This weighting method has been established as an

effective method to extrapolate the findings from a sample study to the entire dataset [86]. Additionally, selected representative cases were examined further with temporal trend-line plots.

4.3 Results and Analysis

Data was automatically collected and processed over a 45 day period, resulting in 1357 registrations from 197 unique patients being analyzed. The distribution of registrations by anatomical region included 506 Head & Neck (HN), 464 Pelvis (PL), and 387 Thoracic-Abdominal (TA). To ensure the validity of our proof-of-concept study, the patient population for this study was kept independent of the patient population used for model training and validation.

The full distribution of observer scores obtained from the 100 case reviews are shown in Table 4.1. Figure 4.2 illustrates the relationship between the observer scores and model predictions for the 100 registrations in the validation set. Cases were binned by average observer score, and a box and whisker plot was constructed to show the distribution of the model predictions for each group.

Table 4.1: Absolute count (N) of the observer scores for each individual expert in our study.

Observer Score (X)	N (Observer 1)	N (Observer 2)	N (Observer 3)
$1 \leq X < 2$	65	77	35
$2 \leq X < 3$	32	20	59
$3 \leq X < 4$	3	2	6
$X = 4$	0	1	0

After confirming the correlation of model predictions to observer scores, the focus shifted to determining an optimal threshold prediction when flagging cases for further investigation. Ideally, a threshold could be found that would catch all cases tagged by the observers as less than ideal, while also minimizing false positives. It was considered a priority to minimize false positives to limit the time required for daily review and avoid inducing alarm fatigue.

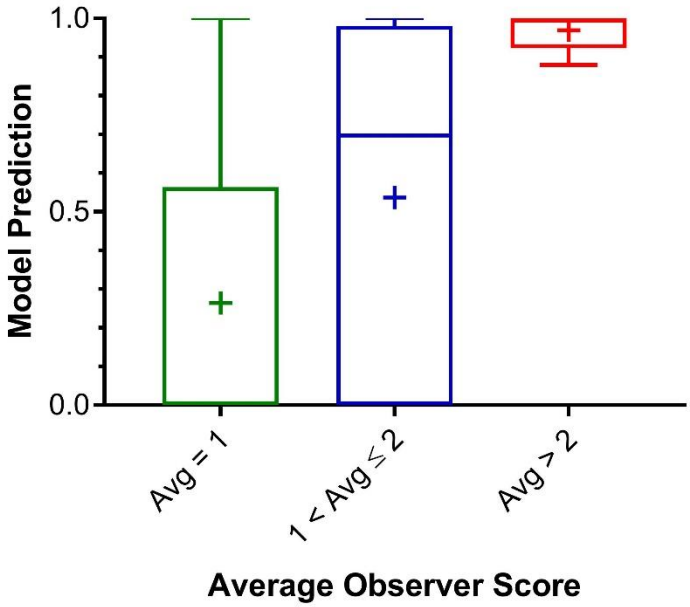


Figure 4.2: Box and whisker plot to show the distribution of the model predictions, grouped by average observer score. The box shows the median, 25th percentile, and 75th percentile, while the whisker shows the minimum and maximum values. The cross within each box represents the mean model prediction for the respective group.

Figure 4.3 plots the receiver operating characteristic (ROC) curves for the validation set, as well as individual ROC curves for each anatomical region. True positives were categorized as cases where the mean observer score was greater than 2, where a score of 2 was considered clinically acceptable but not perfect. From the ROC analysis, it is apparent that using a threshold

of 0.87 achieves 100% sensitivity while minimizing false positives. Applying this threshold to the validation data set of 100 cases, 40 would be flagged for further investigation. Note, this proportion is not reflective of a broader patient population since the validation dataset included the 50 cases receiving the highest model prediction scores (mean prediction of 0.91 (Top50) vs. 0.05 (the rest) – p-value < 0.0001). Inspecting the observer scores for the stratified validation dataset shows an average of 1.65±0.51 for model predictions ≥0.87, and 1.33±0.33 for model predictions <0.87. A two-tailed t-test results in a p-value of 0.0002 between these cohorts.

4.4 Discussion

To explore how the proposed tool could potentially impact clinical workflow, we present case studies for discussion. With most clinics transitioned to electronic medical records (EMR), it is trivial for software tools to compare values between database entries and highlight discrepancies. The task is more difficult and nuanced for image review, and requires both broad clinical knowledge and patient-specific insight to determine relevance and priority. The proposed deep learning pipeline aims to provide a quantitative analysis of daily pre-treatment CBCT alignment, and has the potential to facilitate the recognition of anomalies. Thresholding may be used to identify a manageable number of datasets for manual review. A trendline may provide added value when used in parallel to the hard-thresholding method.

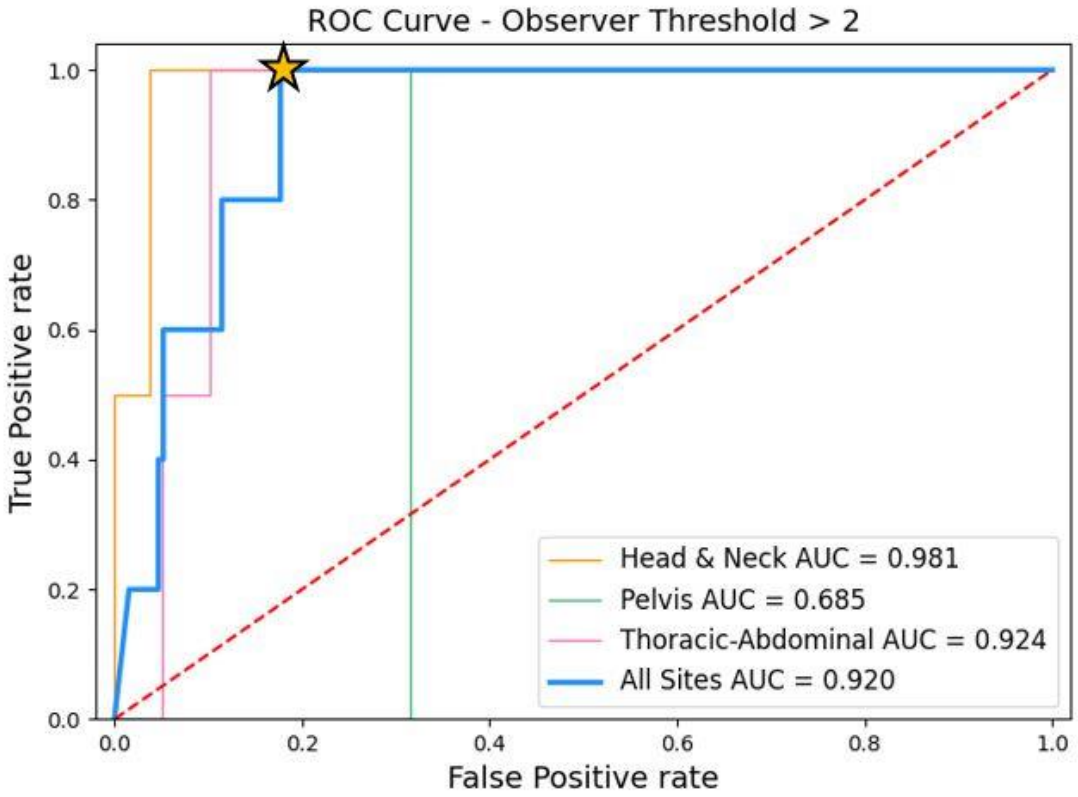
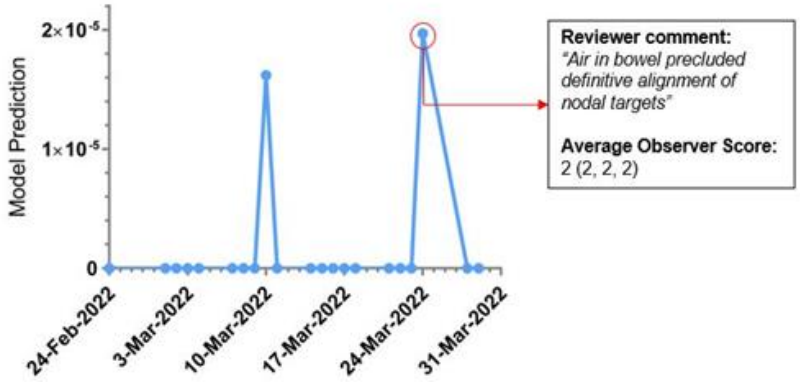


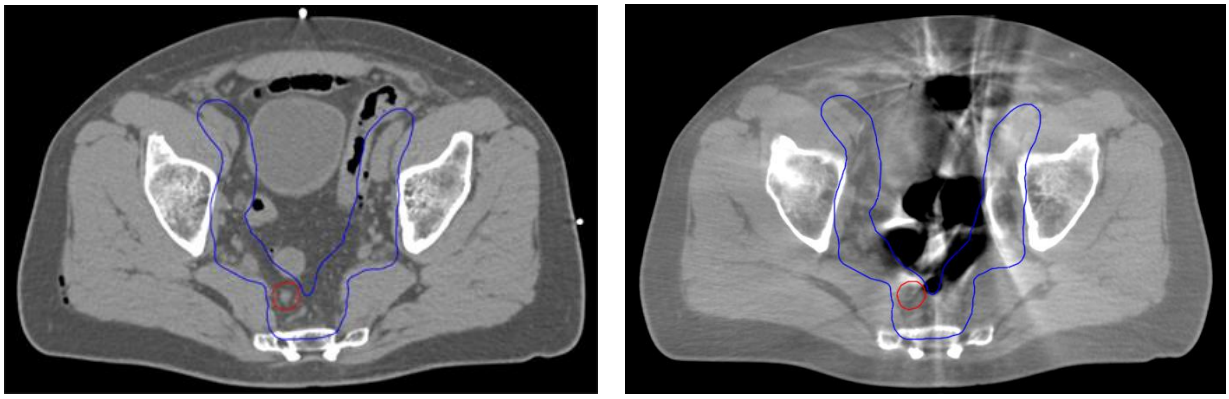
Figure 4.3: Weighted Receiver Operating Characteristic (ROC) curves obtained using a mean observer threshold > 2. The bold blue curve represents the results from the entire 100 patient validation dataset, and the other curves represent the results from each respective anatomical region only. The Area Under Curve (AUC) is also given for each curve. The model prediction threshold (0.87) leading to a sensitivity of 100% and a specificity of 82% (depicted by the yellow star) was obtained and used for further analysis.

One example is illustrated in Figure 4.4. The timeline in Figure 4.4(a) plots the model prediction from each fraction by date over the patient's treatment course. While none of the model prediction scores approach our global threshold to be flagged for further inspection, there are clearly two fractions where the deep learning model identified an increased probability of misalignment. One of these fractions was included in our validation dataset, and an observer commented, "Air in bowel precluded definitive alignment of nodal targets." Figures 4.4(b) and 16(c) display the anatomy from this fraction, where increased air cavities in the bowel resulted in

artifacts on the CBCT image which made target identification and pre-treatment setup more difficult.



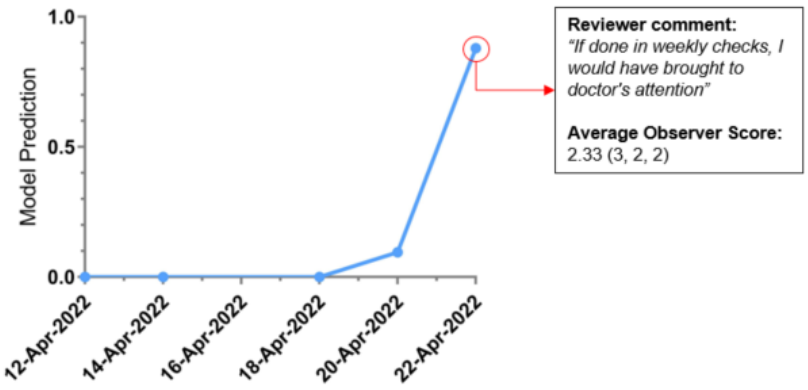
(a)



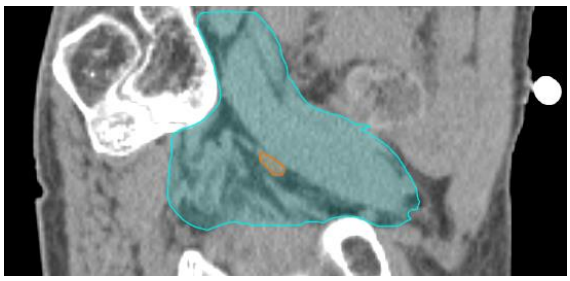
(b) **(c)**

Figure 4.4: This diagram depicts a case where the deep learning model would not have flagged the registration due to the low prediction score (2×10^{-5}). However, the trendline in (a) suggests a relatively high escalation in the model prediction on that particular day (March 24th 2022) as compared to the previous treatment days, demonstrating the potential value the trendline can provide when used in parallel to the hard-thresholding method. Images (b) and (c) are select axial slices from the planning CT and pre-treatment CBCT (March 24th 2022), respectively. The presence of gas in the bowel resulted in artifacts on the CBCT image and inhibits identification of targets within the 50 Gy planning tumor volume (PTV) coverage (blue contour) and the 62.5 Gy gross node PTV coverage (red contour).

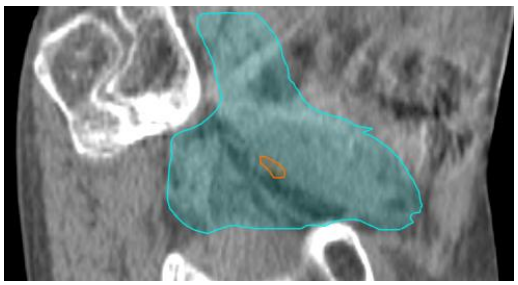
For the case illustrated in Figure 4.5, the trendline shows consistently good alignment for the first few fractions, then a relative increase for fraction 4, before a dramatic jump on fraction 5. Reviewing the images, there is a clear shift of the bony anatomy between the plan (Figure 4.5(b)) and the CBCT (Figure 4.5(c)). Additionally, the lymph node boost volume is difficult to visualize but may be posterior to the nodal PTV in the CBCT image. Fraction 5 was included in the validation dataset, and one of the observers noted, "If done in weekly checks, I would have brought to doctor's attention."



(a)



(b)



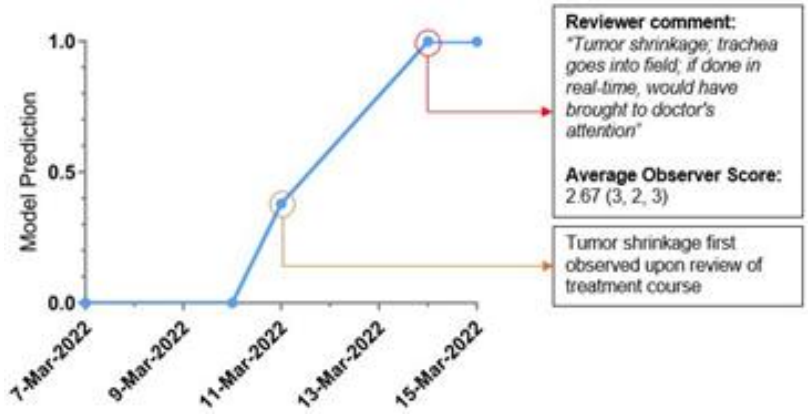
(c)

Figure 4.5: Highlight of a case which obtained a relatively high mean observer score (2.33) and a high model prediction score (0.88). The trendline in (a) shows the model prediction scores of the registrations performed over the course of the patient's treatment. The red circled point represents the case which was reviewed by expert observers for validation. Images (b) and (c)

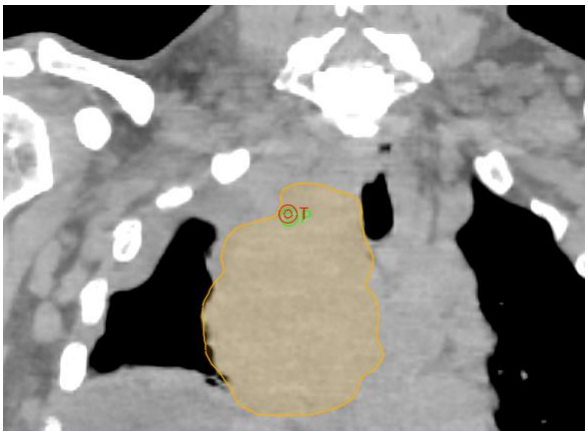
are select sagittal slices from the planning CT and setup CBCT (April 22nd 2022), respectively. Differences in the 25Gy planning tumor volume (PTV) coverage (blue contour) and in the gross node PTV coverage (orange contour) can be observed between (b) and (c).

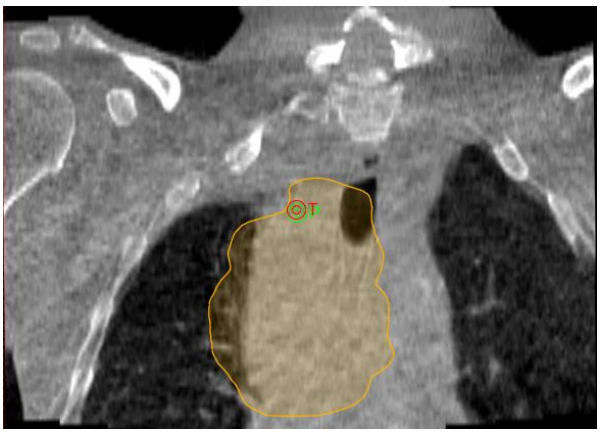
To contrast the case in Figure 4.5, which displayed subtle differences that could be attributed to user judgment during registration at the treatment console, the case illustrated in Figure 4.6 shows drastic anatomic changes. The patient had a large mediastinal mass that shrunk significantly over a 5-fraction treatment course, as well as pleural effusion that showed improvement. The fourth fraction was included in the validation dataset, and the tumor shrinkage was noted with two observers scoring the case "3" and one commenting, "Tumor shrinkage; trachea goes into field. If done in real-time, would have brought to doctor's attention." The model prediction timeline shows a clear progression, for longer treatment courses or more drastic anatomic changes, this could be a valuable tool to anticipate re-planning.

While a deep learning model can identify and quantify differences between the planning CT image and the daily CBCT image, a human observer is still required to review the flagged cases and judge whether the differences are clinical relevant and actionable. To further aid the observer, a debugging tool was incorporated into the proof-of-concept pipeline, which overlays a heatmap of the model activation on the patient anatomy, indicating the areas contributing to the misalignment prediction. Figure 4.7 demonstrates this debugging tool on a bilateral neck case. Figure 4.7(a) shows a sagittal slice the CT-CBCT fusion with target contours overlaid. A colormap fusion is displayed in Figure 4.7(b), with the planning CT in the green channel and the CBCT in the red and blue channels. The activation heatmap is overlaid on the planning CT in Figure 4.7(c). This heatmap could be used as a debugging tool for the end-user to visualize the model prediction.



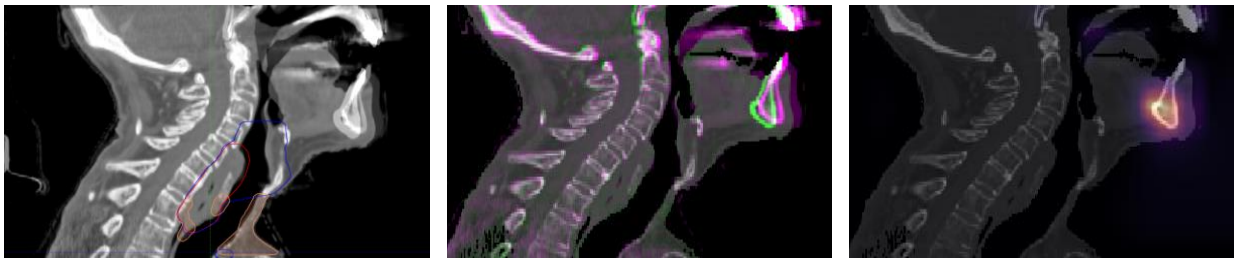
(a)





(b) **(c)**

Figure 4.6: Highlight of a case which obtained a relatively high mean observer score (2.67) and a high model prediction score (1.0). The trendline in (a) shows the model prediction scores of the registrations performed over the course of the patient's treatment. The red circled point represents the case which was reviewed by expert observers for validation. The other registrations were reviewed post-analysis for comparison. (b) shows a selected coronal slice from the planning CT, with the Planning Tumor Volume (PTV) shown as the yellow overlay and the treatment isocenter shown as the red target. (c) shows the pre-treatment CBCT from March 14th, 2022.



(a) **(b)** **(c)**

Figure 4.7: Debugging the misalignment predictions with a model activation heatmap. (a) shows the CT-CBCT fusion with target contours overlaid, (b) shows a colormap fusion of planning CT (green) and CBCT (purple), and (c) shows the activation map of our deep learning model overlaid on the planning CT. The mandible is clearly misaligned in this case, and the heatmap shows a hotspot for model activation at the mandible. This feature provides an avenue for the user to better understand the reason behind the model's misalignment predictions.

The case studies above illustrate the potential uses of the pipeline. Nevertheless, some limitations apply to the current proof-of-concept system. The deep learning model, at least in its current iteration, does not incorporate clinical context and determine relevance of a misalignment. Additionally, as with many computer vision tools, it will preferentially focus on high contrast image features such as bone and tissue-air interfaces. Lastly, while intended to be vendor agnostic, utilizing only DICOM networking protocol to interface with the clinical database, the proof-of-concept pipeline was developed referencing only a single vendor's DICOM conformance statement and the built-in DICOM queries reflect that. Connecting to another vendor R&V system, or even a different version of the same vendor's R&V system would almost certainly require further development of the DICOM handling.

4.5 Conclusion

Due to time and manpower constraints, it is often not feasible for physicists to perform in depth examination of every pre-treatment alignment registration. As discussed in the

introduction, image guidance is one of the most pivotal steps in the treatment workflow, with a significant risk that a mistake could lead to a mistreatment.

In this work, we presented a proof-of-concept clinical implementation of an automated pipeline for AI-assisted CBCT alignment retrospective review. The purpose of the pipeline was to provide quantitative assessment and visualization tools to improve the efficiency and efficacy of periodic chart reviews. The predictions of the deep learning model were validated against expert observers, and demonstrated that a prediction threshold could be identified to stratify pre-treatment images with a statistically significant correlation to the observer scores. In addition to the validation study, we demonstrated through anecdotal examples how these tools could be beneficial to the clinical workflow through quantification of daily alignments and patient/plan specific timelines to identify trends and flag anomalies. Effective visualization of this data, made quickly accessible and easily digestible, can expedite the image review component of periodic physics chart checks.

Future work will aim to leverage the speed of deep learning inference to move this system from retrospective to real-time, integrating directly with the treatment machine to interlock the beam if the AI-model flags a potential setup misalignment. Such a system would ideally require the operator to either revise the alignment or acknowledge the interlock before proceeding with treatment. An automated AI-assisted tool could allow for independent, quantitative review of every alignment registration in the future.

Chapter 5: Results of an AI-Based Image Review System to Detect Patient Misalignment Errors in a Multi-Institutional Database of CBCT-Guided Radiotherapy Treatments

5.1 Introduction

In radiation oncology, primary treatment goals are the precise delivery of radiation such that the dose to the target is optimized for better tumor control, while the doses to the organs at risk are minimized to limit side effects. Studies have shown that incidents regarding incorrect dose delivery are still prevalent even with safety procedures and technologies, such as image-guidance. [9-11] As reported earlier, between 2014 and March 2023, a total of 3,730 therapeutic radiation incidents were reported to the Radiation Oncology Incident Learning System (RO-ILS) Portal, with 18.4% of those being identified as having severe or moderate severity scores. [9]

According to the American Association of Physicists in Medicine (AAPM) Task Group 100, the patient positioning step within the external beam radiotherapy (EBRT) workflow is a high severity and high-risk failure mode, ranking in the top 20% most hazardous steps in the entire external radiotherapy workflow following a failure modes and effects analysis (FMEA). [12] Ezzell et al. have shown that in a cohort of 336 critical events submitted to RO-ILS between 2014 and 2016, 34 such errors occurred due to the wrong shift performed at the treatment table, with 29 of them reaching the patient. [15]

With the widespread adoption of stereotactic radiosurgery/stereotactic body radiation therapy (SRS/SBRT) and the introduction of ultrahigh-dose-rate treatments such as FLASH, proper patient setups become even more critical due to their high-dose per fraction. In a 2017 study of RO-ILS SRS/SBRT events, Hoopes et al. have found that one of the most common

event types was the incorrect shift and alignment of the patient. [13] McGurk et al. further reinforced this finding when they discovered, in a 2023 study involving four institutions in the United States, a patient who was wrongly aligned for one of five of their multilevel spine SBRT fractions. [14] Those findings have shown that efforts are needed to find and analyze cases of reported and unreported patient setup incidents so that the characteristics and causes of these events can be understood and the treatment workflow can be consolidated accordingly to enhance patient safety.

The American College of Radiology (ACR)-American Society for Radiation Oncology (ASTRO) Practice Parameter for Image-Guided Radiotherapy, the AAPM Task Group (TG)-275 and the Medical Physics Practice Guidelines (MPPG) 11.a all provide recommendations on how to mitigate alignment-based failure modes and promote incident learning as a way to reduce future events. [8-10] However, those reports note that many components of the current safety checks, including those involving patient setups, are heavily human-reliant and are therefore prone to be overlooked due to the fast-paced working conditions in the clinic. For instance, McGurk et al. have shown through a Human Factor Analysis and Classification System that 95.2% of 189 reported SBRT safeguard failures occurred due to human errors. [14]

As recommended by TG-275 and MPPG 11.a, the use of automation during these safety checks can act as a safety barrier and help in the analysis of bulk data for efficient incident learning by identifying error pathways that may not be easily detected by a human reviewer. Our group has previously developed deep learning-based algorithms for the detection of simulated gross misalignments for 2D planar x-ray image guidance [33] and cone beam computed tomography (CBCT) image guidance [85]. In prior studies, those algorithms were trained and tested on simulated patient misalignment errors, showing high sensitivity (>85.5%) in detecting

the simulated misalignment errors for a fixed specificity of 95%. Furthermore, an automated pipeline using the CBCT-based patient misalignment detection algorithm was developed to facilitate offline image reviews and showed promising results when validated against expert medical physicists through a feasibility study. [87] In this work, we will be applying the previously developed AI-based patient misalignment detection pipeline to perform a bulk retrospective patient setup error search on CBCT-guided radiotherapy treatments at two radiotherapy centers.

The primary goal of this study is to perform a measurement of the rate of gross patient setup misalignment errors in CBCT-guided radiotherapy at two large academic centers. To the best of our knowledge, this is the first study to apply an AI-based image review system for a bulk retrospective patient misalignment error search on CBCT-guided radiotherapy treatments and to report previously unknown patient setup misalignments from the two institutions. The following points highlight the major contributions of this work:

a) A manual retrospective patient misalignment search is infeasible due to the large number of cases to be reviewed. By performing this AI-assisted image review, an absolute gross patient misalignment error rate in CBCT-guided radiotherapy at two radiotherapy facilities was determined, which is an important aspect of understanding radiotherapy safety.

b) While previous studies have shown promising results for gross patient setup misalignment detection, the model validations were only performed on simulated errors [85] and no gross patient misalignment error was found during the proof-of-concept implementation study [87]. This study focused on detecting and reporting real-world

incidents, which is a fundamental step towards a robust validation of the real-world clinical performance of the tool, as suggested by the AAPM Task Group 273 [88].

5.2 Materials and Methods

5.2.1 A Retrospective Error Search using the AI-based pipeline

The DQR software described in Chapter 2 Section 2.2.2 was used to interact with the ARIA image management system to collect registrations performed between 2016 and 2017 at the University of California Los Angeles Medical Center (UCLA), and between September 2021 and September 2022 at the Virginia Commonwealth University Medical Center (VCU). Those images were automatically sent to the EDA for planning CT-setup CBCT registration analysis, as shown on Figure 5.1. As described in Chapter 2, the EDA was trained separately on HN, TA and PL images. Lower extremities (including glutes, thigh, knee and calf) and upper extremities (including forearm, upper arm, and shoulder) were not used in training due to their scarcity. In the retrospective search, those extremity cases were assigned to the HN, TA or PL class by the Anatomical Region Labeling (ARL) model [121] described in Chapter 3, which would then redirect the scans to the corresponding error detection model dealing with features resembling the most to the extremity site. Cases in which the imaged anatomy overlapped the HN, TA and PL regions were assigned to the closest matching region by the ARL.

Registrations resulting in a score greater than their respective tr_{90} (threshold leading to a sensitivity $\geq 90\%$), were automatically flagged for human review. Additionally, observing the trends in the misalignment probability prediction over the patient treatment course can also add

value in detecting anomalies. Hence, registrations resulting in a considerable jump in misalignment score compared to the adjacent treatment fractions scores (ratio of predictions > 10^4) were also identified; among those cases, registrations with a score in the range of $tr_{99} \leq$ score < tr_{90} were flagged for review. The tr_{90} and tr_{99} threshold values are described in Table 2.2.

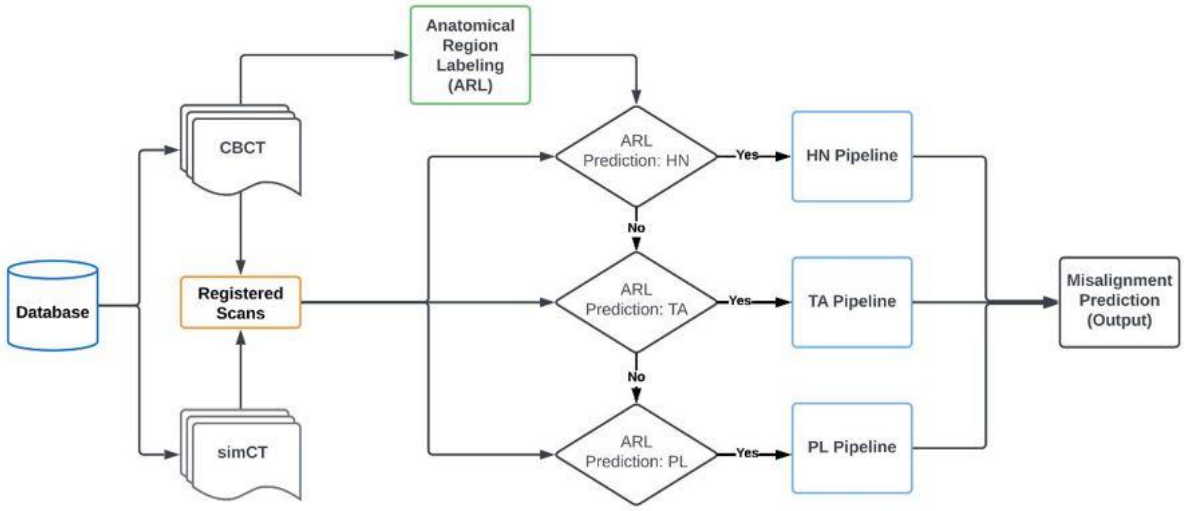


Figure 5.1: Illustration of the error detection algorithm used in this retrospective patient setup misalignment error search. SimCT: planning CT, HN: Head & Neck, TA: Thoracic-Abdominal, PL: Pelvis.

Flagged cases were subsequently reviewed by a human expert and classified as error (true-positive) or no-error (false-positive). The reviewing process and final judgement of each case were performed by two clinical medical physicists with more than 15 years of experience in the field of radiation oncology. During the case reviews, a true-positive event was one that considerably deviated from the correct simCT-CBCT registration at the target (> ~1cm), leading to a gross tumor volume (GTV) under-coverage and subsequently, a considerable deviation in

the prescribed dose to the target. This definition corresponds to a level of significance that is at least reportable to the institutions' incident learning systems. The gross patient misalignment error rate in CBCT-guided radiotherapy at the two institutions was subsequently calculated based on the number of true-positive events found. Given that the incidents that occurred are rare, independent and discrete, with the number of registration analysis being very large, the variability in the error rate was determined using a Poisson approximation to the binomial distribution.

Additionally, for each incident found, a dosimetric analysis was performed over the corresponding treatment course to understand the dosimetric impact of the misalignment at the clinical target volume (CTV). The D95 (minimum dose covering 95% of the target volume) values were calculated for the single mis-delivered treatment fractions, as well as for the whole treatment course (Accumulated D95). Each D95 value was compared to the corresponding prescribed dose in this assessment. The dosimetric analysis was performed on MIM software (MIM Software Inc., OH, United States).

Within the date range of this retrospective study, three patient setup incidents involving CBCT guidance were known at the two institutions. These known setup incidents were the result of off-by-one vertebral body misalignments and had been submitted to RO-ILS as part of the institutions' quality and safety protocols. To validate the real-world clinical performance of EDA, the true-positive cases found following human review were compared to the known incidents.

5.3 Results and Analysis

During the retrospective study, 11,747 and 5,865 registrations (from 1,801 and 613 patients, respectively) were processed by EDA from institutions 1 and 2, respectively. Using the hard-thresholding method and the trending analysis method described in Section 5.2.1, 1,028 and 334 events (from 470 and 161 patients, respectively) were flagged from UCLA and VCU, respectively. Further details regarding the number of the flagged cases are shown in Table 5.1. As compared to performing a fully-manual retrospective review, requiring 880 hours of human effort (assuming an average of 3 minutes per fraction), the AI-aided review only required an average of 68 hours of human effort, hence being considerably less laborious.

Table 5.1: Summary of the AI-assisted retrospective patient setup error search performed at the two radiotherapy sites.

Radiotherapy Site	Date Range	Treatment Region	Number of registrations processed by EDA	Number of registrations flagged for human review		
				Via Thresholding	Via Trending Analysis	Total
UCLA	Jan 2016 - Dec 2017	HN	4,583 (480*)	167	104	271 (117*)
		TA	3,252 (580*)	250	199	449 (192*)
		PL	3,912 (741*)	258	50	308 (161*)
VCU	Sep 2021 - Sep 2022	HN	1,897 (172*)	12	7	19 (13*)
		TA	1,860 (283*)	86	87	173 (102*)
		PL	2,108 (158*)	123	19	142 (46*)
Total			**17,612 (2,414)**	**896**	**466**	**1,362 (631)**

**Represent the number of patients. HN: head & neck; TA: thoracic-abdominal; PL: pelvis; EDA: error detection algorithm*

Among the 1,362 of flagged events, seven off-by-one vertebral body misalignment incidents (true-positives) were found, as shown in Figure 5.2, translating to an absolute gross patient misalignment error rate of 0.04% ± 0.02% in CBCT-guided radiotherapy. Three of the

seven events (Figures 5.2e-5.2g) were the known incidents, and the other four events (Figures 5.2a-5.2d) were previously-unreported incidents. Of those seven cases, five were caught via the hard-thresholding method (Figures 5.2a-5.2d and 5.2g), and two were caught via the trending analysis (Figures 5.2e and 5.2f).

Those seven mis-delivered fractions were from four individual treatment courses, with the magnitude of misalignment ranging from 1.85 cm to 2.5 cm at the clinical target volume (CTV). For each incident, the percent dose deviation of the CTV resulting from the misalignment (as compared to the prescribed dose per fraction) ranged from 44% to 99%. Further details about the treatments and delivered doses (including accumulated doses) are presented in Table 5.2 and Appendix A.2 (case presentations).

The rest of the flagged registrations from the two institutions, while often demonstrating some imperfections such as patient rotations, patient weight loss, soft-tissue differences (caused by tumor growth/shrinkage or deformable organs) and substantial CBCT image artifacts, were found to not be severe enough to be reportable and be labeled as incidents. Some of those false-positive cases are highlighted on Figure 5.3, showing the EDA's potential at flagging cases containing not only systematic shifts from the registration, but also other clinically-relevant imperfections on the registered scans.

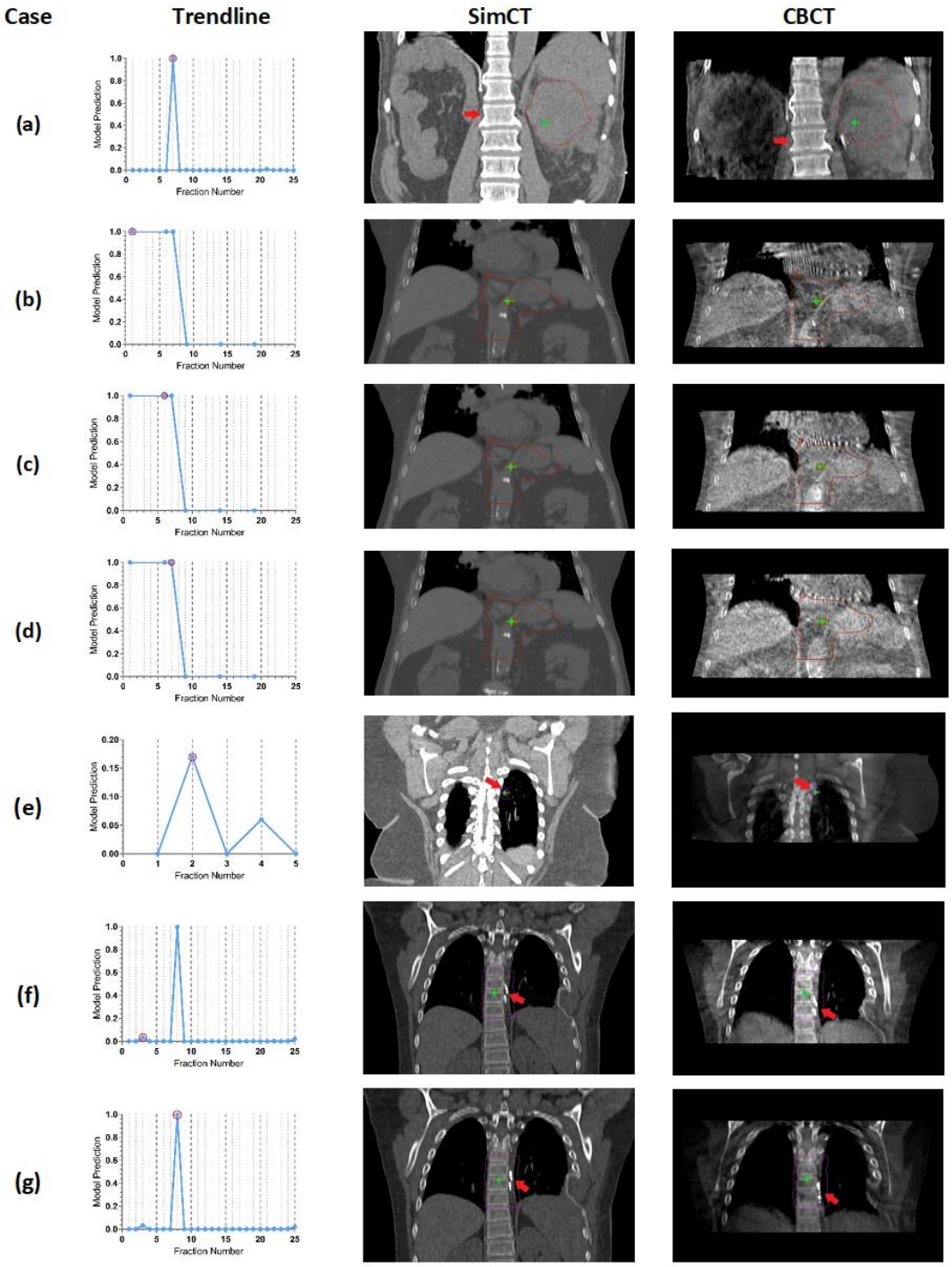


Figure 5.2: The seven incidents (a-g) found during the AI-assisted retrospective error search. For each case, the trend in the model predictions over the treatment course is shown, with each blue dot representing a treatment fraction and the incident circled in red. Additionally, selected coronal planes of the simCT and CBCT (at the corresponding slice location) are displayed for each incident. The contours present on the simCT and CBCT images represent the planning

target volume (PTV) used during treatment, the green star represent the treatment isocenter and the red arrows highlight landmarks that reveal the misalignments (if present).

Table 5.2: Summary of the dosimetric analysis performed on the treatments where patient misalignment incidents were found during the retrospective study.

Treatment description	Number of fractions with patient misalignment	Magnitude of misalignment	Was incident(s) reported or known?	D95 for the CTV, single misaligned fraction	Accumulated D95 for the CTV, with patient misalignment(s)
Abdomen IMRT, with a dose prescription of 87.5 Gy over 25 fractions (3.5 Gy per fraction).	1	2.4 cm	No	1.35 Gy	86.5 Gy
Stomach IMRT, with a dose prescription of 45 Gy over 25 fractions (1.8 Gy per fraction).	3	2.2-2.5 cm	No	1.00 Gy	36.78 Gy*
Lung SBRT, with a dose prescription of 50 Gy over 4 fractions (12.5 Gy per fraction).	1	2.1 cm	Yes	0.13 Gy	37.50 Gy
Spine IMRT, with a dose prescription of 45 Gy in 25 fractions (1.8 Gy per fraction).	2	1.85-2.1 cm	Yes	0.22 Gy	37.0 Gy

**Patient did not complete treatment course (21/25 fractions delivered). IMRT: intensity modulated radiation therapy; SBRT: stereotactic body radiation therapy; D95: dose covering 95% of the target volume; CTV: clinical target volume.*

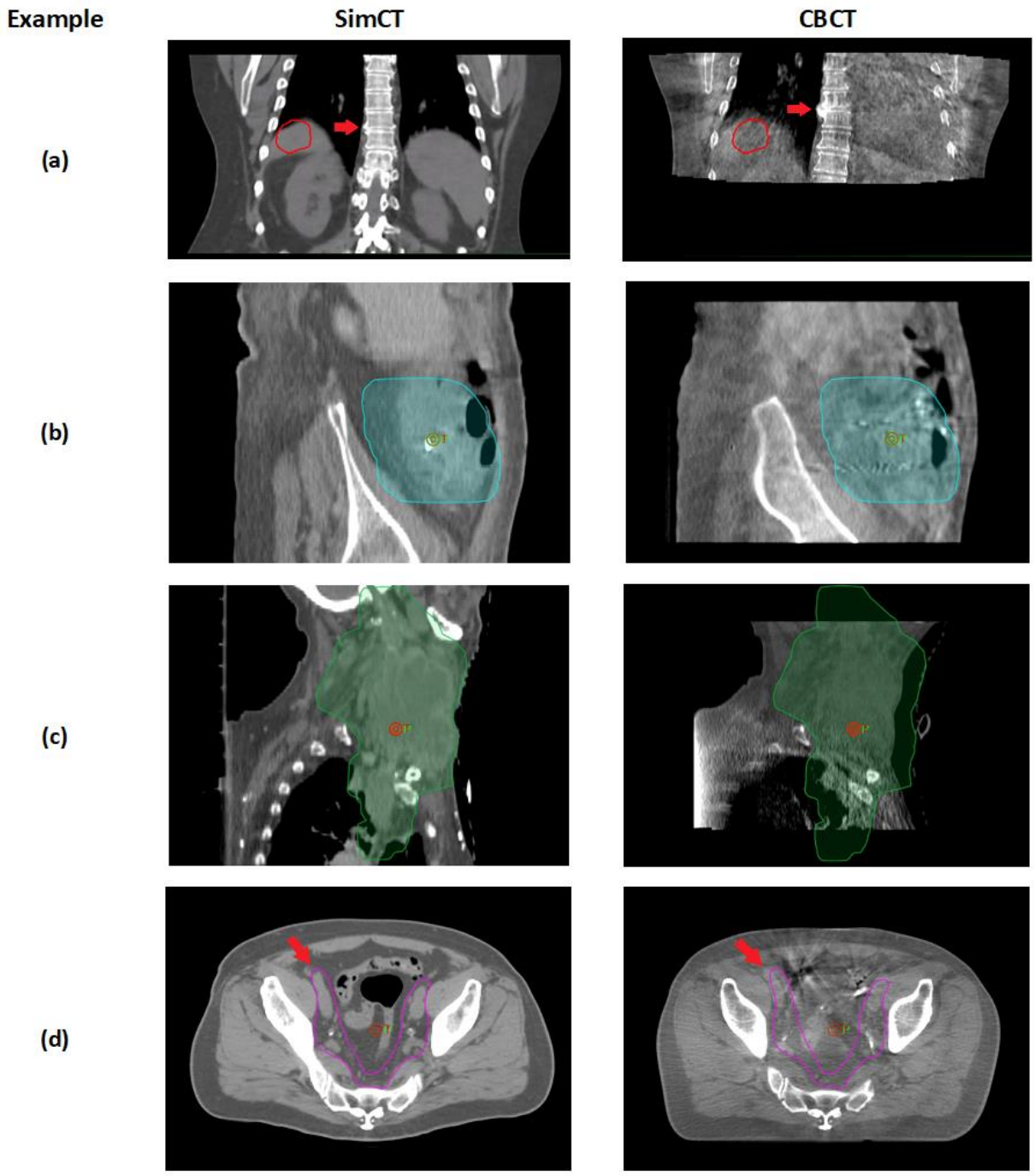


Figure 5.3: Select examples of false-positive cases, which show imperfections in the patient alignment but were judged to be clinically acceptable. The contour present on each image represent the planning tumor volume (PTV). (a) One of the three cases where the alignment was off by one vertebral body (highlighted by the red arrows) but the alignment at the PTV (red contour) was found to be adequate. The patient was undergoing a 5-fractions SBRT liver treatment with a prescription dose of 50 Gy (10 Gy/fraction) and similar observations were made

on three of the five fractions. (b) Considerable organ changes (bowel) resulted in shift in overall registration (see pelvic bone). However, the PTV coverage was judged to be acceptable. (c) Tumor shrinkage causing alignment to be imperfect and an increased dose to the lung. (d) Differences in bowel content and hip rotation causing some imperfections regarding the PTV coverage of the nodes (see red arrows).

As for the out-of-domain scans (i.e., extremity cases), it was found that they constituted less than 3% of the whole dataset analyzed in this study. From those cases, it was observed that lower extremity scans (including glutes, thigh, knee, and calf treatments), were generally sent to the PL model (75.7%, 22.1%, and 2.2% of the scans were sent to the PL, TA, and HN pipeline, respectively). For the upper extremity scans (including forearm, upper arm and shoulder), 37.5%, 42.5%, and 20% of the scans were sent to PL, TA and HN pipeline, respectively.

5.4 Discussion

In this work, a deep learning-based patient setup error detection algorithm (EDA) was used to aid in a bulk retrospective incident search for the CBCT-guided radiotherapy treatments performed between 2016 and 2017 at UCLA and between September 2021 and September 2022 at VCU. Initial model training and testing were performed on a separate dataset composed of simulated errors (true-positive) and clinically performed registrations (true-negatives). A receiver operating characteristic analysis was performed to obtain target thresholds which were used to flag cases for human review. A hard-thresholding method, as well as a model prediction trending analysis over individual patients' treatment courses, were used to identify cases for review.

Our results showed that CBCT-guided radiotherapy is indeed a very reliable and safe treatment modality, with an absolute gross patient misalignment error rate of 0.04% ± 0.02% at the two institutions. Of the 17,612 registrations analyzed by EDA, 1,362 cases were flagged and

investigated by human experts. Seven incidents were found during the case reviews. The three human-reported errors which occurred at the two institutions were detected during this study, validating the real-world performance of EDA in detecting gross patient setup misalignment incidents. Four additional misalignment errors found during the case reviews were previously unreported, which highlights the utility of automation in incident detection and learning within external beam radiation therapy.

Following a dosimetric analysis on the incidents, it was shown that the dosimetric impacts resulting from the patient misalignments were quite significant, with the dose deviations at the CTV ranging from 44% to 99% less than the prescribed doses per fraction. It can also be observed, from Figures 21b-21d, that the off-by-one vertebral body misalignment led to considerable dose to the heart which could have caused serious side effects to the patient. Those observations highlight the severe harm such patient misalignment may cause to the patient, and hence the need to minimize this failure mode within the external beam radiotherapy domain. It is also alarming that, in some of the cases, the error reached the patient even though two imaging modalities were used for patient alignment (see Appendix A.2). Additionally, for the SBRT incidents, the registrations were reviewed by at least two individuals (including the physician as per the institution's procedure), and the error still reached the patient.

The remaining flagged registrations, while often demonstrating some imperfections, were found to not be severe enough to be considered incidents. Those included registrations showing some patient rotation, tumor shrinkage, patient weight loss or substantial bowel/bladder differences, which often resulted in an imperfect overall alignment or potentially an increase in dose to adjacent organs-at-risk. In some cases, those images were flagged by the treating physician for follow-up remediation (e.g. improve bladder filling at the next fraction). There

were also instances where materials to boost surface dose (e.g. skin bolus) or shield organs-at-risk (e.g. clam-shell scrotal shield) were not included in the simCT images but were inserted during treatment, resulting in a mismatch with the CBCT which included such material. Additionally, cases containing substantial image artifacts and treatments of the extremities (e.g. leg, arm, and shoulder) were regularly flagged for review. Registrations involving palliative care patients were also occasionally flagged for review and often showed imperfections in the alignment, most probably due to difficulties in setting up the patient for treatment. However, all of those cases were judged to have reasonable PTV alignment as per the institutional practices (as defined by each expert reviewer, based on their own institution's set of practices and policies) and were labeled as false-positives.

Among those, there were also three cases (from the same 5-fraction liver treatment course) where the vertebral body alignment of the patient was off by one, but the alignment of the PTV was found to be reasonable. This could have represented a difference in breathing phase in the right lung between the planning CT scan and the treatment, with a collapsed left lung occurring between simulation and treatment as a contributory factor. One such example is shown and described in Figure 5.3.

One current limitation of EDA is that it uses select 2D slices from the whole 3D scans during the registration analysis. The choice to use 2D slices instead of the whole 3D scans such that the tool could be easily implemented on current computer systems in the clinic, without having a sizeable memory requirement. While a 3D model would have been able to capture more features, it is currently deemed impractical due to its memory requirements. However, with the rise in computation technologies and easier access to high end GPUs, the 3D EDA could be a more effective and practical approach in the future, as compared to the 2D EDA.

As seen during the image review, another limitation of EDA was its application to images with characteristics that were not well represented in the training dataset. Such cases include significantly-limited FOV CBCT scans (for example, when less than half of the thorax is present for a shoulder treatment) and extremity scans. During initial model training and testing, extremity images were not included in the datasets as too few of those cases were present for optimal model training. In addition to being scarce, their error modes were more complicated to be simulated as misalignments might only occur along the axis of the extremity site, such as the arm and leg. A larger training dataset containing more cases with extremities and additional error modes (such as combinations of rotations and translations) could benefit the EDA and reduce the false-positive rate, thereby improving the generalizability and robustness of the algorithm.

Furthermore, a heuristic model prediction-ratio approach was also applied during the trending analysis to target the 99% sensitivity threshold while minimizing excess false-positives. While this approach proved useful in the identification of errors during our retrospective study (2 of the 7 incidents detected), it may not be robust to all outliers. For example, it will fail in single-fraction treatment cases and in anomalous cases where the model prediction-ratio criteria are not met. Future studies will include ways to more robustly analyze the trends in the model predictions through AI-based clustering methods [89-90], which may help to identify deviations in image alignment when compared to patients with similar disease and treatment sites. However, the results obtained in this study demonstrate that the current trending analysis approach is acceptable and is complementary to the hard-thresholding method for patient setup error detection.

While this study highlights the high reliability and low patient setup incidence rate in CBCT-guided radiotherapy, it also exposes some safety gaps present within the current

workflow, with four of the seven incidents observed in this study going under the radar of both the safety and quality assurance checks at two adequately resourced radiotherapy centers. Additionally, the risk of similar unreported or undetected incidents may be higher in under-resourced radiotherapy centers where the lack of resources may translate to a decrease in safeguards. [91] Nevertheless, the results obtained from this study still emphasize the high reliability and safety of CBCT-guided radiotherapy and commend the current safety practices present throughout the workflow.

5.5 Conclusion

In external beam radiotherapy, gross patient setup errors are infrequent but are considered "never-events" due to their severe consequences. This study employed a deep learning-based patient setup error detection algorithm (EDA) to facilitate a comprehensive retrospective analysis of all cone-beam computed tomography (CBCT)-guided radiotherapy treatments administered at UCLA between 2016 and 2017 and at VCU between 2021 and 2022. Out of the 17,612 CBCT registrations (from 2,414 patients) assessed by EDA, 1,362 (from 631 patients) were flagged as potential incidents and subsequently reviewed. Following a thorough investigation of the flagged cases, seven incidents involving patient setup errors were identified. All of the three reported errors which occurred at the two institutions were found during this study, validating the real-world performance of EDA in detecting gross patient setup misalignments. The other four incidents observed during the case reviews were found to be previously unreported errors. While the results obtained highlight the reliability and safety of CBCT-guided radiotherapy (with an absolute gross patient misalignment incidence rate of 0.04% ± 0.02% at the two institutions), the

incidents which occurred also expose safety gaps still present within the patient alignment process.

Chapter 6: An unsupervised anomaly detection framework for CBCT setup images and registrations in image-guided radiotherapy using a variational auto-encoder

6.1 Introduction

External beam radiotherapy plays an important role in cancer treatment, for both curative and palliative intent. However, this treatment technique necessitates precise patient positioning and accurate delivery of therapeutic doses to the target tissues while minimizing exposure to surrounding healthy structures. Cone Beam Computed Tomography (CBCT) has emerged as a fundamental tool in modern radiotherapy systems, enabling three-dimensional imaging for precise treatment guidance and allowing 3D conformal and adaptive radiotherapy treatment for better tumor control while reducing side effects. [6, 92-94]

Yet, anomalies within CBCT-guided radiotherapy can potentially indicate deviations in treatment setup or delivery, posing a risk to treatment efficacy and patient safety. One such anomaly is the wrong registration of the CBCT with respect to the simulation Computer Tomography (simCT) leading to the wrong patient setup and hence a radiation dose mis-delivery. Other anomalies may include substantial tumor shape and size changes, soft-tissue variations, and suboptimal image quality which may cause deviations from the intended treatment and difficulty in the patient setup.

Traditionally, the detection of those anomalies in CBCT images has relied on manual review by trained professionals. However, this approach is labor-intensive, time-consuming, and

susceptible to human error, particularly in identifying subtle or infrequent anomalies. [95] Hence, there is a growing need to develop automated methods capable of efficiently and accurately detecting such anomalies.

Recently, a fully unsupervised deep learning technique using variational autoencoders (VAE) [110] has surfaced in the realm of anomaly detection [96-97]. The principal concept of VAEs is to compress the input data into a lower-dimensional latent space using an encoder and subsequently decode the latent space using a decoder to generate data that resembles the input. During the encoding part, the model would learn the underlying distribution of the input data, and subsequently use this distribution to generate the new data. When anomalies are rare, this approach assumes that the VAE will not be able to accurately reconstruct the anomalous data. Hence, the dissimilarity of the input data to the output data can be used as a measure for anomaly detection, where a large deviation would indicate an anomaly.

This VAE concept has been used in various anomaly detection studies, including anomalous network traffic detection [98], anomalous spatial and temporal signal detection [99], and videos anomaly detection [100]. In the medical imaging field, the VAE method was successfully used in a multitude of applications, including pathology detection and classification [102], disease diagnosis [103-104], anomalous anatomy segmentation [105-106], and organ segmentation quality assessment [101]. In the field of radiation therapy, Huang et al. used an autoencoder in conjunction with a clustering method to identify abnormal breast cancer radiotherapy treatment plans, with a sensitivity of 94.7% for a specificity of 69.0%. [107] However, the unsupervised VAE anomaly detection has yet to be applied to identify anomalies on the patient setup images, such as the CBCT, used during image-guided radiotherapy.

In this work, we propose an unsupervised anomaly detection framework (ADF) specifically designed for CBCT-guided radiotherapy. Our framework is grounded on the principle of leveraging VAEs to inpaint CBCT scans, exploiting the inherent degradation in inpainting accuracy when anomalous features are present. By quantifying the disparity between actual and inpainted CBCT images, our ADF generates anomaly scores indicative of potential treatment deviations.

The development and evaluation of our ADF are conducted using a comprehensive dataset comprising clinically registered simCTs and setup CBCTs obtained from a cohort of patients undergoing CBCT-guided radiotherapy treatment. Through systematic experimentation, including simulated misalignments and translational errors, we aim to demonstrate the efficacy and robustness of our approach in detecting anomalies in CBCT images.

By introducing automated anomaly detection into the realm of CBCT-guided radiotherapy, our framework holds promise in augmenting existing quality assurance protocols and enhancing patient safety. The subsequent sections of this paper will go into the methodology employed, the experimental setup, and the evaluation of our proposed ADF, followed by a discussion of the results and their implications for clinical practice.

6.2 Materials and Methods

6.2.1 The Anomaly Detection Framework (ADF)

Due to the memory requirements in processing 3D medical images, employing the traditional Variational Auto-encoder (VAE) for this task has been found to be impractical on

currently available hardware at our institution. During the experimental and development phase of the ADF, it has been found that for an input of size 128x128x64x2 pixel^4, the traditional VAE necessitates a latent space in the order of 10^5 for acceptable CBCT reconstruction, resulting in excessive GPU memory requirements during model training.

To mitigate this issue, an alternative inpainting technique using a VAE with skip connections was adopted, as shown in Figure 6.1. This image inpainting method has been previously proposed for anomaly detection tasks [108-109] and offers several advantages over the traditional VAE, including better reconstruction capability. Notably, skip connections enable more efficient information flow between layers, allowing a smaller latent space, and reducing the computational burden and memory requirements during training. By incorporating skip connections, the VAE can therefore effectively capture high-level features and nuances in the simCT-CBCT image pairs while maintaining reasonable memory usage.

This approach ensures that the ADF remains computationally feasible and scalable, making it suitable for this task. Additionally, the use of skip connections enhances the model's ability to detect anomalies by preserving important spatial information and improving inpainting accuracy, even in the presence of unusual image features.

The input for the ADF consisted of simCT-CBCT pairs, with two quadrants of the CBCT intentionally missing. This configuration simulates the scenario where a portion of the CBCT image is corrupted or unavailable, requiring the VAE model to inpaint the missing regions on the CBCT based on surrounding information or learned patterns. By inpainting the missing quadrants, the ADF reconstructs a complete CBCT image, enabling the detection of anomalies through discrepancies between the actual and inpainted CBCT scans.

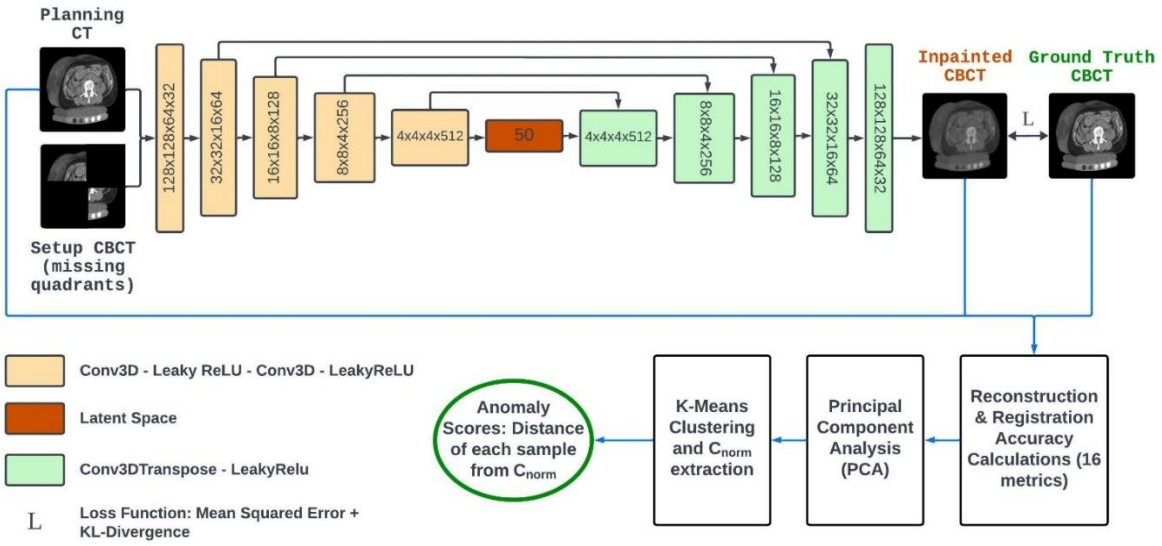


Figure 6.1: Depiction of the VAE-based anomaly detection framework (ADF). The blue lines represent processes that are performed during the test phase only.

6.2.2 Dataset Description

Under an IRB approved protocol (UID 18-001430), simCT and setup CBCT pairs were collected from 614 patients undergoing radiotherapy treatment at the University of California, Los Angeles Medical Center (UCLA) between January 2016 and December 2016. The dataset collection was performed using an in-house DICOM query and retrieval (DQR) application programming interface using the pynetdicom Python package. The treatments at UCLA had been performed on three TrueBeams and one NovalisTx linear accelerator treatment machine (Varian Medical Systems, California, United States). CBCT scans were acquired using the on-board imager of each machine. For each simCT-CBCT pair, the corresponding REG file and RTStruct file were also collected and used during the image pre-processing step in our

implementation.Using the unique patient identifiers, the dataset was partitioned into a training, validation and test set using a 70:10:20 split, and is further described in Table 6.1.

While anomalies may occur in CBCT-guidance, those are rare occurrences. Hence, to evaluate the ADF's ability in flagging anomalies, such cases had to be simulated. In CBCT-guided radiotherapy, patient setup misalignment is a type of anomaly which is hazardous and still occurring in clinics. [13-15, 95] Hence, for each patient in the test set, one translational alignment error was simulated by shifting the CBCT with respect to the CT by 20mm in a randomly chosen direction. Additionally, during this time frame, five known patient setup misalignment incidents, from three different treatment courses, occurred and were present in the dataset. The data from those patients were therefore kept in the test set in order to evaluate the ability of the proposed algorithm to identify real-world setup incidents, in addition to simulated incidents.

Table 6.1: Description of the patient dataset used in the development of the Anomaly Detection Framework (ADF).

	Number of Patients	**Number of clinically performed registrations**	**Number of known patient misalignment incidents**	**Number of simulated misalignments**	**Total number of registrations**
Training	442	3724	0	0	3724
Validation	58	581	0	0	581
Testing	114	884	5	114	1,003

6.2.3 Data Pre-Processing

To ensure consistency and compatibility across the dataset, all scans underwent resampling to a uniform voxel size of 1x1x1.5 mm^2 using a cubic spline interpolation method. Furthermore, due to limitations in GPU memory and computational resources, the original images were cropped around the treatment isocenter, resulting in volumes of 128x128x64 $pixel^3$. The planning tumor volume (PTV) contour for each case was also extracted from its respective RTStruct file, producing a binary mask image. This binary mask followed identical resampling and cropping as the simCT and CBCT images. The PTV contour will eventually be used (as described in Section 6.2.5) to obtain PTV-masked images and derive PTV-based image similarity metrics in order to detect discrepancies or anomalies at the PTV level.

6.2.4 Variational Autoencoder Model Training

The VAE, as depicted in Figure 6.1, was trained on a subset of the dataset consisting of paired CT-CBCT images from 442 patients. During training, the model aimed to minimize a combination of mean squared error (MSE) and Kullback–Leibler (KL) divergence, serving as the loss function. [111] The MSE is used to minimize the reconstruction error between the inpainted CBCT and ground truth CBCT, while KL divergence forces the distribution of the latent space towards a normal distribution. The loss function can be represented as in Equation 8 below.

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\Big(MSE(x_i, y_i) + \beta \times KL\big(P(z \mid \dot{x}_i) \mid Q(z \mid \dot{x}_i)\big)\Big) \qquad [8]$$

Where θ represents the parameters of the VAE, N is the number of training samples, x_i is the ground truth CBCT image, y_i is the reconstructed CBCT image, β is the weight for KL divergence (in this work a value of 0.1 was assigned as the β factor), $P(z \mid \dot{x}_i)$ is the posterior distribution of latent variables given input $\dot{x}_i$, and $Q(z \mid \dot{x}_i)$ is the distribution of latent variables predicted by the encoder.

6.2.5 Reconstruction & Registration Accuracy Measures

Following training, the VAE was applied to the unseen test dataset containing both normal and anomalous (patient setup misalignment) cases. Eight image similarity metrics, based on four commonly applied image similarity measures in the medical imaging field [112-115], were employed to quantify the accuracy of the inpainted CBCT scans (reconCBCT) compared to the ground truth CBCT scans (gtCBCT), both in terms of the reconstruction quality and registration quality with respect to the simCT. Those eight metrics were calculated on the whole images, as well as on the PTV-masked images, for a total of 16 image similarity outputs. The eight metrics employed are as follows:

1. Mean Squared Error (MSE): MSE measures the average squared difference between corresponding pixel intensities in the inpainted CBCT (reconCBCT) and the ground truth CBCT (gtCBCT). A lower MSE indicates higher similarity between the two images, suggesting better reconstruction accuracy.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \qquad [9]$$

2. Structural Similarity Index Measure (SSIM): SSIM quantifies the similarity of structural patterns between two images. It considers luminance, contrast, and structure similarity, providing a comprehensive assessment of image quality. A higher SSIM value indicates greater similarity between reconCBCT and gtCBCT.

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad [10]$$

3. Mutual Information (MI): MI measures the amount of information shared between two images. It captures both linear and nonlinear relationships between pixel intensities, indicating the degree of dependency between reconCBCT and gtCBCT. Higher MI suggests stronger correlation and better alignment between the images.

$$MI = \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right. \quad [11]$$

4. Gradient Magnitude Similarity (GMS): GMS evaluates the similarity of gradient magnitudes between reconCBCT and gtCBCT. It emphasizes edge information and structural details, crucial for medical image analysis. A higher GMS value signifies better preservation of edges and fine details in the reconstructed image.

$$GMS = \frac{2\sigma_x\sigma_y + C_3}{\sigma_x^2 + \sigma_y^2 + C_3} \quad [12]$$

5. Delta MSE (ΔMSE): ΔMSE calculates the absolute difference in MSE between the reconCBCT and gtCBCT, when compared to the simCT. It assesses the relative change in registration accuracy between the gtCBCT and reconCBCT, highlighting deviations from expected behavior.

$$\Delta\text{MSE} = \left\| MSE_{simCT-reconCBCT} - MSE_{simCT-gtCBCT} \right\| \quad [13]$$

6. Delta SSIM (ΔSSIM): ΔSSIM calculates the absolute difference in SSIM between the reconCBCT and gtCBCT, when compared to the simCT. It assesses the relative change in registration and structural accuracy between the gtCBCT and reconCBCT when compared to the simCT, highlighting structural dissimilarities.

$$\Delta\text{SSIM} = \left\| SSIM_{simCT-reconCBCT} - SSIM_{simCT-gtCBCT} \right\| \quad [14]$$

7. Delta MI (ΔMI): ΔMI measures the absolute difference in MI between PredCBCT and GTCBCT, and the corresponding simCT. It captures alterations in information content and dependency, providing insights into changes in image characteristics due to anomalies between the simCT and paired CBCT.

$$\Delta\text{MI} = \left\| MI_{simCT-reconCBCT} - MI_{simCT-gtCBCT} \right\| \quad [15]$$

8. Delta GMS (ΔGMS): ΔGMS calculates the absolute difference in GMS between reconCBCT and GTCBCT, and their corresponding gtCT image. It assesses variations in edge preservation and structural fidelity, indicating anomalies or irregularities when compared to the registered simCT.

$$\Delta GMS = \left\| GMS_{simCT-reconCBCT} - GMS_{simCT-gtCBCT} \right\| \quad [16]$$

Here, x_i represents the ground truth CBCT image, y_i represents the inpainted CBCT image. μ_x and μ_y are the mean values of x and y, σ_{xy} is the covariance of x and y, σ_x and σ_y are the standard deviations of x and y, and C_1, C_2, and C_3 are constants to stabilize the divisions.

6.2.6 Anomaly Score Calculation

After computing the image similarity metrics between the inpainted CBCT scans (reconCBCT) and the ground truth CBCT scans (gtCBCT) for all test instances, anomaly scores were derived to quantify the degree of deviation from normality in the CBCT images. For the first step, a Principal Component Analysis [116] was applied to reduce the dimensionality of the 16 similarity measures calculated. This transformation aims to capture the most significant variations in the data while minimizing information loss. By projecting the high-dimensional similarity measures onto a lower-dimensional space, PCA facilitates visualization and analysis of the data's underlying structure.

Following PCA, a K-means clustering algorithm [117], using the K-Means++ initializer [118], was employed to partition the data into two distinct clusters based on their similarity

patterns. K-means clustering aims to group the similarity measures into clusters such that the intra-cluster similarity is maximized while the inter-cluster similarity is minimized. This step helps identify clusters of data points with similar characteristics, allowing for the detection of anomalous instances or groups.

In practice, anomalous cases and errors are expected to be rare in the field of radiotherapy. [9, 95] Using this assumption, it is expected that most datapoints be near each other, leading to a denser cluster, with the few anomalous cases being further away from the center of that cluster. Hence, using the K-means clustering algorithm, the centroid (C_{norm}) of the denser cluster was extracted, which represents the central tendency of the majority of the data points. The anomaly score for each test case was subsequently calculated based on its euclidean distance from C_{norm}. This distance metric serves as a measure of how much the test case deviates from the typical or expected behavior observed in the majority of the dataset, with a higher distance indicating a greater divergence from normality, suggesting the presence of anomalies or irregularities in the simCT-CBCT image pair.

6.2.7 Performance Evaluation and Implementation Details

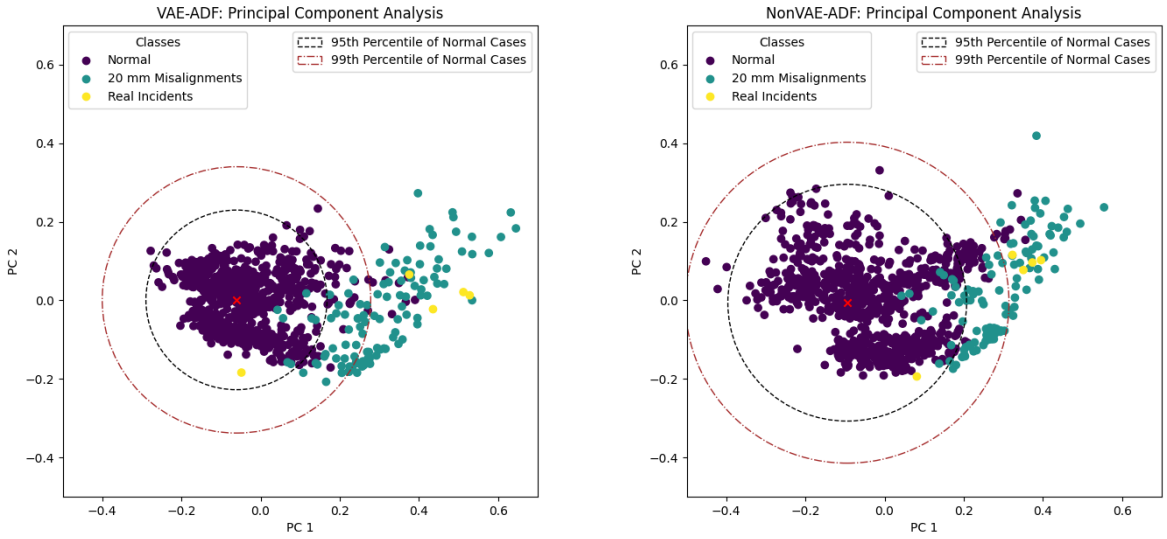
Using the anomaly scores calculated, a receiver operating characteristic (ROC) curve [47] was built to assess the performance of the ADF in classifying the normal and anomalous (error) cases from the test dataset. For comparative measures, the metric and anomaly score calculations were repeated using the simCT and gtCBCT only, excluding the use of the VAE (NonVAE-ADF). This would act as a baseline to understand the utility of the VAE in the ADF. The area-under-the ROC curves were subsequently calculated and the performance of each algorithm on the test set was compared.

All experiments and analyses were implemented using the Python programming language. The Tensorflow 2.13 framework was utilized for VAE implementation, while scikit-learn[†], a machine learning library, was employed for the PCA and k-means clustering algorithms. The VAE was trained using an Adam Optimizer [45] with a starting learning rate of 10^{-5}. During training, the VAE was evaluated after each epoch using the validation set, and the learning rate was reduced by a factor of 0.5 if the validation loss did not improve for 10 consecutive epochs. The VAE was trained until the validation loss did not improve for 50 consecutive epochs, or for a maximum of 200 epochs. The model achieving the highest validation accuracy was then saved. The experiment was performed on a workstation comprising of four NVIDIA GeForce GTX Titan X (NVIDIA Corp, Santa Clara, USA) with 12GB VRAM in each (total 48GB VRAM), and 62GB RAM.

6.3 Results

The scatter plots in Figure 6.2 below show the results obtained following the PCA on the outputs of the VAE-based ADF and the NonVAE-based ADF. For each plot, the 95^{th} and 99^{th} percentile distances of the normal cases from C_{norm} were calculated and reported.

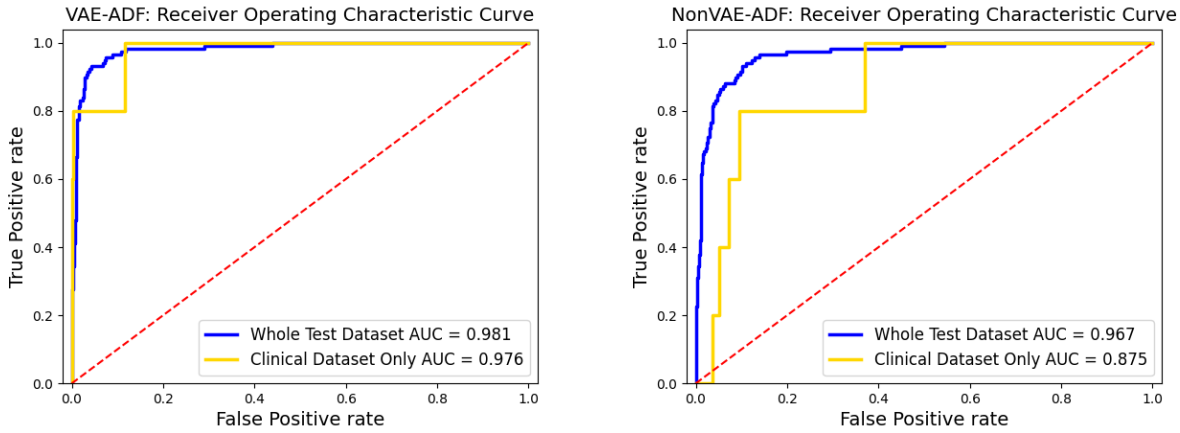
[†] https://scikit-learn.org/stable/



(a) (b)

Figure 6.2: Scatter plots obtained following a principal component analysis (PCA) on the similarity measure calculations between the output CBCTs and ground truth CBCTs (or simCT) in the test dataset for (a) the VAE-based ADF, and (b) the nonVAE-based ADF. The red cross on each plot represents the centroid, C_{norm}, of the denser cluster found using a K-Means clustering algorithm.

Using the anomaly scores obtained from both the VAE-based ADF and the nonVAE-based ADF, receiver operating characteristic curves were produced and the area under each ROC curve was calculated and reported in Figure 6.3. For each algorithm, two ROC curves were produced: one on the whole test dataset (including simulated errors and real incidents), and another one on clinically performed registrations only (i.e., excluding simulated errors, but including real incidents). The second dataset aimed at mimicking a real-world dataset where very few cases contain anomalies or errors. This allowed the evaluation of the algorithms' abilities at catching those real-world incidents, while minimizing false-positive cases.



(a) (b)

Figure 6.3: Receiver Operating Characteristic (ROC) curves obtained from the anomaly scores calculated using (a) the VAE-based ADF, and (b) the NonVAE-based ADF. The blue curves show the results for the whole test dataset (including simulated errors and real incidents) and the yellow lines show the results for clinically performed registrations only (including real incidents but excluding simulated errors).

The principal goal of the ADF is to be able to localize the anomalous cases (in this case, translational alignment errors), while minimizing false-positives. During the development of our tool, a large focus was placed on the ADF's ability to catch off-by-one vertebral-body misalignments with a threshold value that leads to less than 5% of false positives, which can be deemed acceptable in comparison to other false-positive interrupts and interlocks in the clinical workflow. For a fixed sensitivity of 95%, the specificities of the VAE-based ADF and the NonVAE-based ADF were found to be 92.6% and 87.2% respectively on the whole test dataset. When applied to only clinically performed registrations (i.e., excluding simulated errors), the VAE-based ADF identified all clinical incidents with a specificity of 88.3%, while the NonVAE-based ADF had a specificity of 62.9% for similar sensitivity. Hence, the results obtained

showcase the utility of the VAE in the ADF for anomaly detection in CBCT-guided radiotherapy.

Upon review of the true-negative (normal) cases obtaining an anomaly score in the 95th percentile range, it was observed that they often showed irregularities such as bowel/bladder content and size differences between simulation and treatment day, patient rotations, and substantial image artifacts (mostly due to photon starvation and breathing motion).

6.4 Discussion

In image-guided radiotherapy, Cone Beam Computed Tomography (CBCT) offers valuable information regarding treatment progress. This includes monitoring patient anatomy changes, tumor size changes, and patient setup errors, which may all affect the intended treatment doses and may require the treatment plan to be adapted. However, manually reviewing the setup CBCT scans is time-consuming, and mental fatigue can result in those anomalies being overlooked. The unsupervised anomaly detection framework (ADF) proposed in this study leverages a variational autoencoder (VAE) to help automatically localize anomalous cases in cone beam computed tomography (CBCT)-guided radiotherapy. The principal goal of the ADF is to identify and localize clinically relevant anomalies, including patient setup uncertainties and errors, which is a critical aspect in ensuring accurate radiotherapy treatment delivery.

The evaluation of the ADF's performance revealed promising results in the detection of translational alignment errors. For a fixed sensitivity of 95%, the VAE-based ADF demonstrated a specificity of 92.6%, outperforming the NonVAE-based ADF, which achieved a specificity of 87.2% on the entire test dataset. Moreover, when restricted to clinically performed registrations, excluding simulated errors, the VAE-based ADF identified all clinical incidents with a

specificity of 88.3%, while the NonVAE-based ADF exhibited a lower specificity of 62.9% for similar sensitivity.

These findings underscore the utility of the VAE-based approach in anomaly detection for CBCT-guided radiotherapy. By effectively capturing subtle deviations and anomalies in the CBCT scans, the VAE-based ADF demonstrates superior performance in localizing clinically relevant incidents while minimizing false positives. It can be seen from Figure 6.4 that the differences between the ground truth CBCT and the reconstructed CBCT is more considerable in the case where a translational shift is applied, as compared to a properly aligned case. This highlights the potential of VAEs as powerful tools in enhancing the accuracy and efficiency of anomaly detection in radiotherapy image guidance.

Despite the promising results, several limitations exist in this work. Firstly, the performance of the ADF may vary depending on the specific characteristics of the patient population and treatment protocols. Additionally, the reliance on simulated translational alignment errors for evaluation does not fully capture the complexity and diversity of real-world clinical scenarios. Further validation studies involving larger and more diverse datasets are warranted to assess the generalizability and robustness of the proposed ADF. Moreover, the computational complexity and resource requirements associated with VAE-based anomaly detection may pose challenges for real-time clinical implementation. Future research efforts should focus on optimizing the computational efficiency and scalability of the ADF to facilitate seamless integration into clinical practice.

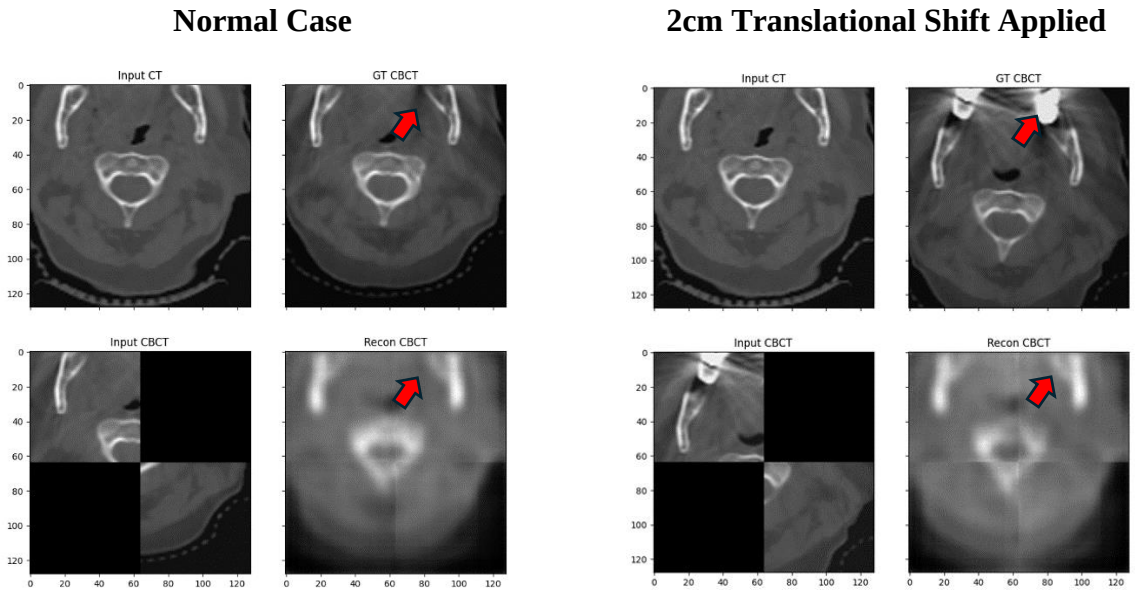


Figure 6.4: Illustration of the inputs, ground truth, and output of the VAE when applied to (a) a non-anomalous case and (b) a simulated anomalous case. The two cases shown above involve the same patient. The red arrows highlight one of the areas where a major difference is seen between the normal and anomalous case.

It is also noteworthy to acknowledge an aspect of our methodology regarding the usage of the latent space of the VAE. Despite its potential as a feature space for anomaly detection, the information present in the latent space was not utilized in the anomaly score derivation. This decision stemmed from our observation that the information contained within the latent space did not contribute significantly to distinguishing between normal and anomalous cases.

This observation may be attributed to the inclusion of skip-connections in our VAE architecture. While skip-connections are beneficial for preserving spatial information and enhancing the inpainting process, they might have limited the discriminative power of the latent

space for anomaly detection. As a result, alternative network architectures that leverage the latent space more effectively could potentially yield improved results.

By exploring alternative architectures or modifying the existing VAE architecture, future research endeavors could unlock the latent space's full potential for anomaly detection in CBCT-guided radiotherapy. Leveraging the latent space through innovative network designs or incorporating additional constraints could enhance its ability to capture subtle deviations indicative of anomalies. This avenue holds promise for improving the sensitivity and specificity of anomaly detection algorithms, ultimately advancing the accuracy and reliability of the ADF.

Despite the challenges and limitations encountered, this study represents a significant step forward in the development of anomaly detection frameworks for CBCT-guided radiotherapy. By harnessing the capabilities of VAEs and innovative inpainting techniques, the potential of unsupervised methods in localizing critical translational alignment errors while minimizing false positives have been demonstrated. The novel anomaly detection framework presented in this study offers a way to automatically identify the patient setup CBCT scans and treatment fractions which are deviating from normality and may require the attention of the physician and/or physicist. We believe that the tool can add value to physicians' routine image reviews and physics chart checks by flagging a select fraction of the most anomalous cases for review. This may not only improve the efficiency of those repetitive and time-consuming processes but can also lead to a positive impact on patient safety and treatment outcomes.

6.5 Conclusion

Anomalies in cone beam computed tomography (CBCT) radiotherapy image guidance can be indicative of treatment deviations. Repetitive manual review of routine images is inefficient and inaccurate at identifying such rare events. By leveraging a variational autoencoder (VAE)-based unsupervised anomaly detection framework (ADF), rare treatment deviations were identified with high accuracy. When validated on a comprehensive dataset of clinically registered simulation CTs and setup CBCTs, including real and simulated errors, ADF demonstrated robust performance at detecting the error cases, achieving an area-under-the-ROC curve of 98.1%. The findings underscore the potential of this approach to enhance the quality and safety of CBCT-guided radiotherapy by accurately identifying both real and simulated patient setup misalignments.

Chapter 7: Conclusions and Future Work

7.1 Summary of work

The goal of specific aim 1 was to develop a fully-automatic deep learning based gross setup error detection algorithm for CBCT-guided radiotherapy. In **Chapter 2**, An error detection algorithm (EDA), composed of 3 distinct densely connected convolutional neural networks (CNNs), was developed using a supervised framework to analyze the registration between the simulation computed tomography (simCT) scan and the setup CBCT scan. As the type of misalignment could vary by treatment site, based in part on the landmarks used during the registration process, each of the 3 CNNs in the EDA was designed to handle a specific body site and its corresponding error type. For the thoracic and abdominal region (TA), the EDA was trained to localize off-by-one vertebral body misalignments. For the head & neck (HN) and pelvis (PL) regions, the EDA was trained on 10 mm translations in randomly chosen directions. The EDA achieved high accuracy in detecting the simulated errors, with areas under the receiver operating characteristic (ROC) curve of 99.6%, 99.4%, and 99.2% for the HN, PL, and TA regions respectively. For a fixed specificity of 99.0%, the sensitivity was 99.0%, 99.4%, and 89.0% for the HN, TA, and PL models respectively.

To fully automate the EDA pipeline, it was essential to identify the treatment region from each incoming simCT-CBCT pair such that it could be sent to the corresponding model. However, there is no robust indicator within the simCT or CBCT dicom headers to identify the treatment region. **Chapter 3** describes a deep learning-based Anatomical Region Labeling (ARL) model that was trained to recognize the treatment region (HN, TA, PL, and extremities)

from a single coronal CBCT slice. During the validation phase the ARL model achieved an overall accuracy of 99.2% in classifying the CBCTs into the 4 distinct regions.

In **Chapter 4**, the EDA (including the ARL) was implemented to interact with the clinical database in a fully-automated fashion through DICOM networking protocol, such that it could analyze incoming simCT-CBCT registrations on a nightly basis. Over a 45-day period, 1357 pre-treatment CBCT registrations from 197 patients were retrieved and analyzed by EDA. The predictions of the EDA were then validated against independent expert observers. Following an ROC analysis, a global threshold for model predictions of 0.87 was determined, with a sensitivity of 100.0% and specificity of 82.0%. This demonstrated that a prediction threshold could be identified to stratify pre-treatment images with a statistically significant correlation to the observer scores. In addition to the validation study, Chapter 4 provided anecdotal examples of how the EDA could be beneficial to the clinical workflow through quantification of daily setup alignments and patient/plan specific timelines to identify trends and flag anomalies.

To address specific aim 2, the fully-automated EDA pipeline was applied for a bulk patient setup error search on clinical CBCT-guided radiotherapy image databases. As described in **Chapter 5**, the EDA was used to analyze all clinically performed simCT-CBCT registrations between 2016 and 2017 at the University of California, Los Angeles Medical Center (UCLA) and between 2021 and 2022 at the Virginia Commonwealth University Medical Center (VCU). A total of 17,612 registrations were analyzed by the EDA, resulting in 7.7% flagged events based on pre-defined thresholds. Three previously reported errors were successfully flagged by the EDA, and 4 previously unreported vertebral body misalignment errors were discovered during case reviews. Those results validated the clinical utility of the EDA for bulk image reviews and

highlighted the reliability and safety of CBCT-guided radiotherapy, with an absolute gross patient misalignment error rate of 0.04% ± 0.02% per delivered fraction.

The goal of specific aim 3 was to develop a 3D and fully-unsupervised error detection framework for anomaly detection in CBCT-guided radiotherapy. In **Chapter 6**, a variational autoencoder (VAE)-based anomaly detection framework (ADF) was proposed. The ADF was developed to output an anomaly score which would be highest for images containing infrequently observed features. When validated on a comprehensive dataset of clinically registered simulation CTs and setup CBCTs, including real and simulated errors, the ADF demonstrated robust performance at detecting the error cases, achieving an area-under-the-ROC curve of 98.1%. When applied to only clinically performed registrations (i.e., excluding simulated errors), the ADF identified all clinical incidents with a specificity of 88.3%. Those results demonstrate the feasibility of a VAE-based anomaly detection framework to detect patient setup anomalies and thus enhance patient safety and treatment outcomes in CBCT-guided radiotherapy.

7.2 Future Directions

The EDA applied in specific aim 1 and specific aim 2 have been developed using 2D convolutional neural networks, which take as input select 2D orthogonal slices from the simCT and CBCT images. Although the 2D images lead to faster computation time, the amount of features captured by the model is limited to the selected slices. A 3D model could capture many more useful features from the entire scans, which could further improve the detection of misalignment errors in CBCT-guided radiotherapy.

Additionally, the EDA was trained on a few error types, which may not fully capture the complexity of cases and errors which may occur in the clinical setting. A larger and more comprehensive patient dataset, containing more error modes (such as wrong patient registrations or rotational errors), could benefit the EDA by making it more robust to patient setup errors. Additionally, in this work, the EDA training was performed on data from only two institutions, which may not be enough to capture the variability in scanning protocol, image quality, and registration techniques across all treatment facilities and treatment machines. Further work in this direction should include a determination of a minimum diversity of cross-institutional data that would lead to an expectation of similar model performance on data from an unseen institution. This would aid in developing a tool which is generalizable to many more institutions and could benefit a wider range of facilities.

While the EDA was applied and validated through retrospective studies in this work, future research endeavors should prioritize prospective validation studies and real-time implementation to assess the EDA's performance in clinical practice accurately. Additionally, considerations should be given to the computational resources and infrastructure required for deploying these algorithms in real-world clinical settings, ensuring their feasibility and scalability for widespread adoption. The aim would be to leverage the speed of deep learning inference to move the EDA from retrospective to real-time, integrating directly with the treatment machine to interlock the beam if the AI-model flags a potential setup misalignment. Such a system would ideally require the operator to either revise the alignment or acknowledge the interlock before proceeding with treatment.

The development of a variational autoencoder (VAE)-based anomaly detection framework (ADF) represents a promising step towards enhancing patient safety in CBCT-guided

radiotherapy. Moving forward, efforts should focus on further optimizing the ADF's performance and scalability. Exploring alternative network architectures, incorporating additional imaging modalities or patient-specific data, and refining the anomaly scoring mechanism could enhance the ADF's ability to detect and classify anomalies with greater precision and reliability. Additionally, conducting prospective clinical validation studies to assess the ADF's performance in real-world settings (e.g. during physics quality assurance checks) would provide valuable insights into its clinical utility and impact on treatment outcomes. Such automated AI-assisted tool could allow for independent, quantitative review of every setup CBCT and registration in the future.

Appendix

A.1 UNet-based Spinal Canal Segmentation in the Error Detection Algorithm (EDA)

For the thoracic and abdominal radiotherapy treatments, it is a common practice to contour the spinal canal as an organ at risk. However, there are a few cases where only part or none of the spinal canal is contoured within the CT volume. During the orthogonal image extraction discussed in Section 2.2, the error detection algorithm relies heavily on the presence of the cord contour on the selected axial slice to obtain the vertebral body position which is used to get the sagittal and coronal images. In the absence of the canal contour, the algorithm would fail in extracting the correct slices, leading to an algorithm failure. Hence, the authors decided to implement a 2D UNet-based spinal canal segmentation (SCS) algorithm that could segment the canal from the selected axial image of the planning CT and avoid the error detection algorithm from failing.

The SCS model was based on the UNet architecture [119], which is composed of a contracting path that captures contextual features from the input, and an expanding path that extract the features obtained. This model was trained and tested using 184 patients' planning CT from institution #1. The patient dataset was split into a training, validation, and test set, as shown in Table B1 below. This dataset split was kept consistent to the one performed during the EDM experiment to avoid training SCS on images that would be used during the validation or testing phase of the EDM. The input to the model was a 150x150 axial image patch automatically extracted about the center of patient body. The binary mask of the spinal canal was obtained from the RT Structure file of each CT dataset and used as ground truth labels during model training and testing.

Table A.1: Description of the dataset used to train, validate, and test the SCS model.

	Number of Patients	Number of Axial Slices
Training Set	147	22,884
Validation Set	15	2,226
Testing Set	22	2,914

The SCS model was implemented using Tensorflow 2.2 with Keras backend. The binary cross-entropy loss function was used during training. The model was trained using Adam Optimizer [45] with a starting learning rate of $5x10^{-4}$. During training, the model was evaluated after each epoch using its validation set, and the learning rate was reduced by a factor of 0.8 if the validation loss did not improve for 5 consecutive epochs. The model was trained until the validation loss did not improve for 20 consecutive epochs, or for a maximum of 200 epochs. The model achieving the highest validation accuracy was then saved. SCM achieved convergence after 5 epochs.

To test the performance of the model, the distance between the centroid of the ground truth contour and the centroid of the predicted contour was calculated for each of the 2, 914 test images. The average and the standard deviation of the calculated distances are reported in Table B2 below. The number of predictions that led to a centroid separation of more than 1 cm was also calculated.

Table A.2: Results of the centroid comparisons between the ground truth contours and the predicted contours.

Average Separation (mm)	1.51
Standard Deviation (mm)	9.49
# images with a separation > 10 mm	61 (2.1%)

The number of slices where the ground truth centroid was found within the region predicted by the SCS model was calculated. Our results show that for 97.4% of the test images, the ground truth centroid was found within the predicted contour. From the results obtained, the SCS was deemed to produce acceptable results such that it can be incorporated in the error detection algorithm as a secondary and independent method of determining the position of the vertebral body for orthogonal slice extraction.

A.2 Case Presentations of Patient Misalignment Incidents Found During the Retrospective Patient Error Search

Case (a): Previously-unknown incident – Abdomen IMRT

The patient was undergoing an intensity modulated radiation therapy (IMRT) treatment to the left abdomen and was prescribed a total dose of 87.5 Gy over 25 fractions (3.5 Gy/fraction). For patient alignment purposes, daily ExacTrac (Brainlab, Munich, Germany) and CBCT imaging were performed, with the bone and target used as registration landmarks for each imaging modality, respectively.

The patient setup incident occurred on the 7th treatment fraction. Upon review of the case, it was found that the physician became aware of the issue on the following day and created a 'Patient Alert' stating "Pt not aligned properly. Please call me to machine today to check setup." However, the incident was reported to neither RO-ILS nor to the in-house incident reporting system, and no adjustment was made to the treatment course.

To determine the dosimetric effect of this previously unknown incident, we performed a cumulative dose-volume histogram (DVH) comparison for the delivered treatment against the planned treatment (no incident), as shown in Figure 3. At the D95 (dose covering 95% of the target volume) a difference of 1.0 Gy was observed between the delivered dose to the CTV (86.5 Gy) and the prescribed dose to the CTV (87.5 Gy). The small deviation (1.1% at D95) suggested that a treatment adjustment might not have been needed based on the conventional practice at the institutions. However, this case remained a sub-optimal treatment and also demonstrated a failure mode in the current incident prevention and incident reporting workflow.

Case (b), (c), (d): Previously-unknown incidents – Stomach IMRT

The patient was undergoing an IMRT treatment to the stomach and was prescribed a total dose of 45 Gy over 25 fractions (1.8 Gy/fraction). CBCT imaging was ordered on a weekly basis, and kV-MV imaging was ordered on a daily basis.

The misalignment incidents occurred on the 1st, 6th and 7th fractions. Those incidents were previously unknown to the institution, and no adjustment was made to the treatment course. It is also important to note that in this case, the patient did not complete the treatment course (21 of 25 fractions completed). Based on the dose analysis performed, the CTV was under-dosed by approximately 44.4% on each misaligned fraction and by approximately 2.7% over the treatment course (21 of 25 fractions). It is also noteworthy that the misalignment also resulted in a higher dose to the heart of the patient.

Case (e): Known incident – Lung SBRT

The patient was undergoing a lung stereotactic body radiation therapy (SBRT) treatment and was prescribed a dose of 50 Gy in 4 fractions (12.5 Gy/fraction) to a lower left lobe lesion. For patient alignment purposes, daily ExacTrac and CBCT imaging were performed, with the bone and target used as registration landmarks for each imaging modality, respectively.

The incident occurred on the second fraction when the patient was misaligned by about 3 cm superior to the PTV. The incident was discovered at the end of the treatment course during a physics chart check and was reported to both RO-ILS and the in-house incident learning system. Following a treatment assessment by the physicists and physician, a fifth fraction of 12.5Gy was delivered to the patient to make up for the positioning error.

Cases (f) and (g): Known incidents – Spine IMRT

The patient was undergoing a spine IMRT treatment and was prescribed a total dose of 45 Gy in 25 fractions (1.8 Gy/fraction) to a para-spinal mass. For patient alignment purposes, daily ExacTrac and CBCT imaging were performed, with the bone and target used as registration landmarks for each imaging modality, respectively.

Two incidents occurred during that same treatment course; on the 3rd and 8th fractions. The incidents were discovered by the physician during image reviews following the 8th fraction and both incidents were reported to RO-ILS and the in-house incident learning system. To correct for the under-dosed region (inferior region of the target), an additional fraction of 1.8 Gy was administered to the patient as a BID (twice-a-day) treatment on the day of the 19th fraction.

References

1. Wouters, B. G. (2018). Cell death after irradiation: how, when and why cells die. In Basic clinical radiobiology (pp. 21-31). CRC Press.

2. World Health Organization. (2021). Technical specifications of radiotherapy equipment for cancer treatment.

3. Bryant, A. K., Banegas, M. P., Martinez, M. E., Mell, L. K., & Murphy, J. D. (2017). Trends in radiation therapy among cancer survivors in the United States, 2000–2030. Cancer Epidemiology, Biomarkers & Prevention, 26(6), 963-970.

4. Folkert, M. R., & Timmerman, R. D. (2017). Stereotactic ablative body radiosurgery (SABR) or Stereotactic body radiation therapy (SBRT). Advanced drug delivery reviews, 109, 3-14.

5. Nieder, C., Grosu, A. L., & Gaspar, L. E. (2014). Stereotactic radiosurgery (SRS) for brain metastases: a systematic review. Radiation Oncology, 9, 1-9.

6. Maund, I. F., Benson, R. J., Fairfoul, J., Cook, J., Huddart, R., & Poynter, A. (2015). Image-guided radiotherapy of the prostate using daily CBCT: the feasibility and likely benefit of implementing a margin reduction. The British journal of radiology, 87(1044), 20140459.

7. Reggiori, G., Mancosu, P., Tozzi, A., Cantone, M. C., Castiglioni, S., Lattuada, P., ... & Scorsetti, M. (2011). Cone beam CT pre-and post-daily treatment for assessing geometrical and dosimetric intrafraction variability during radiotherapy of prostate cancer. Journal of Applied Clinical Medical Physics, 12(1), 141-152.

8. Thengumpallil, S., Smith, K., Monnin, P., Bourhis, J., Bochud, F., & Moeckli, R. (2016). Difference in performance between 3D and 4D CBCT for lung imaging: a dose and image quality analysis. Journal of applied clinical medical physics, 17(6), 97-106.

9. ASTRO (2023). RO-ILS Aggregate Report Q1 2023. Available at: https://www.astro.org/Patient-Care-and-Research/Patient-Safety/RO-ILS/RO-ILS-Education. Accessed June 20, 2023.

10. Doroszczuk, B., Bardet, M. C., Covard, F., & Javay, O. (2019). ASN Report on the state of nuclear safety and radiation protection in France in 2018+ Abstracts.

11. Smith, S., Wallis, A., King, O., Moretti, D., Vial, P., Shafiq, J., ... & Delaney, G. P. (2020). Quality management in radiation therapy: A 15 year review of incident reporting in two integrated cancer centres. Technical innovations & patient support in radiation oncology, 14, 15-20.

12. Huq, M. S., Fraass, B. A., Dunscombe, P. B., Gibbons Jr, J. P., Ibbott, G. S., Mundt, A. J., ... & Yorke, E. D. (2016). The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management. Medical physics, 43(7), 4209-4262.

13. Hoopes, D. J., Ford, E. C., Ezzell, G. A., Dicker, A. P., Chera, B. S., & Potters, L. (2017). Incident learning for stereotactic radiation therapy from RO-ILS: Radiation Oncology Incident Learning System. International Journal of Radiation Oncology, Biology, Physics, 99(2), S46-S47.

14. McGurk, R., Naheedy, K. W., Kosak, T., Hobbs, A., Mullins, B. T., Paradis, K. C., ... & Mazur, L. M. (2023). Multi-institutional stereotactic body radiation therapy incident learning: evaluation of safety barriers using a human factors analysis and classification system. Journal of Patient Safety, 19(1), e18-e24.

15. Ezzell, G., Chera, B., Dicker, A., Ford, E., Potters, L., Santanam, L., & Weintraub, S. (2018). Common error pathways seen in the RO-ILS data that demonstrate opportunities for improving treatment safety. Practical radiation oncology, 8(2), 123-132.

16. Shah, M., Halalmeh, D. R., Sandio, A., Tubbs, R. S., & Moisi, M. D. (2020). Anatomical variations that can lead to spine surgery at the wrong level: part II thoracic spine. Cureus, 12(6).

17. ASN NEWSLETTER FOR RADIOTHERAPY PROFESSIONNALS. (2018) Patient safety - Paving the way for progress: #12 Patient repositioning imaging: vertebra identification error. https://www.french-nuclear-safety.fr/Media/Files/00-Publications/Patient-safety-12.-Patient-repositioning-imaging-vertebra-identification-error

18. Luh, J. Y., Albuquerque, K. V., Cheng, C., Ermoian, R. P., Nabavizadeh, N., Parsai, H., ... & Hartford, A. (2020). ACR–ASTRO Practice Parameter for Image-guided Radiation Therapy (IGRT). American journal of clinical oncology, 43(7), 459-468.

19. Ford, E., Conroy, L., Dong, L., de Los Santos, L. F., Greener, A., Gwe-Ya Kim, G., ... & Wells, M. (2020). Strategies for effective physics plan and chart review in radiation therapy: Report of AAPM Task Group 275. Medical physics, 47(6), e236-e272.

20. Xia, P., Sintay, B. J., Colussi, V. C., Chuang, C., Lo, Y. C., Schofield, D., ... & Zhou, S. (2021). Medical Physics Practice Guideline (MPPG) 11. a: Plan and chart review in external beam radiotherapy and brachytherapy. Journal of applied clinical medical physics, 22(9), 4-19.

21. Pallotta, S., Marrazzo, L., Ceroti, M., Silli, P., & Bucciolini, M. (2012). A phantom evaluation of Sentinel™, a commercial laser/camera surface imaging system for patient setup verification in radiotherapy. Medical physics, 39(2), 706-712.

22. Pallotta, S., Simontacchi, G., Marrazzo, L., Ceroti, M., Paiar, F., Biti, G., & Bucciolini, M. (2013). Accuracy of a 3D laser/camera surface imaging system for setup verification of the pelvic and thoracic regions in radiotherapy treatments. Medical Physics, 40(1), 011710.

23. Schöffel, P. J., Harms, W., Sroka-Perez, G., Schlegel, W., & Karger, C. P. (2007). Accuracy of a commercial optical 3D surface imaging system for realignment of patients for radiotherapy of the thorax. Physics in medicine & biology, 52(13), 3949.

24. Lamb, J. M., Agazaryan, N., & Low, D. A. (2013). Automated patient identification and localization error detection using 2-dimensional to 3-dimensional registration of kilovoltage x-ray setup images. International Journal of Radiation Oncology* Biology* Physics, 87(2), 390-393.

25. Jani, S. S., Low, D. A., & Lamb, J. M. (2015). Automatic detection of patient identification and positioning errors in radiation therapy treatment using 3-dimensional setup images. Practical Radiation Oncology, 5(5), 304-311.

26. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature medicine, 25(1), 24-29.

27. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

28. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet Classification with Deep Convolutional Neural Networks. Paper presented at the Advances in neural information processing systems.

29. Razzak, M. I., Naz, S., & Zaib, A. (2018). Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics.

30. Cai, L., Gao, J., & Zhao, D. (2020). A review of the application of deep learning in medical image classification and segmentation. Annals of translational medicine, 8(11).

31. Suzuki, K. (2017). Overview of deep learning in medical imaging. Radiological physics and technology, 10(3), 257-273.

32. Sahiner, B., Pezeshk, A., Hadjiiski, L. M., Wang, X., Drukker, K., Cha, K. H., ... & Giger, M. L. (2019). Deep learning in medical imaging and radiation therapy. Medical physics, 46(1), e1-e36.

33. Petragallo, R., Bertram, P., Halvorsen, P., Iftimia, I., Low, D. A., Morin, O., ... & Lamb, J. M. (2023). Development and multi-institutional validation of a convolutional neural network to detect vertebral body mis-alignments in 2D x-ray setup images. Medical physics, 50(5), 2662-2671.

34. Charters, J. A., Luximon, D., Petragallo, R., Neylon, J., Low, D. A., & Lamb, J. M. (2024). Automated detection of vertebral body misalignments in orthogonal kV and MV guided radiotherapy: application to a comprehensive retrospective dataset. Biomedical Physics & Engineering Express.

35. Zelefsky, M. J., Kollmeier, M., Cox, B., Fidaleo, A., Sperling, D., Pei, X., ... & Hunt, M. (2012). Improved clinical outcomes with high-dose image guided radiotherapy compared with non-IGRT for the treatment of clinically localized prostate cancer. International Journal of Radiation Oncology* Biology* Physics, 84(1), 125-129.

36. Margalit, D. N., Chen, Y. H., Catalano, P. J., Heckman, K., Vivenzio, T., Nissen, K., ... & Ng, A. K. (2011). Technological advancements and error rates in radiation therapy delivery. International Journal of Radiation Oncology* Biology* Physics, 81(4), e673-e679.

37. Fraass, B. A. (2012). Impact of complexity and computer control on errors in radiation therapy. Annals of the ICRP, 41(3-4), 188-196.

38. Hendee, W. R., & Herman, M. G. (2011). Improving patient safety in radiation oncology. Medical physics, 38(1), 78-82.

39. Mazur, L. M., Mosaly, P. R., Hoyle, L. M., Jones, E. L., Chera, B. S., & Marks, L. B. (2014). Relating physician's workload with errors during radiation therapy planning. Practical radiation oncology, 4(2), 71-75.

40. Belletti, S., Dutreix, A., Garavaglia, G., Gfirtner, H., Haywood, J., Jessen, K. A., ... & Thwaites, D. (1996). Quality assurance in radiotherapy: the importance of medical physics staffing levels. Recommendations from an ESTRO/EFOMP joint task group. Radiotherapy and oncology, 41(1), 89-94.

41. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

42. Mesko, S., Wang, H., Tung, S., Wang, C., Pasalic, D., Chapman, B. V., ... & Phan, J. (2020). Estimating PTV margins in head and neck stereotactic ablative radiation therapy (SABR) through target site analysis of positioning and intrafractional accuracy. International Journal of Radiation Oncology* Biology* Physics, 106(1), 185-193.

43. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

44. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2019). Advances in Neural Information Processing Systems 34 (NeurIPS 2021).

45. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

46. Janocha, K., & Czarnecki, W. M. (2017). On loss functions for deep neural networks in classification. arXiv preprint arXiv:1702.05659.

47. Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

48. Goutte, C., & Gaussier, E. (2005, March). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European conference on information retrieval (pp. 345-359). Berlin, Heidelberg: Springer Berlin Heidelberg.

49. Guilford, J. P. (1954). Psychometric methods.

50. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21, 1-13.

51. Busscher, I., Ploegmakers, J. J., Verkerke, G. J., & Veldhuizen, A. G. (2010). Comparative anatomical dimensions of the complete human and porcine spine. European Spine Journal, 19, 1104-1114.

52. Schulze, R., Heil, U., Groβ, D., Bruellmann, D. D., Dranischnikow, E., Schwanecke, U., & Schoemer, E. (2011). Artefacts in CBCT: a review. Dentomaxillofacial Radiology, 40(5), 265-273.

53. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis, 53, 197-207.

54. Agarwal, J. P., Krishnatry, R., Panda, G., Pathak, R., Vartak, C., Kinhikar, R. A., ... & Deshpande, D. D. (2019). An Audit for Radiotherapy Planning and Treatment Errors From a Low–Middle-Income Country Centre. Clinical Oncology, 31(1), e67-e74.

55. Izewska, J., Andreo, P., Vatnitsky, S., & Shortt, K. R. (2003). The IAEA/WHO TLD postal dose quality audits for radiotherapy: a perspective of dosimetry practices at hospitals in developing countries. Radiotherapy and oncology, 69(1), 91-97.

56. Izewska, J., Vatnitsky, S., & Shortt, K. R. (2006). Postal dose audits for radiotherapy centers in Latin America and the Caribbean: trends in 1969-2003. Revista Panamericana de Salud Pública, 20(2-3), 161-172.

57. Ford, E. C., Terezakis, S., Souranis, A., Harris, K., Gay, H., & Mutic, S. (2012). Quality control quantification (QCQ): a tool to measure the value of quality control checks in radiation oncology. International Journal of Radiation Oncology* Biology* Physics, 84(3), e263-e269.

58. Posiewnik, M., & Piotrowski, T. (2019). A review of cone-beam CT applications for adaptive radiotherapy of prostate cancer. Physica Medica, 59, 13-21.

59. Moazzezi, M., Rose, B., Kisling, K., Moore, K. L., & Ray, X. (2021). Prospects for daily online adaptive radiotherapy via ethos for prostate cancer patients without nodal involvement using unedited CBCT auto-segmentation. Journal of applied clinical medical physics, 22(10), 82-93.

60. Fu, Y., Lei, Y., Wang, T., Tian, S., Patel, P., Jani, A. B., ... & Yang, X. (2020). Pelvic multi-organ segmentation on cone-beam CT for prostate adaptive radiotherapy. Medical physics, 47(8), 3415-3422.

61. Dai, X., Lei, Y., Wang, T., Dhabaan, A. H., McDonald, M., Beitler, J. J., ... & Yang, X. (2021). Head-and-neck organs-at-risk auto-delineation using dual pyramid networks for CBCT-guided adaptive radiotherapy. Physics in Medicine & Biology, 66(4), 045021.

62. Dai, X., Lei, Y., Wynne, J., Janopaul-Naylor, J., Wang, T., Roper, J., ... & Yang, X. (2021). Synthetic CT-aided multiorgan segmentation for CBCT-guided adaptive pancreatic radiotherapy. Medical physics, 48(11), 7063-7073.

63. Gueld, M. O., Kohnen, M., Keysers, D., Schubert, H., Wein, B. B., Bredno, J., & Lehmann, T. M. (2002, May). Quality of DICOM header information for image categorization. In Medical imaging 2002: PACS and integrated medical information systems: design and evaluation (Vol. 4685, pp. 280-287). SPIE.

64. Samara, E. T., Fitousi, N., & Bosmans, H. (2022). Quality assurance of dose management systems. Physica Medica, 99, 10-15.

65. Wada, Y., Morishita, J., Yoon, Y., Okumura, M., & Ikeda, N. (2020). A simple method for the automatic classification of body parts and detection of implanted metal using postmortem computed tomography scout view. Radiological physics and technology, 13, 378-384.

66. Roth, H. R., Lee, C. T., Shin, H. C., Seff, A., Kim, L., Yao, J., ... & Summers, R. M. (2015, April). Anatomy-specific classification of medical images using deep convolutional nets. In 2015 IEEE 12th international symposium on biomedical imaging (ISBI) (pp. 101-104). IEEE.

67. Lee, H., Huang, C., Yune, S., Tajmir, S. H., Kim, M., & Do, S. (2019). Machine friendly machine learning: interpretation of computed tomography without image reconstruction. Scientific reports, 9(1), 1-9.

68. Ouyang, Z., Zhang, P., Pan, W., & Li, Q. (2022). Deep learning-based body part recognition algorithm for three-dimensional medical images. Medical Physics, 49(5), 3067-3079.

69. Lechuga, L., & Weidlich, G. A. (2016). Cone beam CT vs. fan beam CT: a comparison of image quality and dose delivered between two differing CT imaging modalities. Cureus, 8(9).

70. Dubec, M., Brown, S., Chuter, R., Hales, R., Whiteside, L., Rodgers, J., ... & Cobben, D. (2021). MRI and CBCT for lymph node identification and registration in patients with NSCLC undergoing radical radiotherapy. Radiotherapy and Oncology, 159, 112-118.

71. Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.

72. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

73. Lahat, G., Lazar, A., & Lev, D. (2008). Sarcoma epidemiology and etiology: potential environmental and genetic factors. Surgical Clinics of North America, 88(3), 451-481.

74. Uemura, T., Näppi, J. J., Hironaka, T., Kim, H., & Yoshida, H. (2020, March). Comparative performance of 3D-DenseNet, 3D-ResNet, and 3D-VGG models in polyp detection for CT colonography. In Medical Imaging 2020: Computer-Aided Diagnosis (Vol. 11314, pp. 736-741). SPIE.

75. Yu, J., Yang, B., Wang, J., Leader, J., Wilson, D., & Pu, J. (2020). 2D CNN versus 3D CNN for false-positive reduction in lung cancer screening. Journal of Medical Imaging, 7(5), 051202-051202.

76. Jaffray, D. A., Langen, K. M., Mageras, G., Dawson, L. A., Yan, D., Mundt, A. J., & Fraass, B. (2013). Safety considerations for IGRT: Executive summary. Practical radiation oncology, 3(3), 167-170.

77. Jaffray, D., Langen, K. M., Mageras, G., Dawson, L., Di Yan, D., Mundt, A. J., & Fraass, B. (2013). Assuring Safety and Quality in Image-guided Delivery of Radiation Therapy. Radiation oncology, 32, 33.

78. Weintraub, S. M., Salter, B. J., Chevalier, C. L., & Ransdell, S. (2021). Human factor associations with safety events in radiation therapy. Journal of applied clinical medical physics, 22(10), 288-294.

79. Paganelli, C., Meschini, G., Molinelli, S., Riboldi, M., & Baroni, G. (2018). Patient-specific validation of deformable image registration in radiation therapy: overview and caveats. Medical physics, 45(10), e908-e922.

80. Rigaud, B., Simon, A., Castelli, J., Lafond, C., Acosta, O., Haigron, P., ... & de Crevoisier, R. (2019). Deformable image registration for radiation therapy: principle, methods, applications and evaluation. Acta Oncologica, 58(9), 1225-1237.

81. McNutt, T. R., Moore, K. L., Wu, B., & Wright, J. L. (2019, October). Use of big data for quality assurance in radiation therapy. In Seminars in Radiation Oncology (Vol. 29, No. 4, pp. 326-332). WB Saunders.

82. El Naqa, I., Ruan, D., Valdes, G., Dekker, A., McNutt, T., Ge, Y., ... & Ten Haken, R. (2018). Machine learning and modeling: Data, validation, communication challenges. Medical physics, 45(10), e834-e840.

83. Xia, J., Mart, C., & Bayouth, J. (2014). A computer aided treatment event recognition system in radiation therapy. Medical Physics, 41(1), 011713.

84. Hadley, S. W., Kessler, M. L., Litzenberg, D. W., Lee, C., Irrer, J., Chen, X., ... & Moran, J. M. (2016). SafetyNet: streamlining and automating QA in radiotherapy. Journal of applied clinical medical physics, 17(1), 387-395.

85. Luximon, D. C., Ritter, T., Fields, E., Neylon, J., Petragallo, R., Abdulkadir, Y., ... & Lamb, J. M. (2022). Development and interinstitutional validation of an automatic vertebral-body misalignment error detector for cone-beam CT-guided radiotherapy. Medical Physics, 49(10), 6410-6423.

86. Scholer, M. J., Ghneim, G. S., Wu, S., Westlake, M., Travers, D. A., Waller, A. E., ... & Wetterhall, S. F. (2007). Defining and applying a method for improving the sensitivity and specificity of an emergency department early event detection system. In AMIA Annual Symposium Proceedings (Vol. 2007, p. 651). American Medical Informatics Association.

87. Neylon, J., Luximon, D. C., Ritter, T., & Lamb, J. M. (2023). Proof-of-concept study of artificial intelligence-assisted review of CBCT image guidance. Journal of Applied Clinical Medical Physics, 24(9), e14016.

88. Hadjiiski, L., Cha, K., Chan, H. P., Drukker, K., Morra, L., Näppi, J. J., ... & Armato III, S. G. (2023). AAPM task group report 273: recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Medical Physics, 50(2), e1-e24.

89. Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., & Decker, S. (2021). Deep learning-based clustering approaches for bioinformatics. Briefings in bioinformatics, 22(1), 393-415.

90. Arunkumar, N., Mohammed, M. A., Abd Ghani, M. K., Ibrahim, D. A., Abdulhay, E., Ramirez-Gonzalez, G., & de Albuquerque, V. H. C. (2019). K-means clustering and neural network for object detecting and identifying abnormality of brain tumor. Soft Computing, 23, 9083-9096.

91. Agarwal, J. P., Krishnatry, R., Panda, G., Pathak, R., Vartak, C., Kinhikar, R. A., ... & Deshpande, D. D. (2019). An Audit for Radiotherapy Planning and Treatment Errors From a Low–Middle-Income Country Centre. Clinical Oncology, 31(1), e67-e74.

92. Yock, A. D., Ahmed, M., Ayala-Peacock, D., Chakravarthy, A. B., & Price, M. (2021). Initial analysis of the dosimetric benefit and clinical resource cost of CBCT-based online adaptive radiotherapy for patients with cancers of the cervix or rectum. Journal of Applied Clinical Medical Physics, 22(10), 210-221.

93. Elsayad, K., Kriz, J., Reinartz, G., Scobioala, S., & Ernst, I. (2016). Cone-beam CT-guided radiotherapy in the management of lung cancer. Strahlentherapie und Onkologie, 192(2), 83.

94. Grégoire, V., Guckenberger, M., Haustermans, K., Lagendijk, J. J., Ménard, C., Pötter, R., ... & Zips, D. (2020). Image guidance in radiation therapy for better cure of cancer. Molecular oncology, 14(7), 1470-1491.

95. Luximon, D. C., Neylon, J., Ritter, T., Agazaryan, N., Hegde, J. V., Steinberg, M. L., ... & Lamb, J. M. (2024). Results of an AI-Based Image Review System to Detect Patient Misalignment Errors in a Multi-Institutional Database of CBCT-Guided Radiotherapy Treatments. International Journal of Radiation Oncology* Biology* Physics.

96. An, J., & Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. Special lecture on IE, 2(1), 1-18.

97. Sun, J., Wang, X., Xiong, N., & Shao, J. (2018). Learning sparse representation with variational auto-encoder for anomaly detection. IEEE Access, 6, 33353-33361.

98. Zavrak, S., & Iskefiyeli, M. (2020). Anomaly-based intrusion detection from network flow features using variational autoencoder. IEEE Access, 8, 108346-108358.

99. Zhang, Y., Chen, Y., Wang, J., & Pan, Z. (2021). Unsupervised deep anomaly detection for multi-sensor time-series signals. IEEE Transactions on Knowledge and Data Engineering, 35(2), 2118-2132.

100. Nayak, R., Pati, U. C., & Das, S. K. (2021). A comprehensive review on deep learning-based methods for video anomaly detection. Image and Vision Computing, 106, 104078.

101. Sandfort, V., Yan, K., Graffy, P. M., Pickhardt, P. J., & Summers, R. M. (2021). Use of variational autoencoders with unsupervised learning to detect incorrect organ segmentations at CT. Radiology: Artificial Intelligence, 3(4), e200218.

102. Ehrhardt, J., & Wilms, M. (2022). Autoencoders and variational autoencoders in medical image analysis. In Biomedical Image Synthesis and Simulation (pp. 129-162). Academic Press.

103. Mansour, R. F., Escorcia-Gutierrez, J., Gamarra, M., Gupta, D., Castillo, O., & Kumar, S. (2021). Unsupervised deep learning based variational autoencoder model for COVID-19 diagnosis and classification. Pattern Recognition Letters, 151, 267-274.

104. Zhang, L., Chen, X., & Yin, J. (2019). Prediction of potential mirna–disease associations through a novel unsupervised deep learning framework with variational autoencoder. Cells, 8(9), 1040.

105. Baur, C., Denner, S., Wiestler, B., Navab, N., & Albarqouni, S. (2021). Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. Medical Image Analysis, 69, 101952.

106. Zimmerer, D., Kohl, S. A., Petersen, J., Isensee, F., & Maier-Hein, K. H. (2018). Context-encoding variational autoencoder for unsupervised anomaly detection. arXiv preprint arXiv:1812.05941.

107. Huang, P., Yan, H., Song, Z., Xu, Y., Hu, Z., & Dai, J. (2023). Combining autoencoder with clustering analysis for anomaly detection in radiotherapy plans. Quantitative Imaging in Medicine and Surgery, 13(4), 2328.

108. Zavrtanik, V., Kristan, M., & Skočaj, D. (2021). Reconstruction by inpainting for visual anomaly detection. Pattern Recognition, 112, 107706.

109. Du, X., Li, B., Zhao, Z., Jiang, B., Shi, Y., Jin, L., & Jin, X. (2024). Anomaly-prior guided inpainting for industrial visual anomaly detection. Optics & Laser Technology, 170, 110296.

110. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

111. Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., ... & Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. ICLR (Poster), 3.

112. Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE signal processing magazine, 26(1), 98-117.

113. Mudeng, V., Kim, M., & Choe, S. W. (2022). Prospects of structural similarity index for medical image analysis. Applied Sciences, 12(8), 3754.

114. Russakoff, D. B., Tomasi, C., Rohlfing, T., & Maurer, C. R. (2004). Image similarity using mutual information of regions. In Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part III 8 (pp. 596-607). Springer Berlin Heidelberg.

115. Penney, G. P., Weese, J., Little, J. A., Desmedt, P., & Hill, D. L. (1998). A comparison of similarity measures for use in 2-D-3-D medical image registration. IEEE transactions on medical imaging, 17(4), 586-595.

116. Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), 37-52.

117. MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

118. Arthur, D., & Vassilvitskii, S. (2007, January). k-means++: The advantages of careful seeding. In Soda (Vol. 7, pp. 1027-1035).

119. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18 (pp. 234-241). Springer International Publishing.

120. Marvaso, G., Pepa, M., Volpe, S., Mastroleo, F., Zaffaroni, M., Vincini, M. G., ... & Jereczek-Fossa, B. A. (2022). Virtual and augmented reality as a novel opportunity to unleash the power of radiotherapy in the digital era: a scoping review. Applied Sciences, 12(22), 11308.

121. Luximon, D. C., Neylon, J., & Lamb, J. M. (2023). Feasibility of a deep-learning based anatomical region labeling tool for Cone-Beam Computed Tomography scans in radiotherapy. Physics and Imaging in Radiation Oncology, 25, 100427.