2020

DATA SCIENCE METHODS FOR STANDARDIZATION, SAFETY, AND QUALITY ASSURANCE IN RADIATION ONCOLOGY

Khajamoinuddin Syed
Virginia Commonwealth University

DISSERTATION

DATA SCIENCE METHODS FOR STANDARDIZATION, SAFETY, AND QUALITY ASSURANCE IN RADIATION ONCOLOGY

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

KHAJAMOINUDDIN SYED

MS, State University of New York at Albany - August 2013 to December 2014

Director: Dissertation Preetam Ghosh,

Professor, Department of Computer Science

Virginia Commonwealth University

Richmond, Virginia

August, 2020

Acknowledgements

I am obliged to my advisor, Dr. Preetam Ghosh, for his constant support, valuable suggestions, and sensible criticism to inspire and improve my work.

I want to acknowledge Dr. Jatinder Palta, Dr. Bridget McInnes, Dr. Lulin Yuan, and Dr. Thang Dinh for serving on my thesis committee.

My sincere thanks to my collaborators Rishabh Kapoor, William Sleeman IV, and Dr. Michael Hagan, for their encouragement, suggestions, and valuable guidance for improving my understanding of the radiation oncology domain.

I shall never forget the enormous and timely help of my senior colleagues Dr. Bhanu Kamapantula and Dr. Jospeh Nalluri.

I have been extremely fortunate in having lab mates like Pratip Rana, Ahmad Al Musawi, and Priyankar Bose. I would also like to thank my fellow doctoral students and friends, notably Mohammed Obedullah Khan, Samantha Mahendran, and Muhammad Haris Rais. Their helping hand was evident at every stage of my stress, anxiety, and venture in my research.

I want to thank my nephew Asad, who has been a constant source of happiness from the day he was born.

Finally, it is an immense pleasure to express sincere gratitude and heartfelt respect to my mother, Gharibunnisa Begum, my wife, Arshiya Anjum, and my family for their boundless love and motivation. Without their affection and moral support, I would not have come up to this stage.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract

DISSERTATION

DATA SCIENCE METHODS FOR STANDARDIZATION, SAFETY, AND QUALITY ASSURANCE IN RADIATION ONCOLOGY

By Khajamoinuddin Syed

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2020.

Director: Dissertation Preetam Ghosh,

Professor, Department of Computer Science

Radiation oncology is the field of medicine that deals with treating cancer patients through ionizing radiation. The clinical modality or technique used to treat the cancer patients in the radiation oncology domain is referred to as radiation therapy. Radiation therapy aims to deliver precisely measured dose irradiation to a defined tumor volume (target) with as minimal damage as possible to surrounding healthy tissue (organs-at-risk), resulting in eradication of the tumor, high quality of life, and prolongation of survival. A typical radiotherapy process requires the use of different clinical systems at various stages of the workflow. The data generated in these different stages of workflow is stored in an unstructured and non-standard format, which hinders interoperability and interconnectivity of data, thereby making it difficult to translate all of these datasets into knowledge that supports decision-making in routine clinical practice. In this dissertation, we present an enterprise-level informatics platform that can automatically extract and efficiently store clinical, treatment, imaging,

and genomics data from radiation oncology patients. Additionally, we propose data science methods for *data standardization*, *safety*, and *treatment quality analysis* in radiation oncology. We demonstrate that our data standardization methods using word embeddings and machine learning are robust and highly generalizable on real-word clinical datasets collected from the nationwide radiation therapy centers administered by the US Veterans' Health Administration. We also present different heterogeneous data integration approaches to enhance the data standardization process. For patient safety, we analyze the radiation oncology incident reports and propose an integrated natural language processing and machine learning based pipeline to automate the incident triage and prioritization process. We demonstrate that a deep learning based transfer learning approach helps in the automated incident triage process. Finally, we address the issue of treatment quality in terms of automated treatment planning in clinical decision support systems. We show that supervised machine learning methods can efficiently generate clinical hypotheses from radiation oncology treatment plans and demonstrate our framework's data analytics capability.

CHAPTER 1

INTRODUCTION

In the domain of radiation oncology, large amounts of data are captured routinely across several clinical systems over the course of patients' treatment. The electronic health record (EHR) systems are used to document clinical data, which are often stored in free text and unstructured format, wherein key data fields become difficult to abstract for any subsequent data mining efforts. For each patient treated by the radiation oncology department, clinical documentation in the EHR includes the following: (1) a detailed initial consultation note; (2) a simulation note describing the treatment simulation procedure; (3) a treatment planning note documenting the prescription and proposed treatment plan; (4) a weekly On Treatment Visit (OTV) note from the staff physician review of patient's treatment progress and documenting acute side effects; (5) a treatment summary or survivorship care plan for the patient and referring provider at completion of therapy; and (6) routine follow-up notes tracking disease outcomes and any late toxicities. These clinical notes are usually dictated on the telephone, transcribed and imported into the EHR as preliminary documents, and edited by the dictation provider before finalization. There is a wealth of information in these clinical notes for big data applications but the challenge is to capture this data in a discrete format as part of the standard clinical workflow.

The dosimetry data from the treatment planning systems (TPS) which includes the treatment plan, images, dose, structure set, and dose-volume information are stored in structured formats by following the DICOM-RT standard and TG-263 nomenclature [1]. Additionally, the radiotherapy treatment management system

(RTMS) contains information regarding the Radiotherapy (RT) dose delivery, fractions, visits. Each of these clinical systems store data for different purposes, in different formats and in different databases and also employ different mechanisms for sharing these data. For example, the radiation oncologist must access the two software systems mentioned above to clinically manage patients. Moreover, most RT product vendors have no incentive in accommodating each other's data or translate their data format into a standardized nomenclature. The lack of inter-connectivity and interoperability of RT software systems have made the process of data sharing and/or transfer cumbersome and difficult. Hence, valuable clinical and radiation treatment data unfortunately remain trapped behind such proprietary software systems.

The second challenge is that if RT data are manually aggregated and stored in one database, it becomes extremely difficult to clean, parse, collate and scale the data intelligibly. This prevents the ability to create a coherent picture of the patient's comprehensive clinical and treatment record into a single format capable of further utilities and data analytics. This is largely because physician's clinical assessments and diagnoses are often stored as free-text notes, making it extremely difficult to extract critical information with enough accuracy on an automated level. Owing to such challenges, many research and operational tasks that deal with the optimization of quality of care, research-based analysis of RT treatments, diagnosis-based research and development of computer-aided diagnostic tools at infrastructural level are difficult to perform. To this end, the National Radiation Oncology Program (NROP) office at the US Veterans' Health Administration (VHA) designed an initiative to develop an integrated enterprise-wide data curation, storage and analytics portal, called HINGE (Health Information Gateway and Exchange). HINGE is electronically connected to the EHR, TMS and TPS with a specific goal of enabling big data analytics in radiation oncology. It is an automatic data aggregator that collates data from dif-

ferent radiotherapy clinical systems/IT applications. It processes the treatment data for quality assessment, predictive analytics and other enterprise-driven clinical informatics solutions within a single online data portal. Additionally, HINGE's design and infrastructure caters to the imminent need for a research-based practice environment and is cognizant of the role of advanced modern computational strategies involving big-data predictive analytics and clinical informatics.

In this dissertation, we present an agile and scalable software architecture of the HINGE system and different data science approaches to solve the issues related to data standardization, patient safety, and treatment quality assurance in radiation oncology.

In Chapter 2, we present the details of the different clinical systems that were used in data collection along with a workflow of the radiation therapy treatment process to motivate a high-level software architecture for the enterprise-level HINGE system. We also present an outline of the data science methods and relevant evaluation metrics that were employed in this dissertation.

In Chapter 3 and 4, we present the different approaches to standardize the radiotherapy structure names. Specific contributions of Chapter 3 are as follows.

1. We present a machine learning approach to standardize the radiotherapy structure names that can automatically convert the arbitrary physician-given structure names to the domain wide TG-263 based nomenclature.

2. We demonstrate that a relatively small amount of data from each treatment center is enough to build a generalizable machine learning model, which a simple text mapping cannot achieve.

3. We establish that our proposed approach is disease site agnostic, i.e., it can be used on multiple disease sites.

4. We also demonstrate that physician-given names hold enough information about the structures that can be utilized to predict the standard names in TG-263.

5. Finally, we create a scalable approach that requires little to no preprocessing.

In Chapter 4, we address the limitations of structure name standardization using solely physician-given names and present an approach that utilizes the geometric information of structures for standardization. Specific contributions of this chapter are as follows.

1. We demonstrate that the use of bony anatomy information along with structures helps in the standardization process using geometric information.

2. We show that even target structure can be identified along with the Organs at Risk (OARs) with the physician-given names.

3. We demonstrate that it is still challenging to predict the standard name with just geometric information in real-world clinical datasets.

4. We finally demonstrate that integrating physician-given structure names with geometric information of structures improves the overall structure name standardization process.

In Chapter 5, we focus on the safety aspects of radiation oncology. We specifically looked at the triage process in incident learning system. In this chapter, we present machine learning approaches to automatically identify incident severity with an overarching goal of automating the incident triage and prioritization process. Specific contribution of this chapter are as follows.

1. We present an approach to automatically identify the severity of the radiation oncology incidents using the textual incident description.

2. We demonstrate that identifying the severity is a challenging problem when it comes to classifying the incidents into the four possible categories using just the incident description. However, merging severity types into two categories (High and Low severities) results in much better classification accuracy considering the incident report data from multiple VHA radiation oncology centers as well as the VCU medical center datasets.

3. We next demonstrate that transfer learning does help in the severity prediction process specifically considering multi-institution data that may each follow a different protocol for recording the incident reports.

4. We show that incident reports are correlated with institutional practices and there is a need for standardized incident reporting guidelines to reduce the subjective incident analysis practices.

Finally, in Chapter 6, we consider the treatment quality component of the radiation therapy process.

1. We present feature engineering methods to analyze the treatment selection practices for High or Intermediate risk prostate cancer patients across 34 different VHA radiation therapy centers.

2. We demonstrate that there is an inherent bias in the treatment selection process at the VHA treatment centers. The selected treatments deviate from the NCCN guidelines and there is little to no correlation for this deviation with specific treatment center attributes such as, number of radiation oncologists, radiation therapists, other staff or treatment resources.

Figure 1 shows the dissertation outline. Chapters 3 and 4 use material from four

different publications [2, 3, 4, 5]. Chapter 5, uses material from [6, 7]. Chapter 6 uses the material from [8, 9].

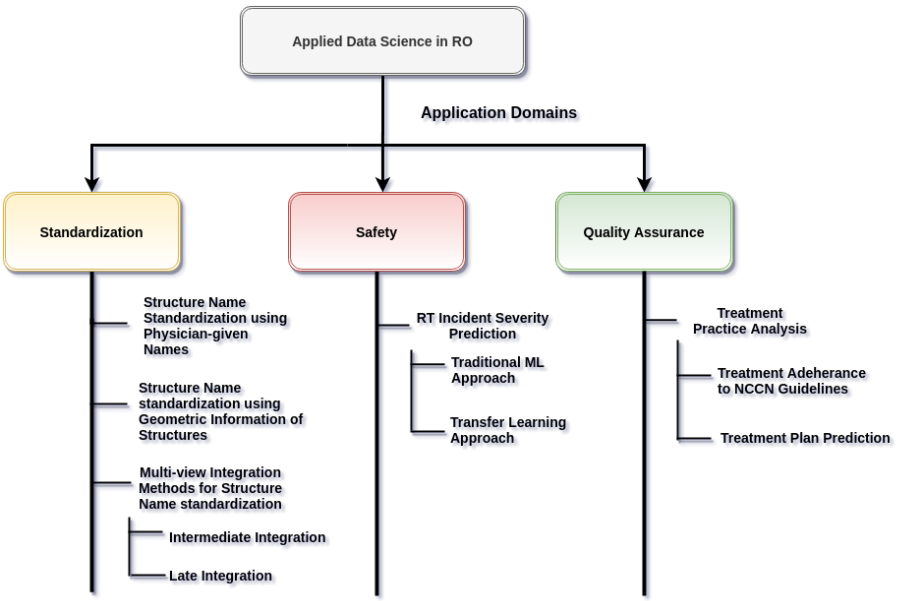


Fig 1: Thesis contribution flow chart. Contributions are done in three domains in radiation oncology: standardization, safety, and quality assurance.

CHAPTER 2

BACKGROUND

2.1 Radiation Therapy Process

It is imperative to understand the layout and structure of RT clinical workflow while building solutions for the issues involved. The RT clinical workflow can be divided into four steps. Figure 2 shows the RT clinical workflow. The steps involved in the RT workflow are mentioned below.

- **Consultation**: In this step, patients meet with a radiation oncologist and they both go through the available treatment options. The radiation oncologist asks a series of questions to determine the best possible treatment options.

- **Treatment Planning**: The radiation oncologist scans the patients using CT or MRI and simulates the radiation treatment to determine the best course of treatment. Simulation involves the identification of the target (tumor) and neighboring anatomical structures to ensure minimal radiation exposure to the healthy tissues.

- **Treatment Delivery**: In this step, the actual treatment is delivered. It involves keeping a record of the treatment delivered and scanning to enable modifications to the treatment according to the patient's response to the treatment.

- **Follow-up**: Once the treatment is complete, radiation oncologists set up a series of follow up meetings with the patients to keep track of the disease. It also helps the patients in providing details on their quality of life, post-treatment,

and receive appropriate care to improve it.

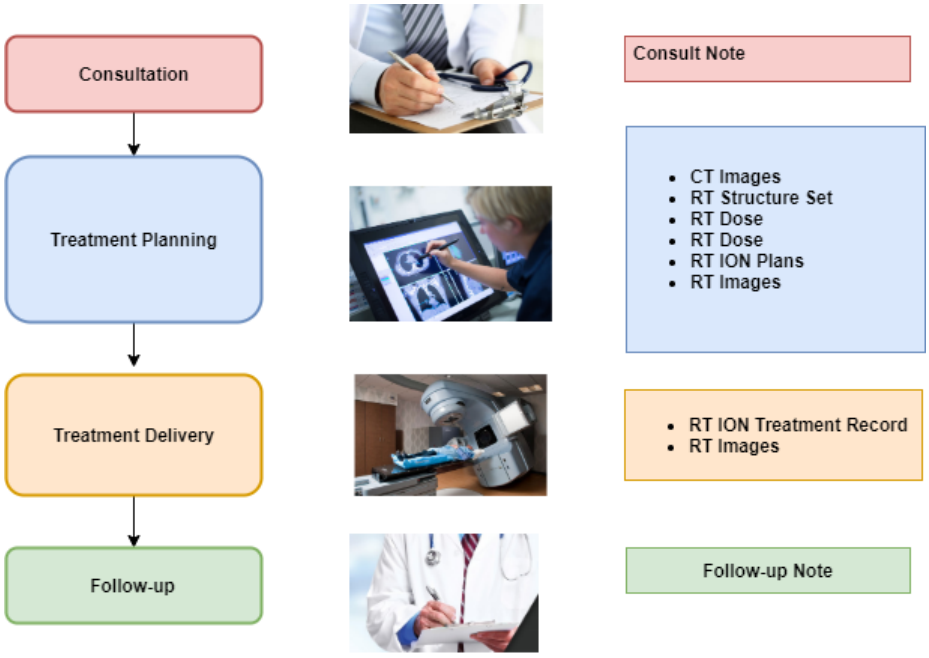


Fig 2: Typical radiotherapy clinical workflow. Four major steps in RT process are shown and type of data generated in each step is shown on the right.

Each patient encounter or a set of encounters is documented in different clinical systems within the department as the patient progresses through the sequential radiotherapy (RT) workflow. Figure 3 shows the clinical workflow and the respective clinical system used in each step of the RT process.

Electronic Health Records (EHRs) are a digital version of a patient's paper chart. EHRs are real-time, patient-centered records that make information available instantly and securely to authorized users. On the other hand, treatment planning systems (TPS) contain information about a prescribed radiation therapy treatment plan by physicians and dosimetrists. Treatment management systems (TMS) use plans generated by the TPS as input and deliver the radiation to the patient. These individual systems often comprise of proprietary software that records and documents

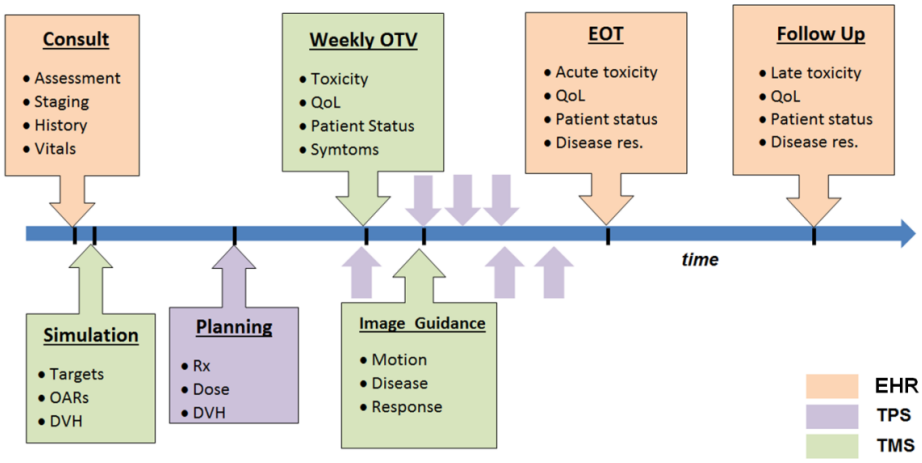


Fig 3: Encounters between physicians and patients during the entire treatment. The information is recorded in different clinical IT systems: EHR, TPS and TMS.

this information. Besides these clinical systems, most of the radiation oncology departments make use of incident learning systems (ILS). This system is used to report incidents that occur at all stages of the RT workflow. A major quality enhancement criteria for the ILS system is to be able to automatically classify incident reports into different severity categories as addressed in this dissertation. This is needed to optimize the operations and resources allocated to attend to incidents of varying severity and improve the overall quality of care.

These independent clinical systems have their own interfaces, proprietary data format and databases which are not interoperable with each other. Data from all these systems need to be extracted which in itself is a giant institutional task since it involves integrating research-based modules (i.e., algorithms, machine learning and natural language processing) with clinical systems (i.e., data). Such a transition from *concepts* to *applications* needs a new layered software infrastructure, as shown in Figure 4 [2].

Treatment data of patients has to be routinely accessed from the clinical tier,

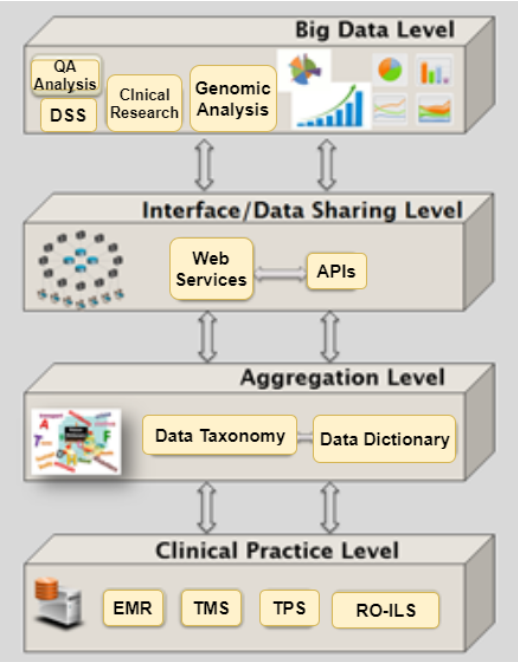


Fig 4: Overview of an informatics-driven clinical infrastructure. Data exchange happens across several tiers which are modularized for specific services.

parsed through the aggregation tier and made available through data sharing interfaces to act as endpoints for the research-based algorithms/applications. In addition to the data acquisition challenges, other important parameters such as, permissions/rights regarding the surrounding architecture, data type, data structures, data rules/restrictions, privacy and compliance, institutional review board (IRB) approvals, data security, etc. have to be resolved too. Building an informatics-driven clinical infrastructure embedded with artificial intelligence (AL) and/or machine learning (ML) tools, requires investment and participation from all the stakeholders and policy makers of the clinical institution.

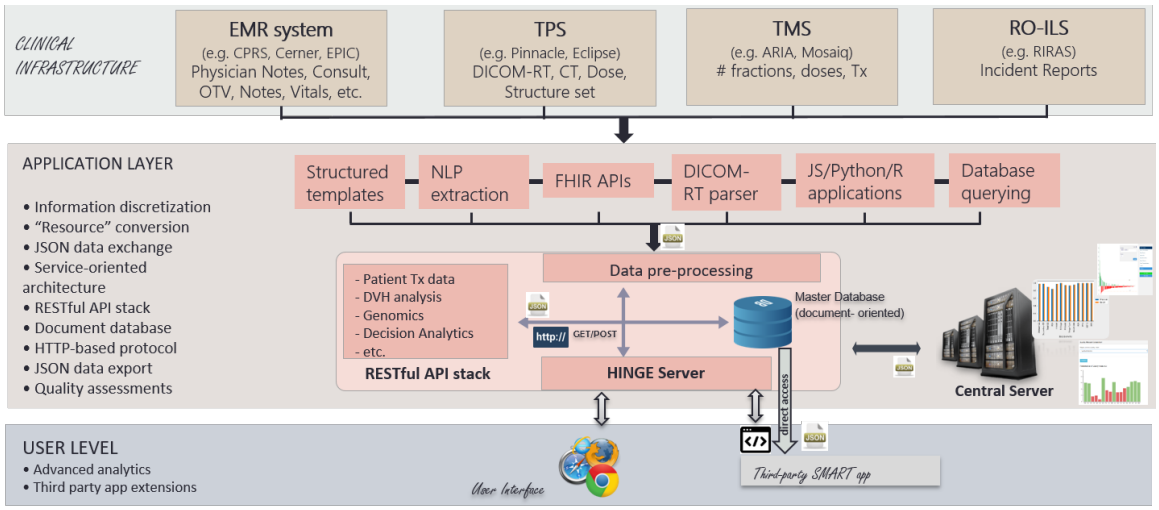


Fig 5: Radiation oncology data curation, standardization, and analytics platform (EMR, TPS, TMS, and RO-ILS).

2.2 Naming Standards

The present lack of radiotherapy structure name standardization in practice not be associated with the actual inexistence of standards. Multiple standards have been proposed, and the widespread attention of the clinical world to the need for naming standards is increasing [10]. In this section, we discuss the one such standard used in RT medical practices.

Ontologies

Ontologies provide a rich framework for defining concepts and inter-relationships among them. The BioPortal [11] is a website that is maintained by the National Center for Biomedical Ontology contains a wide variety of medical ontologies that are publicly accessible. Ontologies are helpful to represent essential components in interoperability and integration into healthcare informatics.

American Association of Physicists in Medicine's Task Group-263

The American Association of Physicists in Medicine (AAPM) is a scientific and professional organization. One of the primary goals of AAPM is to identify and implement improvements in patient safety for the medical use of radiation in imaging and radiation therapy.

In 2018, AAPM released the final report of its task group numbered 263 (TG-263) [12], with a focus on identifying a comprehensive nomenclature standard for RT, which could be efficiently and proficiently used in every medical institution in the United States. Task group developed a comprehensive nomenclature system of all the concepts after reviewing the already available medical ontologies and the recent development in standards for nomenclature in RT [13, 14]. Special consideration was given to practical limitations (like characters supported by vendors' solutions) and the utilization of names to minimize the chance of communication errors. Essential RT concepts that were not covered in other medical ontologies are covered in detail. TG-263 is not an ontology but can be considered as a set of simple naming guidelines and conventions. As a result, TG-263 names are short but easy to understand and interpret, even without a strong anatomy background. When possible, it provides the most closely matching Foundational Model of Anatomy (FMA) identifier of the structure, thus providing the direct linking between the FMA and TG-263. An accurately standardized TG-263 clinical dataset is more useful for medical purposes. Also, it will make it easier to use semantic web technologies, thanks to the integration with the FMA ontology. The standard structure names continuously updated and made publicly accessible [1]. With the easy to follow guidelines and tremendous adoption of medical practices, a new challenge has emerged in the radiation oncology domain, updating retrospective DICOM datasets with standardized structure names

compliant with the TG-263 standard.

2.3 Machine Learning Algorithms

Here, we will briefly describe the supervised machine learning algorithms that have been used in this dissertation.

In a supervised machine learning approach, algorithms know the correct labels of the data it is trying to learn. We have used supervised classification algorithms that try to learn a patterns to categorize the data points into two or more categories. Below are some of the classification algorithms used in this dissertation.

Logistic Regression (LR)

Logistic regression is a simple linear classification algorithm that takes in a vector and converts it to the probability ranging between 0 and 1. It uses a sigmoid function to convert the value. For binary classification, a cutoff value is used to decide the class label. It is easy to interpret due to its linear nature. Even though it is predominantly used for binary classification, it can also be used for multi-class classification.

Support Vector Machines (SVM)

Support vector machines make use of a hyperplane or a set of hyperplanes to distinctively classify the data points. Linear SVM makes use of maximum-margin hyperplanes to classify the linearly separable datapoints [15]. Alternatively, non-linear SVM uses the function to map the input vector to a high-dimensional or infinite-dimensional vector space and determines the hyperplane in the new space to classify the data points [16]. It has been previously observed that SVMs have consistently outperformed many other classifiers in text categorization problems, and they are less susceptible to imbalanced datasets [17].

k-Nearest Neighbors (kNN)

(kNN) [18] is a simple but powerful machine learning algorithm that can be used for both supervised and unsupervised learning. This algorithm finds the k-nearest neighbors in a dataset (with n samples) when compared to a new example. The distances between examples are calculated on each feature with a distance metric such as Euclidean, Manhattan, or Mahalanobis. The only parameter for kNN is the value k itself. According to [19], choosing k to be $\sqrt{n}$ is a good option, although other values may be better depending on the properties of the dataset and application.

Unlike many machine learning algorithms, the traditional kNN algorithm does not require a training phase as the queries are simply compared against the examples in the existing dataset. Although the brute force kNN will produce the true k-nearest neighbors, it will also have poor computational performance as the number of example queries or the underlying dataset becomes large.

Random Forests (RF)

Random Forests consists of multiple decision trees, but each tree can only be split based on the randomly selected subset of features from the randomly selected samples. For each tree, different subset of samples and subset of features are selected randomly. For classification, majority voted label is considered as the predicted label [20].

2.4 Machine Learning Model Training Process

Training a machine learning model involves a lot of experimentation, such as selecting different algorithms and selecting appropriate hyperparameters. The final selected model needs to be optimized by choosing a different set of hyperparameters.

Each set of hyperparameters with training data leads to a different model. Since we are interested in selecting the best performing model from this set, we need to compare their performance. The dataset is divided into the following three sets to train and select the best machine learning model.

- **Training Set:** This contains the instances and labels used for training the model.
- **Validation Set:** This is used to calculate the performance of the model and hyper-parameter selection.
- **Testing Set:** This set is used to test the predictive performance of the final selected model. Test set samples are never seen by the model either at training or validation, thus mimicking the real-world data.

To correctly estimate the model performance, we assume that training, validation, and test sets are coming from the same distribution. In classification, to maintain a similar distribution across each set, data is split in such a way that an equal percentage of instances from each class are in each set, which is also known as a stratified split.

2.5 Evaluation Metrics

The performance of a model is evaluated by comparing the model's prediction against the actual (true) class. In classification, a confusion matrix is used to describe the model's predictions.

Figure 6 shows the confusion matrix for a binary classification task. In binary classification, data points are divided into two classes; Positive (P) and Negative (N) class. Model predictions are categorized into the following components:

FIGURE Page

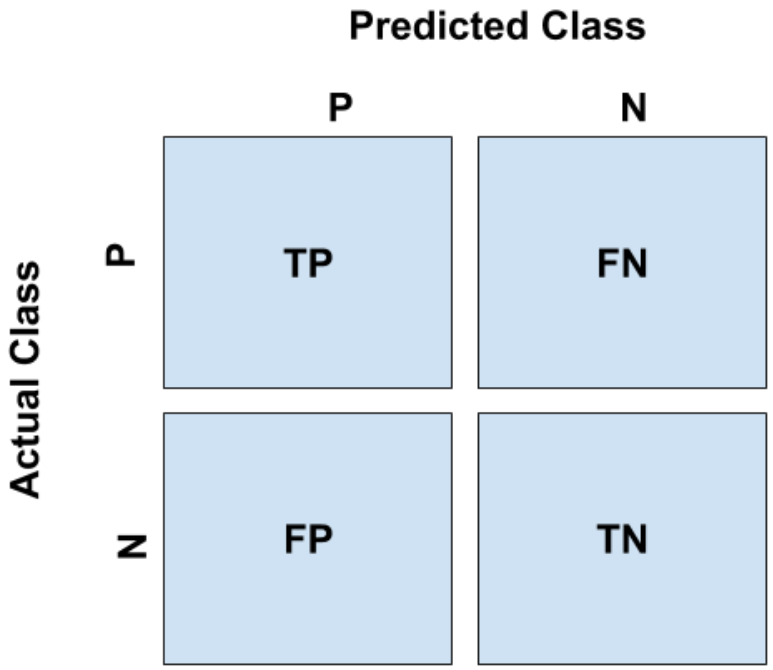


Fig 6: A diagrammatic representation of binary classification confusion matrix.

True Positive (TP): This is when the model predicted a positive class and the actual class is also positive.

False Positive (FP): The model predicted a positive, but the actual class is negative.

False Negative (FN): This is when the model predicted a negative, but the actual class is positive.

True Negative (TN): The model predicted a negative and the actual class is also negative.

Using the confusion matrix components, different types of classification metrics are calculated. The mathematical expressions of each of these metrics are shown below.

$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

$$F_1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

For multi-class classification, an overall model's performance can be calculated using these metrics. A macro-averaged metric computes results for each class independently and then takes the average of all the classes to calculate the overall average metric. In contrast, a micro-average aggregates the contributions of all classes to compute the overall metric. We note that in classification tasks such as ours, in which each structure name is mapped to precisely one label (as in the structure name stan-

dardization problem), accuracy is the same as the micro-averaged F_1-Score. A micro-averaged F_1-Score and overall accuracy metric do not disproportionately penalize a classifier for performing poorly on the less frequent classes, whereas macro-averaged F_1-Score is heavily influenced by how well the classifier performs on the less frequent classes. Hence the performance of a rare class and a more frequent class are equally important.

Accuracy measures how well a classifier performs overall, whereas macro-averaged precision, recall, and F_1-Scores better capture how well a classifier can identify cases that it does not often see, which is extremely important in real-world settings.

CHAPTER 3

RADIOTHERAPY STRUCTURE NAME STANDARDIZATION USING PHYSICIAN-GIVEN NAMES

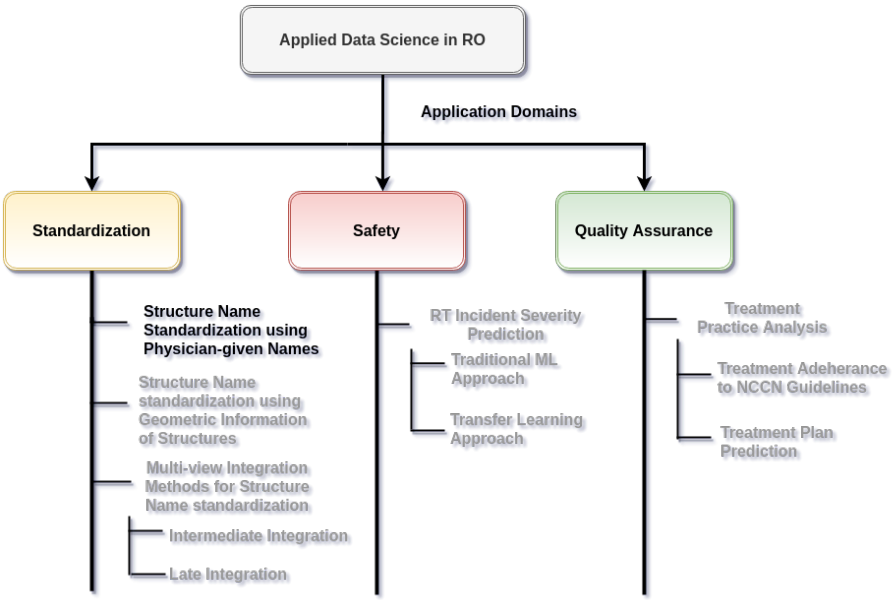


Fig 7: Thesis contribution, Chapter 3 contributions are highlighted.

3.1 Introduction

Radiation therapy is a type of cancer treatment that uses high intensity energy beams to kill cancer cells and shrink the tumor. In order to treat cancer, the radiation oncologist delineates the tumorous region or target volume on a computed tomography (CT) or magnetic resonance imaging (MRI) dataset. Additionally, the normal organs, known as organs-at-risk (OAR) volumes are delineated to spare and estimate radiation doses and reduce possible side effects. These delineated volumes are known

as structures. Radiation oncology team members, such as radiation physicists and dosimetrists, delineate other types of structures termed as "planning organs at risk volume" (PRV). These structures are used strictly in the treatment planning process and take into account the mobility of the organs at risk, and therefore, a surrounding margin is added to these structures to compensate for geometric uncertainties. All delineated structures are given names that are usually written in free text as identifiers, but the lack of standardized nomenclature has created inconsistencies in naming the structures. Figure 8 shows a representative CT image overlaid with its defined structures. The left side of the figure shows the physician-transcribed names of the structures delineated on the right side.

The use of standard nomenclature is an essential step for the construction and use of informatics-based tools to automatically extract pertinent data from electronic medical records in support of clinical trials, data-pooling initiatives, and clinical practice improvement. It also provides a foundation for the development of software tools to automate data extraction, analysis, data submission, exchange, and quality assurance (QA) [21, 22].

To address these issues, the American Association of Physicists in Medicine (AAPM) has released a Task Group 263 (TG-263) report with the standardized nomenclature for structures names [12]. This report was developed in collaboration with stakeholders from both multi-institutional and multi-vendor organizations. The American Society for Radiation Oncology (ASTRO) and AAPM have identified the following as the main challenges in RT structure name standardization [12]:

- Vendor-based challenges that originate from the inter-vendor variation on software architecture. Each vendor has a particular character set for naming the structures; limited allowable character sets, however, hinder the interoperability.

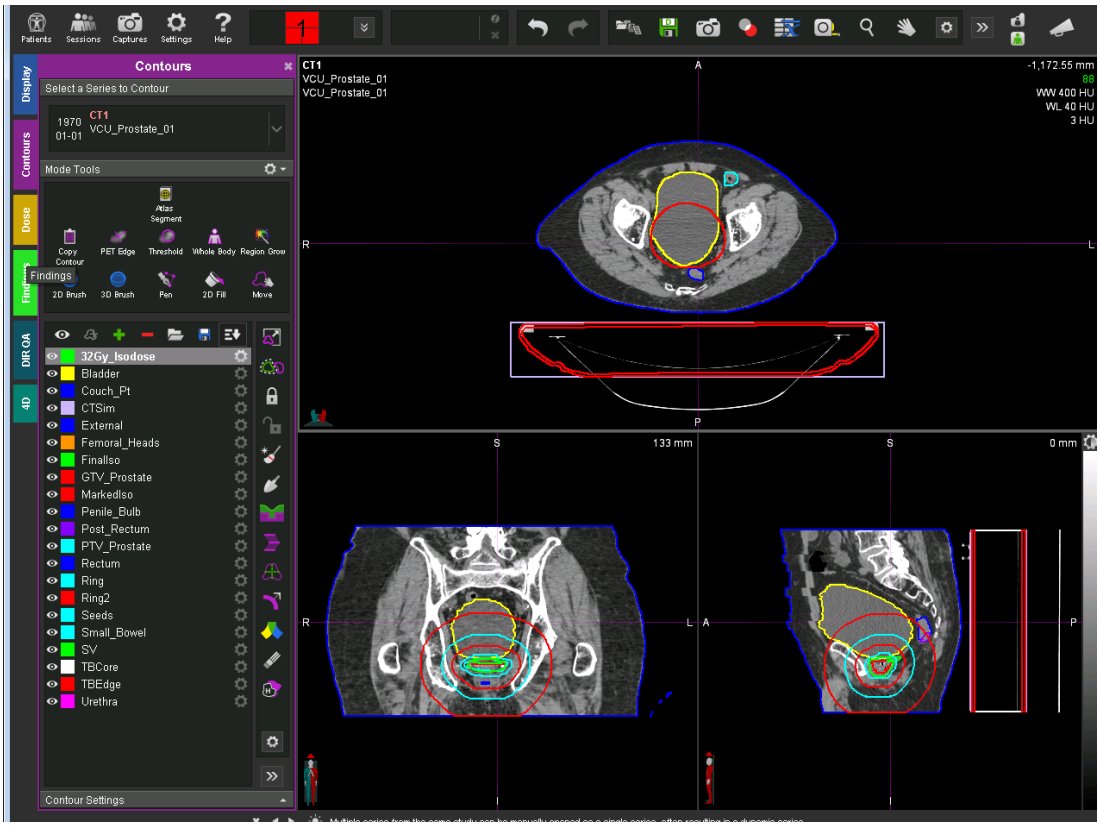


Fig 8: A representative CT image overlaid with its defined structures. The left side of the figure shows the physician-transcribed names of the structures delineated on the right side. The physician-transcribed names and structures delineated can be matched by the color.

- Multi-institutional-based challenges that may arise from the lack of participation, oversight, and guidelines in creating a standardized nomenclature.
- Single institutional challenges include data governance issues, costs, and difficulties in implementing new nomenclatures, making them compatible with existing treatment modalities, and training the institutional staff to follow the standards.
- Clinical staff challenges may encompass the lack of guidelines or a detailed

schema to follow.

Strict adherence to a standardized nomenclature will help to achieve future standardization, but it cannot address retrospective data standardization. Manually relabeling inconsistent names with the corresponding standardized TG-263 names is one way to correct retrospective data; however, generating such mappings for multi-center data is slow, time consuming, inefficient, hard to generalize, and challenging to scale. This sets the stage for machine learning (ML) based methods that may be able to overcome some of these limitations. To address each of the issues mentioned above, we propose a methodology to retrospectively standardize the radiotherapy structure names using a combination of ML and natural language processing (NLP) techniques.

The main contributions of this chapter are:

- Proposing a novel automated machine learning approach to standardize the physician-given structure names to the domain wide utilized TG-263 standard names.
- Demonstrating that a relatively small amount of data from each center is enough to build a generalizable machine learning model, which a simple text mapping cannot achieve.
- Establishing that the approach is disease site agnostic; it can be used on multiple disease sites.
- Demonstrating that physician-given names hold enough information about the structures that can be utilized to predict the standard name.
- Creating a scalable approach that requires little to no preprocessing.

3.2 Related Work

The existing techniques for structure name standardization can be broadly classi-

fied into three categories: expert-based, ontology-based, and machine learning based.

Previous works in the RT community to retrospectively standardize structure names mostly use the physician provided names (free-text labels) or geometric information such as volume, area, and location of the structures. The recently published works to standardize structure names using physician-given names are illustrated as below.

A research team in Australia recently proposed an expert-based approach to standardize radiotherapy structure names as per the TG-263 standard recommendations [23]. In this study, a panel of experts developed a mapping and structure synonym set for 36 structures from their clinical database. With their method, they were able to map 99% of the relevant structures and relabel the names correctly. However, the major limitation of this approach are scalability and generalizability; data used in this project were from a single academically focused institution that could enforce the local standards, and the mappings were dependent on inputs provided by experts. This method is also center specific; mappings from one institute may not be useful to the other institute.

A different team in the Netherlands has proposed an ontology-based RESTful web service to standardize the structure names [24]. However, this approach was more focused on building a linked data than a technique for structure name standardization. The authors used the mappings provided by the institutions to generate centralized mappings, thereby creating a common terminology for linked data.

There are few works that have proposed machine learning based approaches to structure name standardization. Unlike expert-based and ontology-based methods, machine learning based methods use either free text labels or geometric information to build learning models for standardization. One such work made use of multiple string similarity measures to generate feature vectors, and these feature vectors were used

as input for the classification algorithm to predict the labels [25]. This paper used neural-network-based methods but lacked the pertinent details for reproducibility of the results. Two other papers proposed methods using geometrical information for structure name standardization [26, 27]. Both of these papers have used a machine learning approach with neural networks to standardize the structure names of the head and neck region. Even though they both showed a high accuracy for identifying the standard names, the major limitation of these works was that they considered only limited OAR structures to build the ML model and Non_OARs were discarded. Removing Non_OAR structures makes it difficult to apply these two approaches in the real-world datasets which contain a mixture of both OARs and Non_OARs.

Expert-based methods have high accuracy but require manual effort from experienced clinicians, which makes scalability and generalizability challenging to achieve. Although ontology-based techniques can help in automating the labeling task, there is a paucity of domain-specific comprehensive ontologies in the radiation oncology domain. Machine learning based methods are well suited for retrospective structure name relabeling but are seldom used in this domain. Additionally, the TG-263 standardization was only completed in 2018 [12], and hence applications of machine learning based methods for structure name prediction are still in their infancy.

3.3 Methods and Materials

3.3.1 Annotation Process

As part of VA-ROQS, teams of domain experts visited each of the 40 VA facilities that performed radiotherapy in-house and extracted patient data from the local EMR and TPS. The original treatment planning data was reloaded in the TPS software, and the associated imaging, dose and structure set information was reviewed. Using the

full treatment planning information, the domain experts then built a data table that mapped original structure set labels to the preferred TG-263 labels. The structure label standardization was originally performed so that dose-related Quality Measures could be compared across all VA facilities; the resulting information was also used as the true labels for the predictive models in our pipeline. The same annotation process was performed on the VCU data with a local expert.

3.3.2 Dataset

Across the United States, the Veterans Health Administration (VA) has 40 centers treating veterans with in-house radiation therapy services. The VA has put together the Radiation Oncology Quality Surveillance Program (VA-ROQS), and as part of this program the treatment quality is assessed from all VA centers [28]. As part of the initial pilot study, data from all 40 centers were manually abstracted from clinical charts, imaging databases, and radiation oncology specific systems, such as treatment planning systems and treatment management systems. Data from up to 20 prostate and 20 lung cancer patients were manually abstracted from each center, resulting in a total of 794 and 754 patients respectively. The collected data included the DICOM (Digital Imaging and Communication in Medicine) structure set files representing anatomical structures of interest and the corresponding DICOM CT image datasets for each patient. For this project, ten lung and nine prostate OAR structures were identified. These structures were manually labeled to their TG-263 standard names, and all other structures, including target and PRVs, were labeled as Non_OAR. The dataset will be further referred to as the VA-ROQS dataset.

We also collected data from the Department of Radiation Oncology at Virginia Commonwealth University (VCU) as an external test dataset, which included DICOM structure set data from 50 randomly selected patients with prostate cancer

and another 50 patients with lung cancer. The same procedure that was used in the VA-ROQS data preparation was also used to label the structures in this dataset with a local expert, which will be referred to as the VCU dataset.

The structure label standardization was originally performed so that dose related Quality Measures could be compared across all VA facilities; the resulting information was also used as the true labels for the predictive models in our pipeline. Assigning standard labels to DICOM structures was a very time consuming process and has motivated us to find a more automated or semi-automated solution for structure label standardization. This automated solution can additionally help in reducing possible human errors in the manual annotation process.

The following prostate and lung OAR structures were considered in this work:

Prostate organs-at-risk structures: Bladder, Rectum, LargeBowel, SmallBowel, Femur_L, Femur_R, SeminalVesicles, PenileBulb, and External.

Lung organs-at-risk structures: Heart, Esophagus, Lungs, Lung_R, Lung_L, SpincalCord, BrachialPlexus, BrachialPlexus_L, BrachialPlexus_R, and External.

Table 1 shows the distributions of lung structures for the VA-ROQS and VCU datasets, while Table 2 shows the distributions of the prostate structures in these two datasets. In both cases, the Non_OAR structures present an overwhelming majority; these Non_OARs include all the structures contoured as a part of treatment planning and delivery and the dose evaluation structures. We also observed similar class imbalances across all VA-ROQS centers' data (see Figures 40 and 41 in Appendix B). Table 3 shows the examples of physician-given names compared to the standard OAR structures, which highlights the variability in the physician-given names. Table 1 also shows the number of unique names found in each Lung structure in the VA-ROQS and VCU datasets, and Table 2 shows physician-given unique names for the prostate structures in VA-ROQS and VCU datasets.

	VA-ROQS		VCU	
	Non Standard Name		Non Standard Name	
Standard Name	**Total Count**	**Unique Count**	**Total Count**	**Unique Count**
Brachial_Plexus	44	11	0	0
Brachial_Plexus_L	59	14	4	5
Brachial_Plexus_R	69	23	5	3
Carina	497	7	33	2
Esophagus	636	28	46	4
Heart	693	21	47	2
Lung_L	553	46	28	10
Lung_R	563	46	27	10
Lungs	439	39	41	10
Non_OAR	8800	3701	577	259
SpinalCord	689	37	50	7
Total	13,044	3973	858	309

Table 1: Lung structure type distribution in VA-ROQS and VCU datasets.

3.3.3 Data Preprocessing

Structure names are short and have a limited character set to use, and the available character set is vendor dependent. As shown in Table 3, even though there is high variability in physician-given structure names for most of the structure types, the character set used is limited. Preprocessing methods need to be selected to ensure that critical information is retained; losing the information might negatively affect the ability to standardize the structure names with high fidelity. Hence, we decided to keep the preprocessing of physician-given names to a minimum by just converting

	VA-ROQS		VCU	
	Non Standard Name		Non Standard Name	
Standard Name	**Total Count**	**Unique Count**	**Total Count**	**Unique Count**
SmallBowel	250	40	47	7
LargeBowel	341	33	6	2
Femur_R	717	62	31	14
Femur_L	711	59	32	16
Rectum	742	14	50	3
Bladder	738	10	50	3
External	597	5	50	1
SeminalVesicles	510	50	28	8
PenileBulb	590	33	47	12
Non_OAR	9869	2886	813	425
Total	15,065	3195	1154	491

Table 2: Prostate structure type distribution in VA-ROQS and VCU datasets.

them to lower case.

3.3.4 Model Selection

After preprocessing the data, the next step is to select the appropriate machine learning method. We experimented with different types of methods to map the physician-given structure names to the TG-263 standardized names. The datasets presented have some unique characteristics that impacted the choices and performances of our algorithms. Structure names are very short in size (varying between 4 and 20 characters), which limits the use of complex machine learning algorithms [29].

TG-263 Standard Name	Physician-Given Names in Dataset
Large Bowel	Colon_Sigmoid, BOWEL LARGE, Bowel, sigmoid colon, Bowel_LG, SIGMOID_COLON, colon, Sigmoid OAR, Bowel NOS, large bowl, Sigmoid AZ, large bowel, Lg bowel, LG BOWEL, COLON_partial, LargeBowel, Sigmoid-AZ, Bowel Large, Rectosigmoid, Sigmoid Colon, LARGE BOWEL, SIGMOID08JUN16
Femur_L	FEMORAL LT, Femur_L, LFH, Femur_LT, Femoral Head Lt, Femoral Head_Lt,Lt Fem Head, FEMUR_L, left_femhead, Femur L, L_FEM HEAD, Lt Femur, Femur_Head_L, Hip Left, Femur-Lt, Hip Left, Femur-Lt, Lt Femoral Head, Fem hd neck Lt, Lt Hip, lt fem head, Femoral Lt, Femoral Head L, FEM HEAD LT, L Fem Hd,Femur Left, Femur l. , lt femoral hd, Left Femoral head JPC,

Table 3: Examples of physician-given RT structure names in VA-ROQS dataset. Standard names on the left and physician-given names on the right.

For better applicability of the machine learning algorithms, we identified the features from the structure names to build the feature vectors, which are necessary for any machine learning algorithm.

Since machine learning algorithms work on numerical data, we converted the text data into numerical features. Numericalization of text data involves two steps [30]: (1) tokenization or feature set generation and (2) vectorizing the features with feature weight calculation techniques. We tried multiple feature generation and feature weight

calculation methods, as discussed next.

We tested the following list of techniques for feature set generation.

1. Bag-of-words (BoW): In this model, text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity [31]. The bag-of-words model has also been used extensively in the NLP domain. For example, bag-of-words features for the physician-given name "emoral head left" are "femoral", "head", and "left".

2. Word NGram: An NGram is a contiguous sequence of n words from a given sequence of text. Given a sentence, we can construct a list of NGrams from it by finding pairs of words that occur next to each other. For example, with a physician-given name, "femoral head left", we can construct bi-grams (NGram of length 2) by finding consecutive pairs of words; "femoral head" and "head left" are bi-grams.

3. Character NGram: In this model, instead of considering a full token or a term, a set of continuously occurring characters is used to build the feature set. These character sets are considered to form NGram features. For example: with the physician-given name "bladder", character tri-gram features are "bla", "lad", "add", "dde", "der".

Assigning appropriate weights to individual features as per their relevance in a given dataset is known as feature weighting. It is generally thought of as a generalization of feature selection, where the presence of a feature serves as the criterion for its extraction. We used various feature weighting methods to build the feature vectors, as shown below.

1. Term presence (tp): In this method the presence or absence of a term in the

given document is encoded as 1 or 0.

2. Term count (tc): This method is an extension of the tp method. Here, term occurrence is considered as the weight; it denotes the number of times a given term appears in a document.

3. Term frequency (tf): In this method, the term occurrence is usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) from giving a measure of the importance of the term t within the particular document d. Thus we have the term frequency, defined as follows [32, 33].

$$\mathrm{tf}_{t,d} = 1 + \log \mathrm{tf}_{t,d} \tag{3.1}$$

4. Term frequency-inverse document frequency (tf-idf): tf-idf is a numerical statistic that reflects how important a word is to a document in a collection or corpus [34]. It involves two parts: First is tf, which is defined as in Equation (3.1). Second is inverse document frequency (idf)), which is a measure of the general importance of the term (obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient).

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t} \tag{3.2}$$

$$\text{tf-idf}_{t,d} = \mathrm{tf}_{t,d} \cdot \mathrm{idf}_t \tag{3.3}$$

In Equations (3.1)–(3.3), tf is term frequency, df is document frequency, t is term, d is document, df_t is number of documents a term (t) appears in, and N is the total number of documents.

5. Word embeddings: Words or phrases from the vocabulary are mapped to vectors

of real numbers. Conceptually, it involves a mathematical embedding from a space with many dimensions per word to a continuous vector space with a much lower dimension; word2vec [35], Glove [36], and fastText [37] are some of the word embedding techniques.

Feature Weighting Example

Here we show the examples of each of these weighting methods. Consider four physician-given names: (1) *large bowel*, (2) *sigmoid colon*, (3) *bowel*, and (4) *bowel lg* . If we consider the bag-of-words model for feature set generation, our feature set will consist of unique tokens from the above mentioned four names, which are { large, bowel, sigmoid, colon, lg }. The total number of documents is four ($N = 4$) (physician-given names). Below are feature vectors with each of the weighting methods for physician-given name "large bowel" as below.

$$
\begin{aligned}
feature\_Set &= \begin{bmatrix} large & bowel & sigmoid & colon & lg \end{bmatrix} \\
tp &= \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \end{bmatrix} \\
tc &= \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \end{bmatrix} \\
tf &= \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 \end{bmatrix} \\
tf - idf &= \begin{bmatrix} 1.301 & 0.087 & 0 & 0 & 0 \end{bmatrix}
\end{aligned}
$$

We used six different classification algorithms—SVM-linear [15], SVM-RBF [16], k-nearest neighbors (KNN) [18], logistic regression [38], random forest [20], and fastText [37]—for initial model selection. All models were built by using scikit-learn machine learning library in python [39]. The best model was selected based on their

performance on the VA-ROQS dataset. Tables 4 and 5 show the performances of these models for the different feature vector methods. One of the objectives of this work was to understand the impact of feature weighting techniques on model performance. A thorough comparison of feature weighting techniques and their effects on structure name standardization is beyond the scope of this study. Nevertheless, we report the observations we made during the initial model selection as below.

Tables 4 and 5 show the machine learning model performance with different feature weighting methods. We observed that the tp, tc, and tf with all combinations of ML algorithms produced the same results. We observed that these three feature weighting techniques produced the same feature vectors, where tp and tc produce the same vector, and tf is a normalized version of the tc. We believe this is because of the unique characteristics of our dataset. Instances (physician-given names) are short, and words within the names are not repeated. The examples shown above indicate the same. As we know from Equation (3.3), the tf-idf feature weighting technique takes the global picture of words into account in the calculations, which changes the weights of the features when compared to other methods. Interestingly, tf-idf did not perform well when compared to the other weighting methods for both prostate and lung disease datasets. In comparison with all weighting methods, the word vector based fastText algorithm consistently outperformed all other algorithms; hence we selected it to build our final model.

3.3.5 Model Evaluation

An essential part of building a machine learning system is to demonstrate its quantifiable generalizability. For example, the critical goal of a machine learning classification algorithm is to create a learning model that accurately predicts the class labels of unseen data samples. Hence the machine learning model should work

Features	Algorithm	Accuracy	Precision	Recall	F_1-Score
tp	SVM_RBF	0.99	0.96	0.97	**0.97**
	SVM_Linear	0.99	0.96	0.97	**0.97**
	Random_Forest	0.98	0.96	0.97	0.96
	Logistic_Regression	0.99	0.97	0.97	**0.97**
	KNeighbors	0.97	0.94	0.96	0.95
tc	SVM_RBF	0.99	0.96	0.97	**0.97**
	SVM_Linear	0.99	0.96	0.97	**0.97**
	Random_Forest	0.98	0.96	0.97	0.96
	Logistic_Regression	0.99	0.97	0.97	**0.97**
	KNeighbors	0.98	0.94	0.97	0.95
tf	SVM_RBF	0.99	0.96	0.97	**0.97**
	SVM_Linear	0.99	0.96	0.97	**0.97**
	Random_Forest	0.98	0.96	0.97	0.96
	Logistic_Regression	0.99	0.97	0.97	**0.97**
	KNeighbors	0.98	0.94	0.97	0.95
tf-idf	SVM_RBF	0.99	0.97	0.96	**0.97**
	SVM_Linear	0.99	0.97	0.97	**0.97**
	Random_Forest	0.99	0.96	0.97	**0.97**
	Logistic_Regression	0.98	0.97	0.96	0.96
	KNeighbors	0.98	0.95	0.97	0.96
Word-vectors	fastText	**0.99**	**0.97**	**0.97**	**0.97**

Table 4: Initial Model Selection Results for VA-ROQS Prostate datasets.

Features	Algorithm	Accuracy	Precision	Recall	F_1-Score
tp	SVM_RBF	0.99	0.95	0.92	0.93
	SVM_Linear	0.99	0.98	1.00	**0.99**
	Random_Forest	0.99	0.96	0.97	0.96
	Logistic_Regression	0.99	0.97	0.97	0.97
	KNeighbors	0.97	0.88	0.93	0.89
tc	SVM_RBF	0.99	0.95	0.92	0.93
	SVM_Linear	0.99	0.98	1.00	**0.99**
	Random_Forest	0.99	0.96	0.97	0.96
	Logistic_Regression	0.99	0.97	0.97	0.97
	KNeighbors	0.97	0.88	0.93	0.89
tf	SVM_RBF	0.99	0.95	0.92	0.93
	SVM_Linear	0.99	0.98	1.00	**0.99**
	Random_Forest	0.99	0.96	0.97	0.96
	Logistic_Regression	0.99	0.98	0.98	0.98
	KNeighbors	0.97	0.88	0.93	0.89
tf-idf	SVM_RBF	0.99	0.94	0.94	0.94
	SVM_Linear	0.99	0.93	0.93	0.92
	Random_Forest	0.99	0.96	0.97	0.96
	Logistic_Regression	0.99	0.94	0.90	0.92
	KNeighbors	0.98	0.89	0.92	0.90
Word-vectors	fastText	**1.00**	**1.00**	**0.99**	**0.99**

Table 5: Initial Model Selection Results for VA-ROQS Lung datasets.

well for classifying future data.

Model validation is an important step in the machine learning process. Evaluation of a model on the training dataset would result in a biased score. Therefore the model is evaluated on the held-out set to give an unbiased estimate of model performance. Just a hold-out set validation is not enough to test the robustness and finalize the model. It is recommended to validate the model on the entire dataset [40, 41]. One such technique is k-fold cross-validation. To that effect, we validated our models in three different ways on the VA-ROQS dataset and tested it on the VCU dataset (external dataset).

Model Validation

1. **70:30**: The VA-ROQS dataset was divided into a 70:30 ratio as the training and validation sets. The split was stratified by TG-263 standard names, which ensured that an equal percentage of data was taken from each standard name for training, validation, and testing, thereby avoiding center-based bias in modeling.

2. ***K*-fold**: The VA dataset was divided into K-folds in such a way that each fold was stratified by standard name. The K-1 fold of the data was used for training, and the remaining fold was for validation. This was repeated until all folds were validated. We performed 5-fold and 10-fold cross-validation to better capture the variance in data folds.

3. **Center-based**: The VA-ROQS dataset came from 40 ($n = 40$) different treatment centers. Data from 39 (n-1) centers were used for training, and one center's data was used for testing. We repeated this process until all centers were tested based on the model trained on the remaining n-1 centers.

Model Testing

Once the model is thoroughly validated and finalized, we need to test it on entirely new data (unseen by the model during training). We built a final model on the VA-ROQS dataset and tested it on the VCU dataset. One of the reasons we choose VA-ROQS for training and VCU for testing was to avoid any overlap of data between the training and test sets.

3.3.6 Evaluation Metrics

The performance of a model can be measured with different evaluation metrics. However, these metrics need to consider the class (structure labels) distribution to evaluate the model accurately. The dataset presented has a high level of class imbalance, as shown in Tables 1 and 2. Hence we evaluated the performance of each model using four distinct metrics—overall accuracy, macro-averaged precision, recall, and F_1-Score. Overall accuracy simply measures the percentage of OARs in the validation set classified correctly. These evaluation metrics were described in Section 2.5.

3.3.7 fastText Classification Algorithm

The fastText text classification algorithm [37] is an extension of the word2vec method, which includes three major steps. First, is generating the word vectors; fastText learns the vector representation of words from subwords (character NGram) [42]. For example, the word "Bladder" with a character NGram of 3 will have fastText representations such as "<bl, bla, lad, add, dde, der, er>" wherein < and > are added to indicate the beginning and end of the word. The technique of breaking the word into character NGram makes it work well with rare words. This helps to find the

vector representation of a word, even if it is not seen in training, and this is done by breaking down the word into character NGrams to get the word embedding. A subword size can be selected with range *minn* and *maxn*, indicating the minimum and maximum length of the subwords to generate. Along with these, fastText also considers *wordNgrams* (word NGram) to build the vector representation. Vector size is selected by setting the *dim* parameter. In Section 3.3.8, we explained the hyperparameter tuning process.

In the second step, word vectors are averaged to form a document vector, and in our method, it represents the vector representation of the complete RT structure. In the third and final step, it passes the averaged vectors through a shallow neural network with one hidden layer and uses the *softmax* function to generate the probability of a structure is one of the standard RT structures. Figure 9 shows the architecture of the fastText supervised classification algorithm.

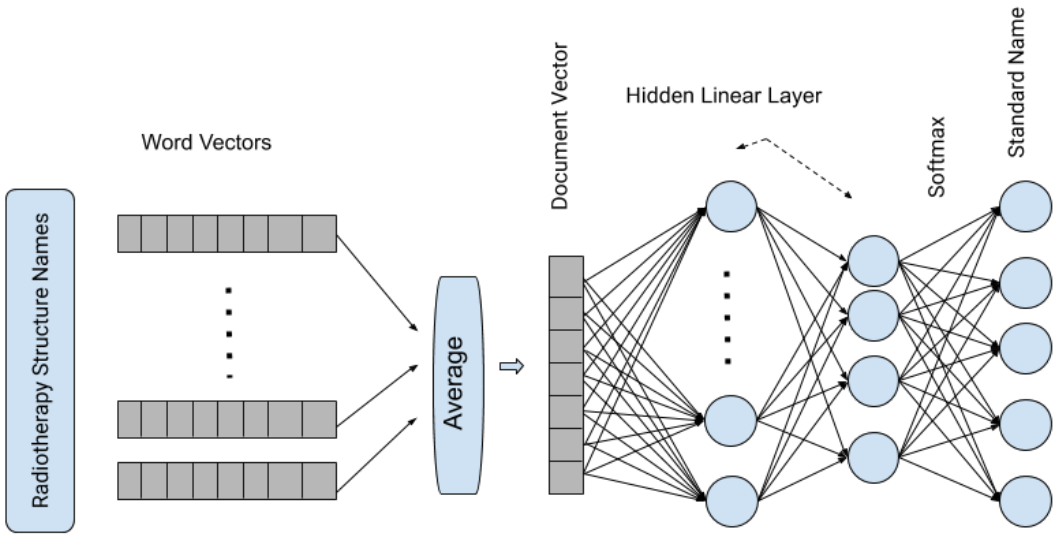


Fig 9: Pictorial representation of fastText supervised classification algorithm.

3.3.8 fastText Hyperparameter Tuning

After the initial selection of models, we chose fastText for further analysis, as it performed better than all other models. To further improve the model's performance, selecting appropriate hyperparameter values is important. The fastText algorithm has many hyperparameters, and we chose eight parameters to optimize, which have an impact on the data dictionary and model training. Out of eight hyperparameters selected for model tuning, two hyperparameters *minn* and *wordNgrams* were kept at fixed values. *wordNgrams* selects the number of consecutive individual words while building a data dictionary. Physician-given names are most likely to have less than three distinct words; to avoid considering the complete given name as a token, we set *wordNgrams* to 2. On the other hand, minn provides the minimum number of consecutive characters to consider as a token. We set *minn* to 2 to capture the more meaningful tokens rather than selecting every character as a token. Table 6 shows the hyperparameters and values tested.

A total of 15,360 combinations of hyperparameters was generated; each combination of hyperparameters was used to build a separate model for each disease type, and so considering the two disease types, overall we created 30,720 models. Models were evaluated with metrics described in Section 3.3.6 on the validation dataset and were recorded separately for each of the diseases types. Figures 10 and 11 show the impact of each hyperparameter on model performance. Boxplots are used to show the distribution of model performance (F_1-Score) for each value of the hyperparameter; the value with the smallest inter-quartile range and highest median was selected. The hyperparameter value was selected based on its performance on both disease type data (prostate and lung). The best values for selected hyperparamters are shown in Table 6 with brief descriptions.

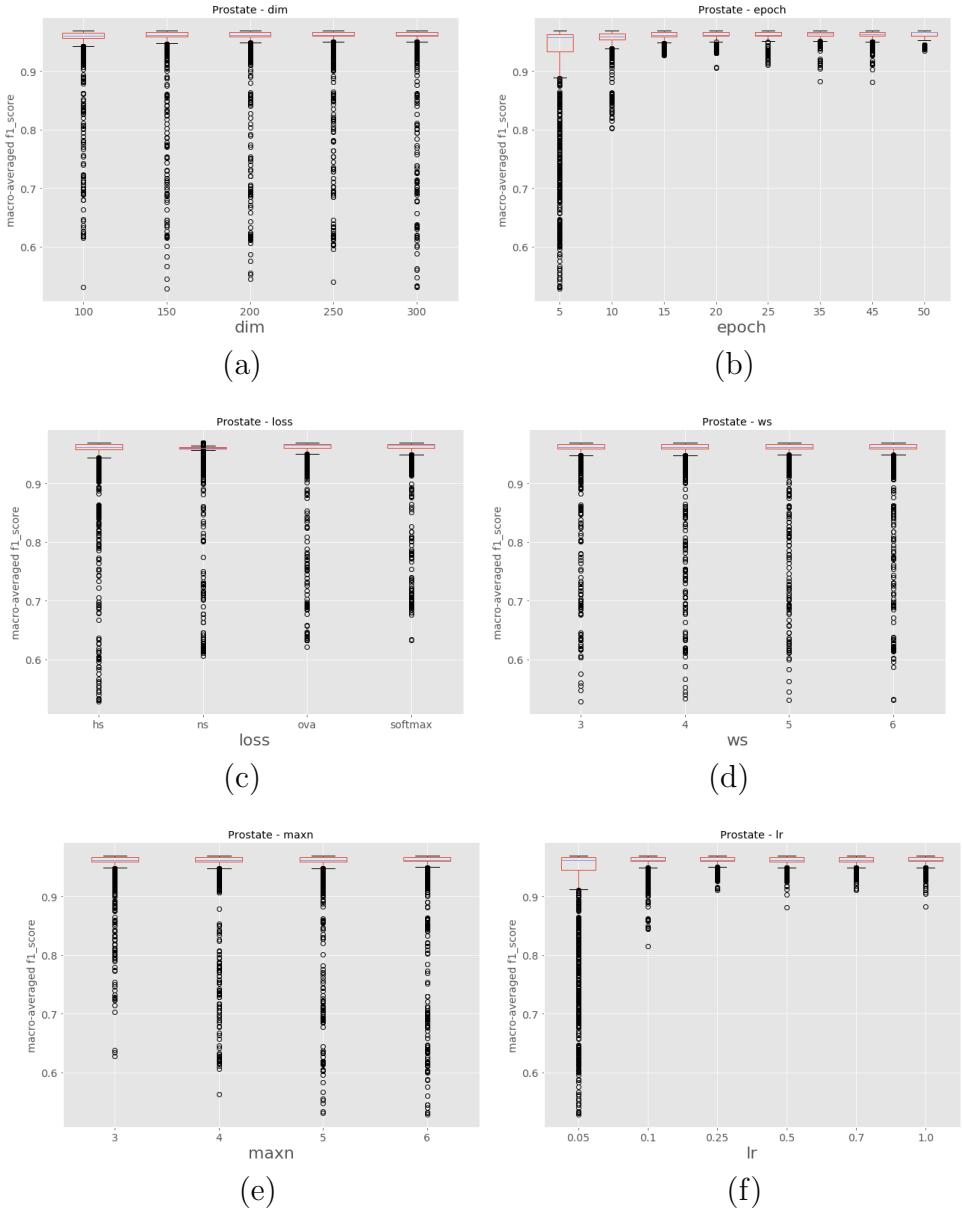


Fig 10: Hyperaparameter Tuning of fasttext for VA-ROQS Prostate cancer dataset. (a) dim: size of vector (b) epoch: number of times a model see's the all of the data while training, (c) loss, (d) ws: context window size (e) maxn: maximum length of character ngram (f) lr: learning rate.

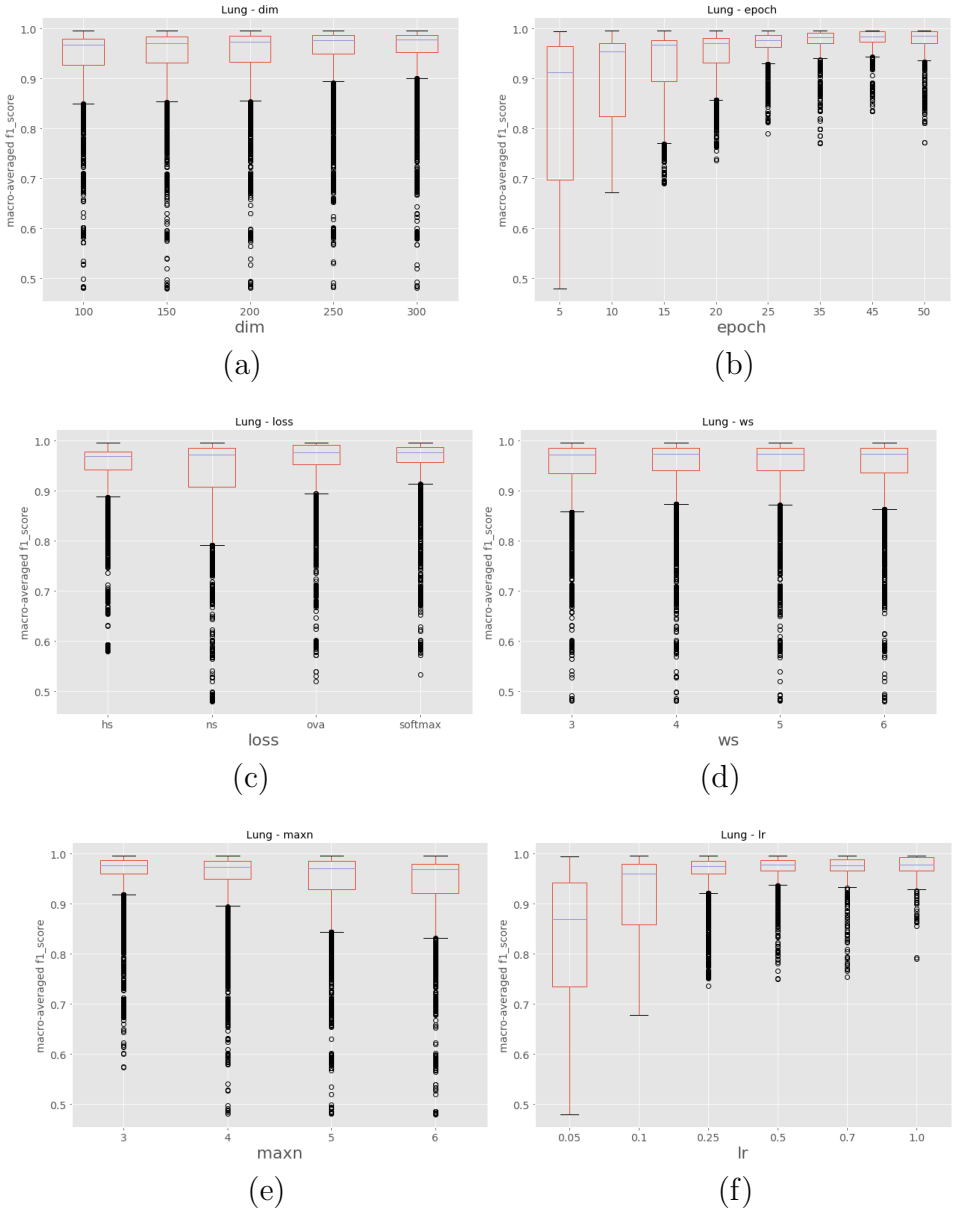


Fig 11: Hyperaparameter Tuning of fasttext for VA-ROQS Lung cancer dataset. (a) dim: size of vector (b) epoch: number of times a model see's the all of the data while training, (c) loss, (d)ws: context window size (e) maxn: maximum length of character ngram (f) lr: learning rate.

3.4 Results

In this section, we present the results of our models for both the VA and VCU datasets. We built models with combinations of feature sets, feature weighting methods, and machine learning algorithms. We observed that among all models, the fastText model performed consistently well on our data. Hence we present the detailed descriptions of results from only the fastText models. Results from the remaining models are shown in the Appendix B. The macro-averaged precision, recall, F_1-Score, and overall accuracy for both prostate and lung datasets for all the validation types are shown in Table 7. Individual class level results are shown in Tables 28, 30, 31, and 29 for prostate and Tables Tables 32, 33, 34, and 35 for lung in the Appendix B.

After fastText was selected as a final model, we tested the robustness of this method with four different validation types. Each of the validation types tested a different aspect of our model performance. Below we describe the results for each of these validation types.

3.4.1 Validation Results

70:30 validation: This validation type was chosen to test the model generalizability when data was split into 70% for training and 30% for testing. We split the data such that 70% of the patients from each center were under the training set and the rest of the patients from each center were under the testing set. We observed that our method was able to generalize well, and our model achieved overall macro-averaged F_1-Scores of 0.97 and 1.0 for prostate and lung datasets respectively. That indicates that our model was able to predict each label correctly. We also observed that our results were consistent across all classes regardless of class imbalance. Figures 12a and 13a show the class-wise results for prostate and lung data.

Parameter	Name	Optimal Value	Values Tested	Description
epoch	number of epochs	50	5, 10, 15, 20, 25, 35, 45, 50	This parameter is used to determine the number of times a model will see the entire dataset
lr	learning rate	1.0	0.05 0.1, 0.25, 0.5, 0.7, 1.0	This determines the step size taken at each iteration while moving toward a minimum of loss function
minn	minimum length of char NGram	2	2	minimum length of subword used to build word vector
maxn	maximum length of char NGram	6	3, 4, 5, 6	maximum length of subword used to build word vector
wordNgrams	maximum length of char NGram	2	2	Along with unique terms consecutive n-terms word vectors are generated
dim	size of the word vector	300	100, 150, 200 250, 300	In ML context word vectors are numerical representations of word. dim indicates the length of the representation
ws	size of the context window	3	3, 4, 5, 6	Word vectors are build in such a waythat it can predict the neighboring words in given text. It helps to encode the semantics of word. Window size indicates the range of words to predict.
loss	loss function	softmax	ns, hs, ova, softmax	A loss function is a measure of how good a prediction model does in terms of being able to predict the expected outcome.

Table 6: fastText hyperparameters and values tested for tuning the model.

Evaluation Type	Disease Type	Validation	Precision	Recall	F_1-Score	ACC
Validation (VA-ROQS)	Prostate	70:30	0.97	0.97	0.97	0.99
		5-fold	0.96	0.96	0.96	0.98
		10-fold	0.96	0.97	0.96	0.98
		VA Center	0.94	0.94	0.94	0.97
	Lung	70:30	1.00	0.99	0.99	1.00
		5-fold	0.98	0.98	0.98	0.99
		10-fold	0.99	0.99	0.99	0.99
		VA Center	0.93	0.93	0.93	0.99
Test (VCU)	Prostate	-	0.94	0.99	0.96	0.98
	Lung	-	0.83	0.89	0.86	0.96

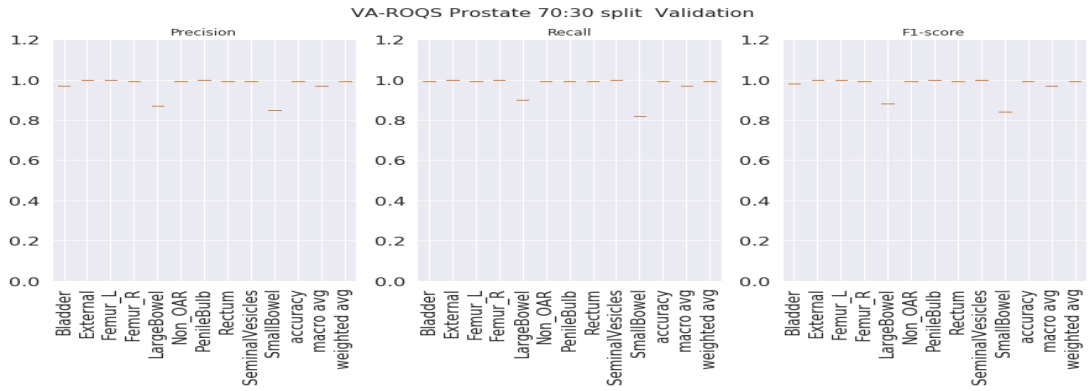
Table 7: Disease specific macro-averaged precision, recall, F_1-Score, and overall accuracy for validation and test sets.

***K*-fold validation:** With this validation type we checked the performance on the complete dataset. Here, we split the data into K-folds using a K value of 5. We observed that the 5-fold cross-validation achieved overall macro-averaged F_1-Scores of 0.96 and 0.98 for prostate and lung datasets respectively. Excellent results from 5-fold validation indicates that our model was able to generalize the overall data and not just on some random split of the data. We also repeated the same process for 10-fold cross-validation and observed that the model achieved similar results with 0.96 and 0.99 macro-averaged F_1-Scores for prostate and lung respectively. We chose to present the 5-fold results here, and the 10-fold cross validation results are presented in the Table 31 for the prostate and Figure 35 in Appendix B for the lung. It is important to see the consistent performance of each label in all folds. Figure 12b for the prostate and Figure 13b for the lung shows that our model has performed consistently well across all folds for each class and provided consistent performance.

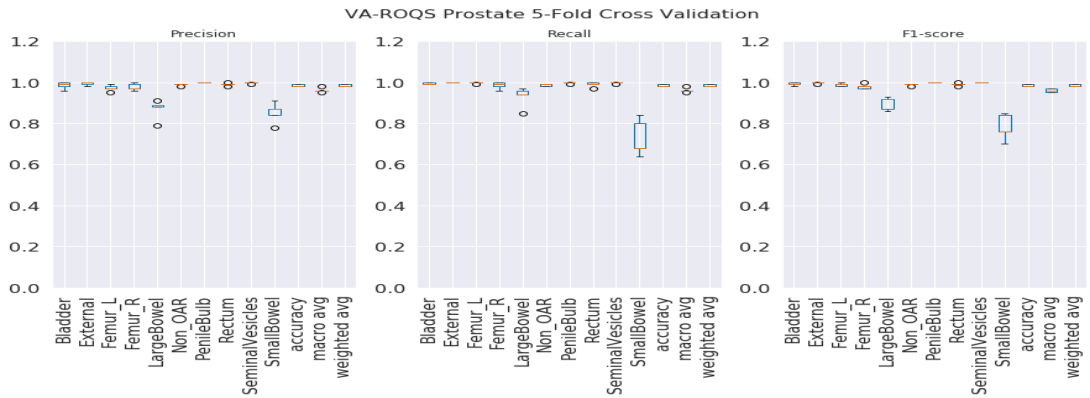
Center-based validation: VA has 40 radiation therapy centers. Even though they all are under one VA management, we believe that there are some differences in their practices. Each center operates as an individual institution at the practice level. In order to test this hypothesis, we trained the model on the data from 39 centers and tested it on one center and repeated this process until all the centers had been tested. We observed that the model achieved 0.94 and 0.93 overall macro-average F_1-Scores for the prostate and lung respectively. Although the model performed well, the performance dropped by 2% for the prostate and around 6% for the lung. This indicates that our model has high performance, but the inherent variance in structure naming practices at the different VA centers caused the model to make some mistakes, which lead to a decrease in performance when compared to the first two validation types.

3.4.2 Test Results

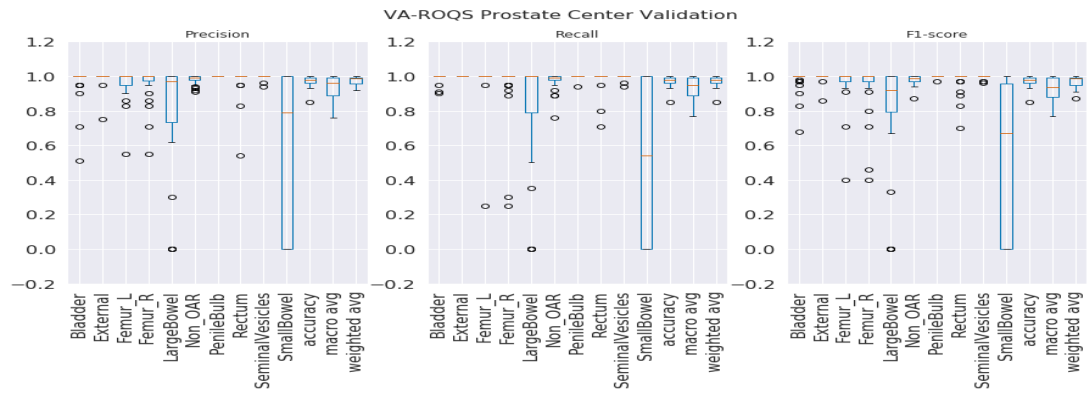
Once the model is finalized after thorough validation methods, it is imperative to check the model's performance on the unseen dataset. Here, the VCU dataset was used as a test set, which was never used in algorithm selection, model training, or validation. The final model was built with hyperparameters selected (see Section 3.3.8) on the entire VA-ROQS dataset. By using the VCU dataset as a test set, we were able to assess two aspects of our model. First, we checked the model's ability to generalize on the unseen data. Second, we checked the generalizability on a dataset from a different source. We observed that our model was able to predict the correct labels with high macro-averaged F_1-Scores of 0.96 and 0.86 for prostate and lung datasets, respectively as shown in Table 7. However, model performance dropped when compared to the model validation results, which indicates that although the model is robust, it is still affected by the change in the data source. We observed



(a)

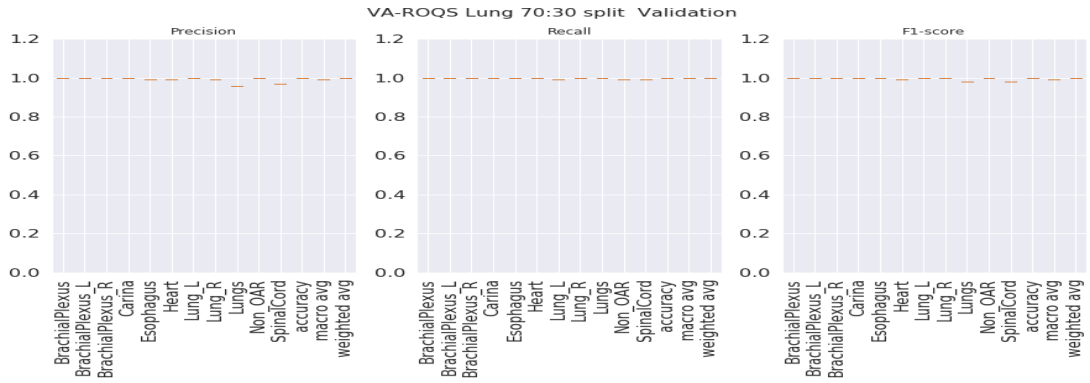


(b)

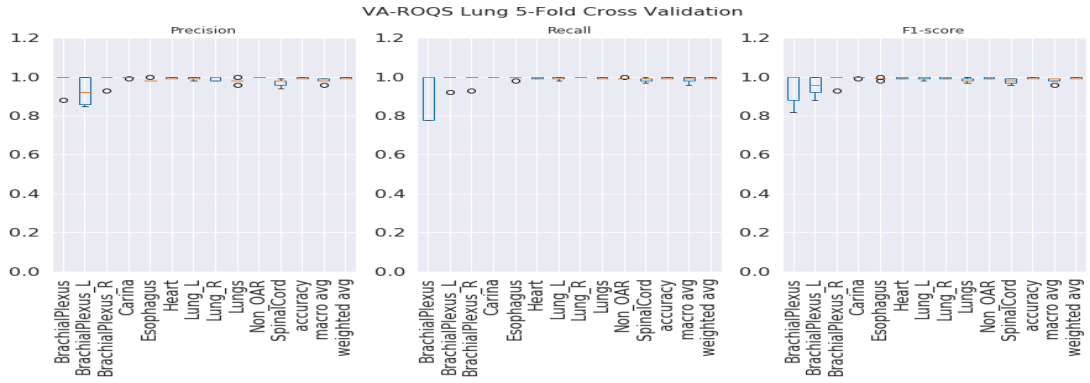


(c)

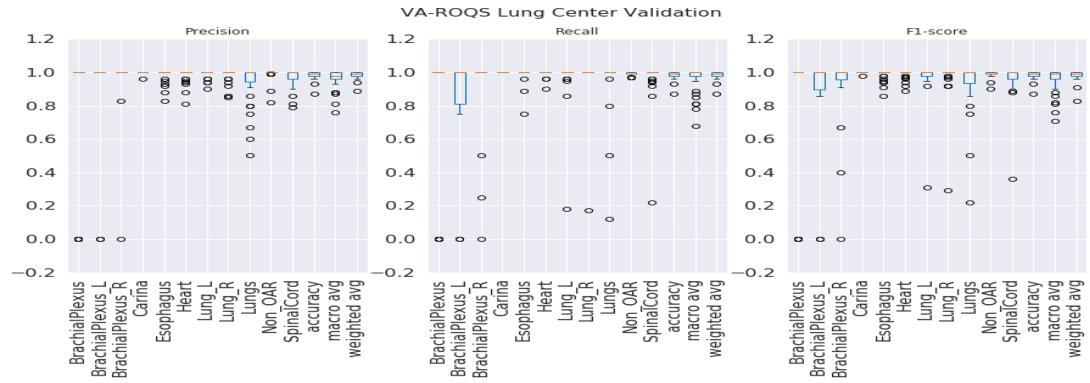
Fig 12: VA-ROQS prostate dataset—cross-validation results: (**a**) VA-ROQS 70:30 split cross-validation, (**b**) VA-ROQS 5-fold cross-validation, (**c**) VA-ROQS center based validation.



(a)



(b)



(c)

Fig 13: VA-ROQS lung dataset—cross-validation results: (**a**) VA-ROQS 70:30 split cross-validation (**b**) VA-ROQS 5-fold cross-validation (**c**) VA-ROQS center based validation.

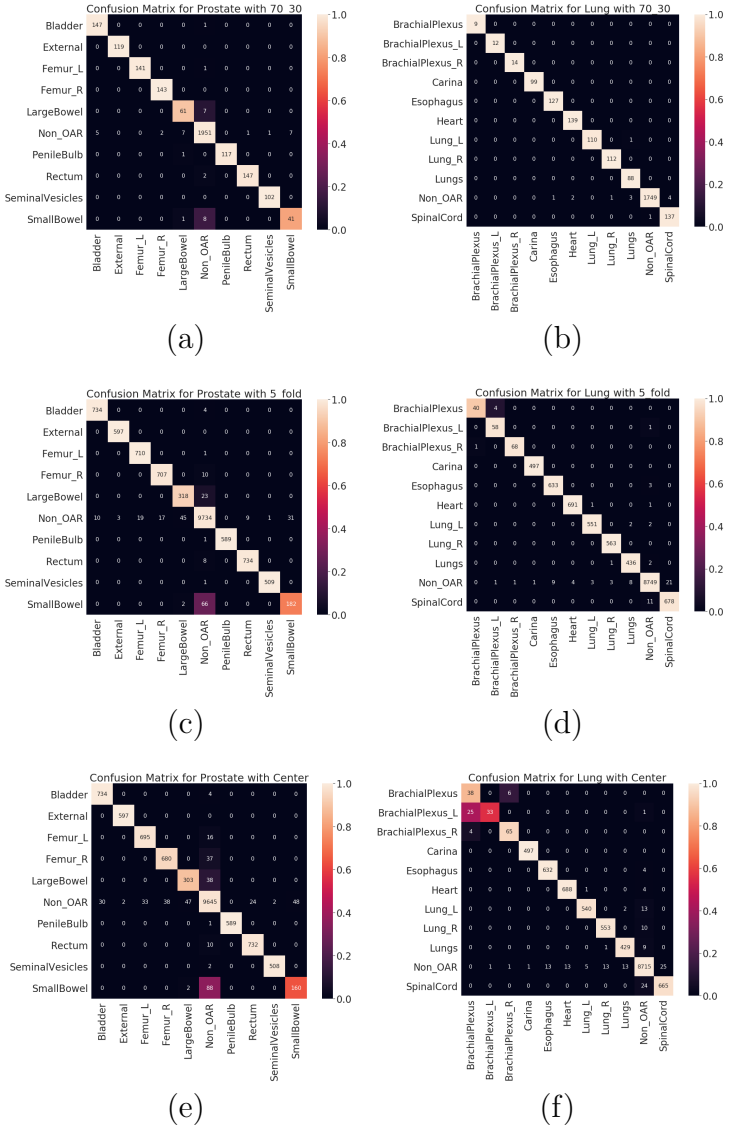
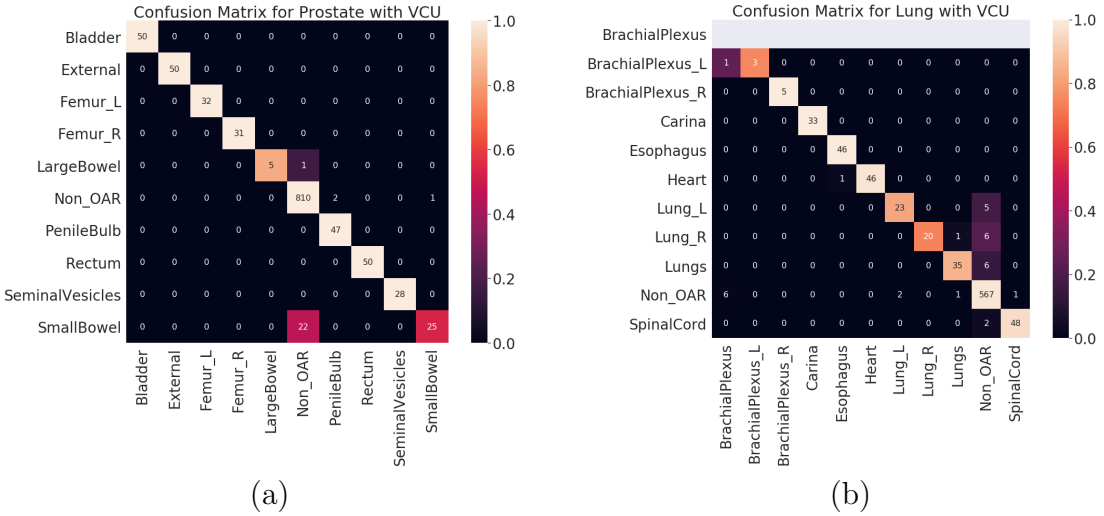


Fig 14: Validation set (VA-ROQS) confusion matrices of different validation types for both prostate and lung. (**a**) Prostate 70:30 split validation. (**b**) Lung 70:30 split validation. (**c**) Prostate 5-fold cross-validation. (**d**) Lung 5-fold cross-validation. (**e**) Prostate VA Center cross-validation. (**f**) Lung VA center cross-validation. Lighter color indicates better prediction. Diagonal indicates the correctly predicted labels.

a drop in overall macro-average F_1-Score due to the one OAR label *BrachialPlexus*; VCU dataset did not have any OARs labeled *BrachialPlexus* but our model predicted the *BrachialPlexus_L* as *BrachialPlexus*. Even if the number of samples is very few, macro-averaged metrics give equal importance to all labels and penalize the overall score regardless of the number of instances of labels in the dataset. Table 8 and Table 9 shows the class-wise results for prostate and lung data. We suspect that it is because VCU is an academic medical center, unlike the VA, and hence the structure-naming practices at VCU differ to accommodate the needs of academic hospitals.



(a) (b)

Fig 15: Test set (VCU) confusion matrices. (**a**) Prostate. (**b**) Lung. Lighter color indicates better prediction. Diagonal indicates the correctly predicted labels.

3.5 Discussion

The proposed radiotherapy structure name standardization methodology is system agnostic. Each of the validation types we presented on the VA-ROQS data demonstrates that our model is robust and works well to identify the correct TG-263 standardized names. We also tested our model with data from outside of the VA

Structure Name	Precision	Recall	F_1-Score	Support
Bladder	1.00	1.00	1.00	50
External	1.00	1.00	1.00	50
Femur_L	1.00	1.00	1.00	32
Femur_R	1.00	1.00	1.00	31
LargeBowel	0.83	1.00	0.91	5
Non_OAR	1.00	0.97	0.98	833
PenileBulb	1.00	0.96	0.98	49
Rectum	1.00	1.00	1.00	50
SeminalVesicles	1.00	1.00	1.00	28
SmallBowel	0.53	0.96	0.68	26
accuracy	0.98	0.98	0.98	1154
macro avg	0.94	0.99	0.96	1154
weighted avg	0.99	0.98	0.98	1154

Table 8: VCU Test Set results of Prostate structures.

Structure Name	Precision	Recall	F_1-Score	Support
BrachialPlexus	0.00	0.00	0.00	7
BrachialPlexus_L	0.75	1.00	0.86	3
BrachialPlexus_R	1.00	1.00	1.00	5
Carina	1.00	1.00	1.00	33
Esophagus	1.00	0.98	0.99	47
Heart	0.98	1.00	0.99	46
Lung_L	0.82	0.92	0.87	25
Lung_R	0.74	1.00	0.85	20
Lungs	0.85	0.95	0.90	37
Non_OAR	0.98	0.97	0.98	586
SpinalCord	0.96	0.98	0.97	49
accuracy	0.96	0.96	0.96	858
macro avg	0.83	0.89	0.85	858
weighted avg	0.96	0.96	0.96	858

Table 9: VCU Test Set results of Lung structures.

system (VCU dataset) which shows that our method works well for data from other institutions.

For the prostate RT structures, we observed that the majority of mistakes made by the model were in classifying *SmallBowel* and *LargeBowel*. This confusion is

attributed to the fact that the same name can be used for both anatomical structures. In Table 10, we can see that "*bowel*" is used to label both *SmallBowel* and *LargeBowel*.

In the VCU Lung dataset validation, accuracy and macro-average F_1-Score dropped when compared to the 70:30 split validation. This drop was caused by the misclassification of the lung and brachial plexus related structures, as shown in Table 11.

3.5.1 Error Analysis

Confusion matrices for all validation types on validation dataset (VA-ROQS) are shown in Figure 15 and for test dataset (VCU) in Figure 14. We performed an error analysis on the test set to understand our model's ability to generalize on unseen data. Error analysis provides insights into the reasoning behind the failure of the model prediction. In this work, a false positive is more expensive than a false negative. Although, this is a multiclass classification problem, wrongly predicted OAR is more expensive than a wrongly predicted Non_OAR. To this end, we divided classification errors into three main categories.

- Type I: When the structure was OAR but predicted as Non_OAR.
- Type II: When the structure was OAR but predicted as the wrong OAR.
- Type III: When the structure was Non_OAR but predicted as OAR.

Type II and III errors are expensive when compared to the type I error, as they produce false-positive OAR. Looking at the predicted and standard labels for physician-given names, we can infer that there is a pattern to errors for a few structures. Table 10 shows the errors made on VCU Prostate dataset. We observe that the majority of the errors come from Type I. The major error was due to the lack of signal in the text label. Just looking at the structure name "bowel" and inferring the "SmallBowel" or "LargeBowel" structures is difficult even for experts.

Error Type	Physician Given Name	TG-263 Standard Name	Predicted Name	Count
Type I	bowel	LargeBowel	Non_OAR	1
	bowel	SmallBowel	Non_OAR	22
Type II	nonptvpenilebulb	Non_OAR	PenileBulb	2
	small bowel	Non_OAR	SmallBowel	1

Table 10: Error analysis of VCU dataset prostate structure.

Error Type	Physician Given Name	TG-263 Standard Name	Predicted Name	Count
Type I	bilatlungs	Lungs	Non_OAR	5
	ptv	Lungs	Non_OAR	1
	lung-l	Lung_L	Non_OAR	1
	lung_l1	Lung_L	Non_OAR	4
	lung-r	Lung_R	Non_OAR	2
	lung_r1	Lung_R	Non_OAR	4
	spinal column	SpinalCord	Non_OAR	1
	spine	SpinalCord	Non_OAR	1
Type II	brachial_plexus	BrachialPlexus_L	BrachialPlexus	1
	esophagus	Heart	Esophagus	1
	lung	Lung_R	Lungs	1
Type III	ipsi_lung	Non_OAR	Lung_L	1
	left lung	Non_OAR	Lung_L	1
	brachial plexus	Non_OAR	BrachialPlexus	1
	brachial_plexus	Non_OAR	BrachialPlexus	2
	lung	Non_OAR	Lungs	1
	plexus	Non_OAR	BrachialPlexus	3
	t7 cord	Non_OAR	SpinalCord	1

Table 11: Error analysis of VCU dataset lung structure names.

In case of Lung, we see that there are many more Type II an III errors made by the model. Table 11 shows all the errors made on the VCU Lung dataset. We can see that majority of the errors were made while predicting the structures related to the lungs (Lung_L, Lung_R, or Lungs) and brachial plexus. For lung-related structures

we see that names containing numerical characters are most likely to be predicted as Non_OARs, as it is common for Non_OAR structures to contain numerical characters. For brachial plexus related structures, we can see that names containing "Plexus" are predicted as BrachialPlexus if there is no other information found to determine it as left or right BrachialPlexus. This also indicates the model errors due to the lack of signal in the input data. We also looked at the errors made by the model from holdout set (70:30 split) validation results. Tables 12 and 13 show the errors made on VA validation set for prostate and lung datasets respectively. We observed a similar pattern of errors for the prostate; the major confusion is between "SmallBowel" and "LargeBowel".

3.5.2 Comparison with Previous Works

Our work differs in many ways when compared to the most recent proposed approaches in the research community. Schuler et al. reported that their approach resulted in a 99% relabel rate [23], but it requires the mappings from the domain expert from the same institute where data are collected. In contrast, our method provides the same success rate with the added advantage of working on arbitrary physician-given names from multiple institutes. Our work is scalable and generalizable to the external dataset. Two other works proposed machine learning based structure name standardization using geometric information [26, 27]; both of those projects reported high accuracy. However, both of them did not use all the structures; instead they used only OARs. Our approach takes all possible structures into account and hence will work on real-world clinical datasets. However, due to the aforementioned limitations of the related work, it is not possible to perform a direct comparison between the accuracies from our approach and those from related work. It should also be noted that our proposed approach is the very first text mining based method

to automatically standardize arbitrary structure names from the DICOM dataset.

Error Type	Physician Given Name	TG-263 Standard Name	Predicted Name	Count
Type-I	bowel	LargeBowel	Non_OAR	5
	bowel large	LargeBowel	Non_OAR	1
	bowel, large	LargeBowel	Non_OAR	1
	bowel	SmallBowel	Non_OAR	6
	bowel (partial)	SmallBowel	Non_OAR	1
	bowel-ptv_sigm	SmallBowel	Non_OAR	1
	fem hd neck l	Femur_L	Non_OAR	1
	p bulb control	PenileBulb	Non_OAR	1
	rectum_om	Rectum	Non_OAR	1
	rectum_wm	Rectum	Non_OAR	1
	bladder min	Bladder	Non_OAR	1
Type-II	sigmoid	SmallBowel	LargeBowel	1
Type-III	sigmoid	Non_OAR	LargeBowel	5
	large bowel	Non_OAR	LargeBowel	1
	colon	Non_OAR	LargeBowel	1
	small bowel	Non_OAR	SmallBowel	4
	sm bowel	Non_OAR	SmallBowel	3
	whole_rectum	Non_OAR	Rectum	1
	bladder, nos	Non_OAR	Bladder	3
	bladder1	Non_OAR	Bladder	1
	femoral head r	Non_OAR	Femur_R	1
	femur r	Non_OAR	Femur_R	1
	vesicle bed	Non_OAR	SeminalVesicles	1

Table 12: Error analysis of VA-ROQS prostate structure names with **70:30** validation.

3.5.3 Limitations

Our proposed model has three limitations. Firstly, we are only predicting the identities of the OARs and labeling them with standard names. However, the target

Error Type	Physician Given Name	TG-263 Standard Name	Predicted Name	Count
Type I	total_lung	Lungs	Non_OAR	2
	spinalcanal	SpinalCord	Non_OAR	2
	cord_0	SpinalCord	Non_OAR	1
	esophagus-kl	Esophagus	Non_OAR	1
Type III	es	Non_OAR	Esophagus	1
	cord	Non_OAR	SpinalCord	1
	cord3	Non_OAR	SpinalCord	2
	l lung lymph	Non_OAR	Lung_L	1
	heart2	Non_OAR	Heart	1

Table 13: Error analysis of VA-ROQS Lung structure names with **70:30** validation.

(tumors) and PRVs are important structures and identifying and labeling them is also crucial for treatment delivery quality assessment. Secondly, we demonstrated that we can train on data from one institution and predict data from another. Our model is also language dependent, as it was trained only on structure names written in English. We believe the model pipeline will work for any language, but inter language models are only possible if they are trained on a mixture of languages. Thirdly, the ML pipeline from data preprocessing to prediction works as a standalone system. In the future, we plan to create a seamless enterprise informatics platform that can automatically collect data from the treatment planning systems and perform automatic structure name standardization on retrospective data.

3.6 Conclusion

In this chapter, we presented an ML approach to standardize the radiotherapy structure names using physician-given names. We observed that the fastText algorithm works best when compared to other feature weighting and classification algorithms. Our method was evaluated with the data from 40 VA radiotherapy cen-

ters and tested on an external dataset from VCU. We demonstrated that our method works well on multiple disease sites and is also generalizable. To the best of our knowledge, this is the first and the only model using the physician-given name to predict the TG-263 standard names using NLP and ML based methods. We also observed that our approach fails in certain conditions, when enough information is not available for the model to infer the correct label. Our approach can be augmented with other available information, such as geometric information of structures. We believe that the proposed structure names standardization methods can help with big data analytics in the radiation therapy domain using population-derived datasets, including standardization of the treatment planning process, clinical decision support systems, treatment quality improvement programs, and hypothesis-driven clinical research.

Contribution summary: In this chapter, we presented a text mining based approach for structure name standardization using physician-given names. Specific contributions of this chapter are as follows.

1. We present a machine learning approach to standardize the radiotherapy structure names that can automatically convert the arbitrary physician-given structure names to the domain wide TG-263 based nomenclature.

2. We demonstrate that a relatively small amount of data from each treatment center is enough to build a generalizable machine learning model, which a simple text mapping cannot achieve.

3. We establish that our proposed approach is disease site agnostic, i.e., it can be used on multiple disease sites.

4. We also demonstrate that physician-given names hold enough information about the structures that can be utilized to predict the standard names in TG-263.

5. Finally, we create a scalable approach that requires little to no preprocessing.

CHAPTER 4

MULTI-VIEW DATA INTEGRATION METHODS FOR RADIOTHERAPY STRUCTURE NAME STANDARDIZATION

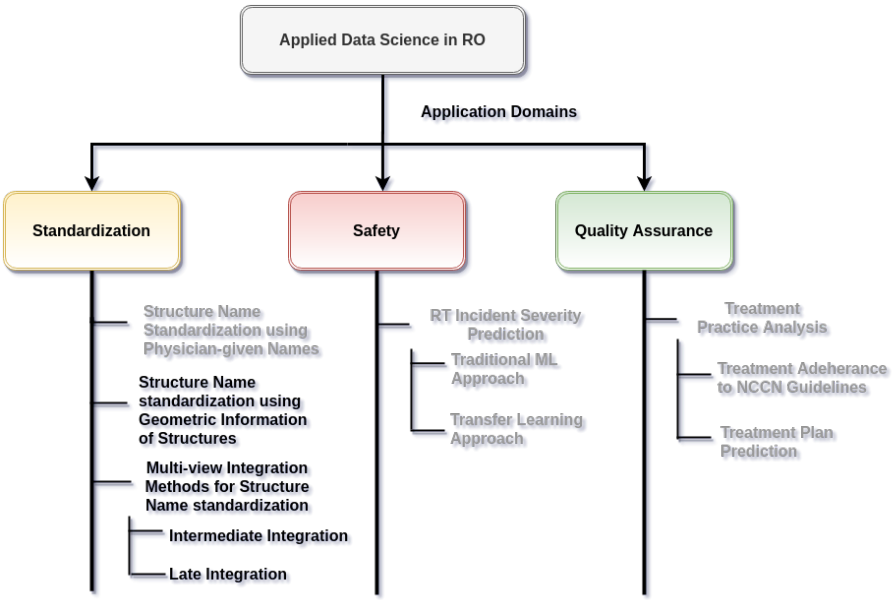


Fig 16: Thesis contribution chart, Chapter 4 contributions are highlighted.

4.1 Introduction

In Chapter 3, we presented structure name standardization using fastText document embeddings. Although our model had performed well, it made some wrong predictions. Our analysis showed that it is because of the use of the same labels for different structures. For example, some radiation oncologists used *Bowel* to label *SmallBowel*, and some had used it to label *LargeBowel*, which creates confusion when data from all patients is used to build the model. The use of the same name for

different structures may be because of the difference in naming practices at different VHA centers. To address such issues, we investigated the use of geometric information of structures for automatically identifying the standard structure names. It was evident from the results shown here that just geometric information may not be enough; hence we also investigated different approaches to integrate the textual labels and geometric information of structures. Such geometric information of structures provides a different view of the structures, which additionally helps in differentiating structures when physician-given names are the same. Since our datasets (views) are heterogeneous, we integrated the approaches at the intermediate and last stages of the machine learning pipeline.

4.2 Methods and Materials

4.2.1 Dataset

Dataset used in this chapter is the same as in Chapter 3 with the following two differences. First, Chapter 3's objective was to identify OARs when arbitrary structure names were given. In this chapter our objective is to identify "PTV" (target structure) along with OARs. Hence, the following prostate and lung structures were considered in this chapter:

Prostate Structures: Rectum, Bladder, Femur_L, Femur_R, LargerBowel, SmallBowel, PTV

Lung structures: Esophagus, Heart, SpinalCord, Brachial_Plexus, PTV

VA-ROQS Dataset: We have utilized the same VA-ROQS data used in Chapter 3, however we considered only structures required for the VA-ROQS project.

VCU Dataset: A new 50 set of patients for each prostate and lung cancer were

selected from VCU databases. These were entirely new patients with no overlap compared to the patients considered for VCU in "Structure Name Standardization using Physician-given names" (Chapter 3). The annotation process explained in Section 3.3.1 is followed to annotate the new dataset. Table 14 shows the distribution of lung and prostate structures for the VA-ROQS and VCU datasets.

4.2.2 Creation of Structure Set

Once a patient has been diagnosed with cancer and radiotherapy is prescribed as part of the treatment, a patient model needs to be created to determine the radiation dose to the target volume, OARs, and the coverage volume. To accomplish this purpose, imaging datasets are acquired. The most commonly used imaging dataset is Computed Tomography (CT), which provides tissue density information and the patient model information by rendering the patient's anatomy. A clinician will delineate the target/tumor region, OARs, and any other structures deemed necessary for the current case on this acquired dataset. This delineation is usually done within the TPS software, which will allow for the creation of the dose delivery treatment plan.

Figure 17 shows the axial, coronal and sagital cut sections of a prostate cancer CT, overlaid with several delineated structures. The imaging and structure set information is in the Digital Imaging and Communications in Medicine (DICOM) format which is the industry standard for the storage and transmission of medical imaging data. This data is traditionally stored as slices on the axial axis but can be rendered on any axis. A clinician will delineate any necessary structures using the delineation tool-sets in the TPS software; often by adding individual points or by using a free-hand drawing tool to create a closed polygon. For a given structure, this process is performed on each imaging slice until the delineation is complete.

	VA-ROQS		VCU	
	Non Standard Name		Non Standard Name	
Standard Name	**Structure Count**	**Unique Count**	**Structure Count**	**Unique Count**
Brachial_Plexus	108	44	4	4
Esophagus	613	26	47	3
Heart	670	20	45	2
Other (Lung)	10,292	3,639	775	317
SpinalCord	681	37	48	6
PTV (Lung)	680	286	36	4
Lung Total	**13,044**	**4,052**	**955**	**336**
Bladder	609	10	50	3
Femur_R	700	62	29	14
Femur_L	694	59	29	13
Rectum	719	14	50	3
SmallBowel	250	40	49	10
LargeBowel	341	34	0	0
Other (Prostate)	11,038	2,799	980	434
PTV (Prostate)	714	236	38	16
Prostate Total	**15,065**	**3,254**	**1,225**	**493**
Grand Totals	**28,109**	**7,306**	**2,180**	**829**

Table 14: Lung structure type distribution in VA-ROQS and VCU dataset.

For the same patient, Figure 18 shows the planning target volume (PTV) (green) and multiple planning related structures (red). The PTV represents the region that will be receiving the prescribed radiation dose. It is also common to have other structures that are very similar to the PTV as presented here, and may include a clinical target volume (CTV), gross tumor volume (GTV), or expansions of the

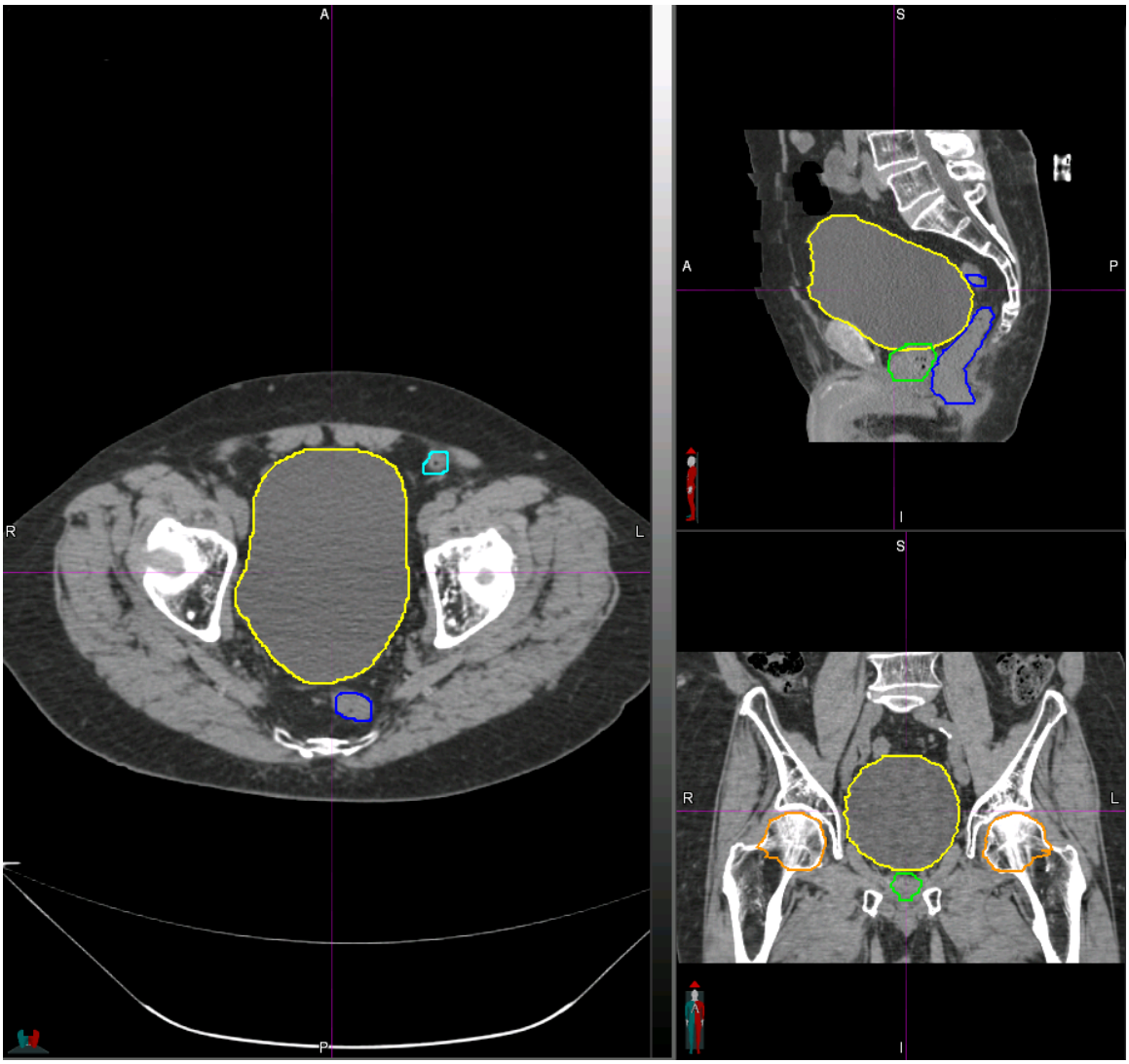


Fig 17: Planning CT from a prostate cancer patient with the following delineated structures: Bladder (yellow), Rectum (blue), Left and Right Femurs (orange), Small Bowel (aqua), PTV (green).

PTV. Also presented in this figure are rings, used for helping to guide the TPS dose optimization process, and implanted marker seeds.

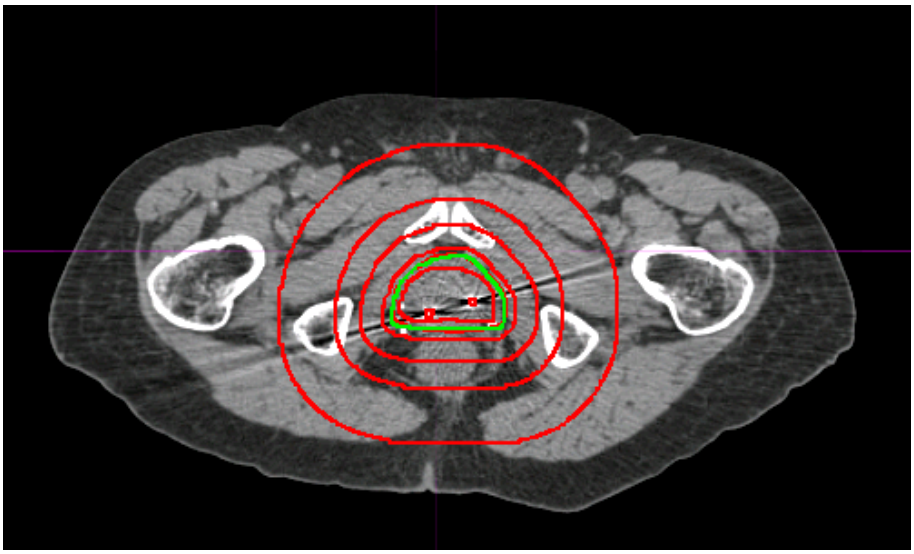
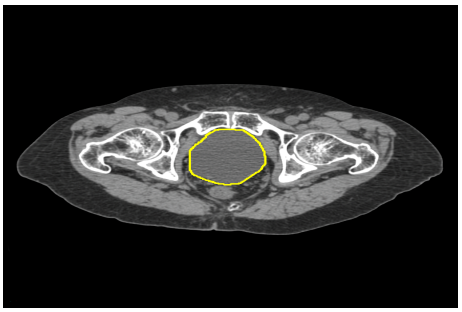


Fig 18: The PTV (green) and multiple other planning related structures (red) delineated on a planning image. These planning structures include rings, implanted seeds, and several interpretations of the tumor volume.

4.2.3 Data Preprocessing

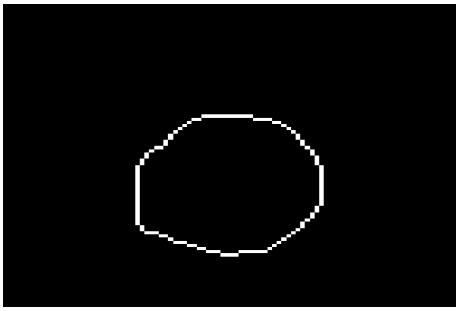
Textual Data Preparation

Structure names are short and have a limited character set to use, and the available character set is vendor dependent. As shown in Table 3, even though there is high variability in physician-given structure names for most of the structure types, the character set used is limited. Preprocessing methods need to be selected to ensure that critical information is retained; losing the information might negatively affect the ability to standardize the structure names with high fidelity. Hence, we decided to keep the preprocessing of physician-given names to a minimum by just converting them to lower case.

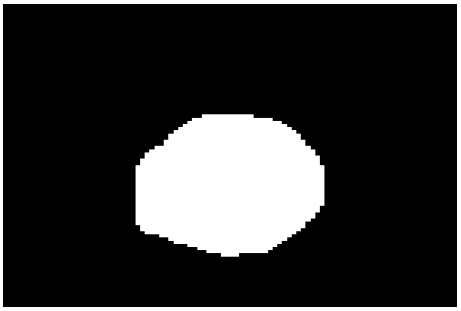


(a) A transverse slice of the original planning image with the bladder structure shown in yellow.

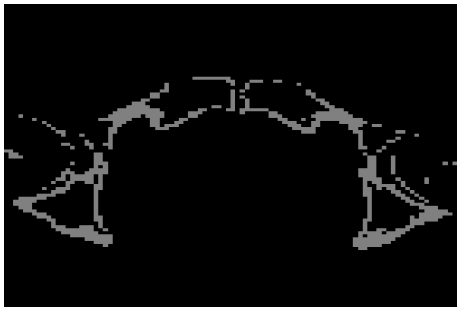
(b) Polygon points from the DICOM bladder structure set delineation. These individual points are interpolated on to the standardized bitmap volume.



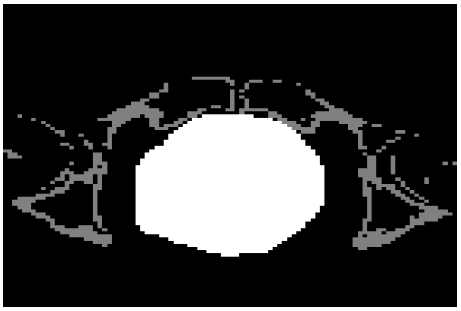
(c) Each sequential point is connected to form a close polygon.



(d) The close polygon is flood filled to create a solid structure.



(e) A density threshold is applied to the planning image such that only voxels that belonged to bony anatomy remain.



(f) Structure set (white) and bony anatomy (grey) data shown together with the same frame of reference.

Fig 19: Workflow for creating structure set and bony anatomy bitmaps for feature vector creation.

Geometric Data Preparation

The following process, shown in Figure 19, takes the DICOM structure set and imaging data and converts it into features vectors to be used as input for the classification algorithms. Figure 19a shows an original DICOM planning image and its associated structure set. The bladder structure delineation is shown in yellow.

Since the planning images available in our dataset did not have consistent voxel count, voxel resolution or origins, a standard grid was needed so that all structure sets could be stored in a consistent manner. The standard grid chosen for this purpose was 96 x 96 x 48 voxels and with a voxel resolution of 2mm x 2mm x 3mm. These parameters were chosen by manually inspecting a number of bitmap examples from each structure type to verify that the bounding box was large enough to cover the structures of interest. It should be noted that this manual step is needed only once as all structure volumes will be interpolated in the same bounding box. Future work is required to determine if such a one-size-fits-all based solution is sufficient, especially considering large structures like the entire lungs. Each original planning image and structure set was programmatically shifted such that the geometric center of the given structure was aligned to the geometric center of this standardized grid.

The automated workflow for creating feature vectors from the imaging and structure set data is demonstrated using one prostate patient as shown in Figure 19a. For each individual structure in the dataset, an empty three-dimensional bitmap object was created with the standard grid dimensions as defined above. Each polygon point in the DICOM structure set is mapped to its corresponding voxel in the new bitmap with a value of 1 as shown in Figure 19b. Then for each transverse slice of the bitmap, the sequential points were connected with new line segments which results in one or more closed polygons per slice as shown in Figure 19c. A flood fill algorithm [43]

was then run on each closed polygon to set all interior values to 1 resulting in a solid bitmap structure shown in Figure 19d. Voxels belonging to the structure in question would then have the value of 1 and all other voxels would remain as 0. The procedure used for generating these bitmaps was derived from the Research Computing Framework package [44].

In addition to the structure set data, imaging data was also used to add spatial context to the location of each structure in the human anatomy. A density threshold was applied to each planning image such that voxels with Hounsfield units (HU) above 1,300 were set to 1 and all others set to 0, leaving only the bony anatomy. While bone density starts around 1,050 HU [45], we have chosen a slightly higher value to focus on the gross skeletal structure and reduce noise from borderline tissue. The resulting bony image was then interpolated to the same standardized grid used by the structures so that both data types were properly aligned. Figure 19e shows just the bony anatomy and Figure 19f shows the bony anatomy and structure set data combined.

To create feature vectors, the 96 x 96 x 48 bitmap object was stretched out into a 442,368 x 1 vector by simply creating an array of each voxel value with increasing x, y, z axis indices. From this bitmap creation process, two datasets per disease site were created: *Without Bones* and *With Bones*.

- ***Without Bones***: The feature vectors were created with only structure set data as shown in Figure 19d. The total length of the feature vector is 442,368.
- ***With Bones***: The feature were created by appending the *No Bones* feature vectors with the bony anatomy data as shown in Figure 19f. The total length of the feature vector is 884,736.

Very long feature vectors make the model training phase slow and susceptible

to the *Curse of Dimensionality* [46]. One popular solution to this problem is to perform feature reduction by either removing features that are not strongly influencing predictions or condensing multiple features in such a way that still preserves importance. We have chosen to use truncated singular value decomposition (SVD) as it uses much smaller matrix multiplications when compared to methods like principal component analysis (PCA) or standard SVD [47]. This approach can approximate the input $m \times n$ matrix as $[m \times k] \times [k \times n]$ where k is the numerical rank [48]. When testing both methods, the truncated SVD ran faster and required less memory while still producing an explained variance within 0.1% of the result from PCA.

Figure 20 shows the explained variance of the disease sites for the *Without Bones* and *With Bones* datasets. All cases show a similar pattern and the cumulative variance curves start to flatten out around 100 features. For that reason, we have chosen 100 as the number of SVD features to use in our experiments as increasing the explained variance by more than a few percent would require at least doubling the total number of features. Initial tests using up to 1,000 SVD features did not improve classifier accuracies (data not shown). The anonymized patient identifier, physician specified label, and the TG-263 standardized label for each structure were added as features, not for model training, but for patient filtering and assessing the model accuracy.

This pipeline can be fully automated allowing for the processing of large Radiation Oncology datasets. While the disease specific bounding boxes should be set manually, it only needs to be done once while all the other modules are done programmatically.

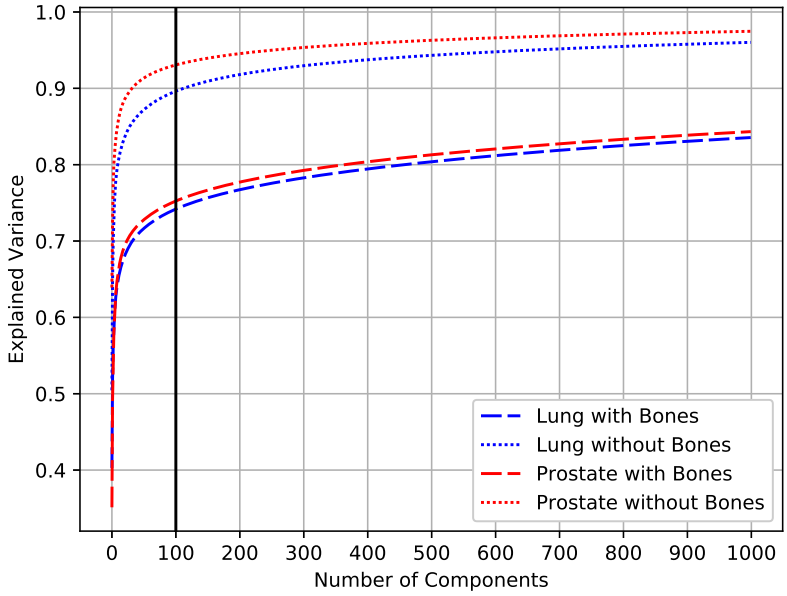


Fig 20: Cumulative explained variance from the number of features created by the SVD process. We have chosen the top 100 features in all models.

4.2.4 Model Selection

4.2.4.1 Single-View

Dataset used in this work is heterogeneous in nature. To properly compare the advantages of utilizing the multi-view heterogeneous data, we built the best possible model with single-view separately. In our previous work, we have thoroughly investigated the different algorithms for standardizing radiotherapy structure names with physician-given names [3] and geometric information [4]. Single-view model selection details are as below.

- Text data (Physician-Given Structure Names): We built structure name standardization models using the combinations of different feature extraction techniques, feature-weighting, and ML-algorithms. We tested NGram (uni-gram,

bi-gram, and tri-grams), character NGrams, and word embedding techniques for feature extraction. For feature weighting, we tried with term presence (tp), term count (tc), term frequency (tf), and term frequency-inverse document frequency (tf-idf) techniques. Finally we compared the six different ML classification algorithms —SVM-linear [15], SVM-RBF [16], k-nearest neighbors (KNN) [18], logistic regression [38], random forest [20], and fastText [37]—for initial model selection. All models were built by using scikit-learn machine learning library in python [39]. Finally, we selected the fastText algorithm for automatically identifying the standard structure names using the physician-given names because it had highest F_1-Score in comparison with other algorithms.

- Image data (3D geometric information of structures): In our prior work, we have also investigated the radiotherapy structure name standardization using geometric information. In order to extract geometric information, we converted the geometric information into binary vectors and selected top 100 components with SVD algorithm. After thoroughly evaluating different algorithms, we used the RF classification algorithm to build our final model because it provided the best F_1-Score.

4.2.4.2 Intermediate Integration

Intermediate Integration involves transforming the all view data into similar feature space and combining them (concatenating) into one. We utilized different techniques to transform them into a similar feature space as discussed below.

- Image Data Transformation: We used truncated singular value decomposition (SVD) as it uses much smaller matrix multiplications when compared to methods like principal component analysis (PCA) or standard SVD [47]. When

testing both methods, the truncated SVD ran faster and required less memory while still producing an explained variance within 0.1% of the result from PCA. We used the first 50 features from this feature set.

- Text Data Transformation: We used fastText algorithm to generate the embeddings (numerical representation) of size 200 for each physician-given structure name.

Thus, a final vector of size 250 is generated by concatenating feature vectors from each view (image and text). This vector is fed into the ML algorithm. We chose SVM with linear kernel to build the final classification model. Figure 21 shows the pictorial representation of our proposed intermediate integration method.

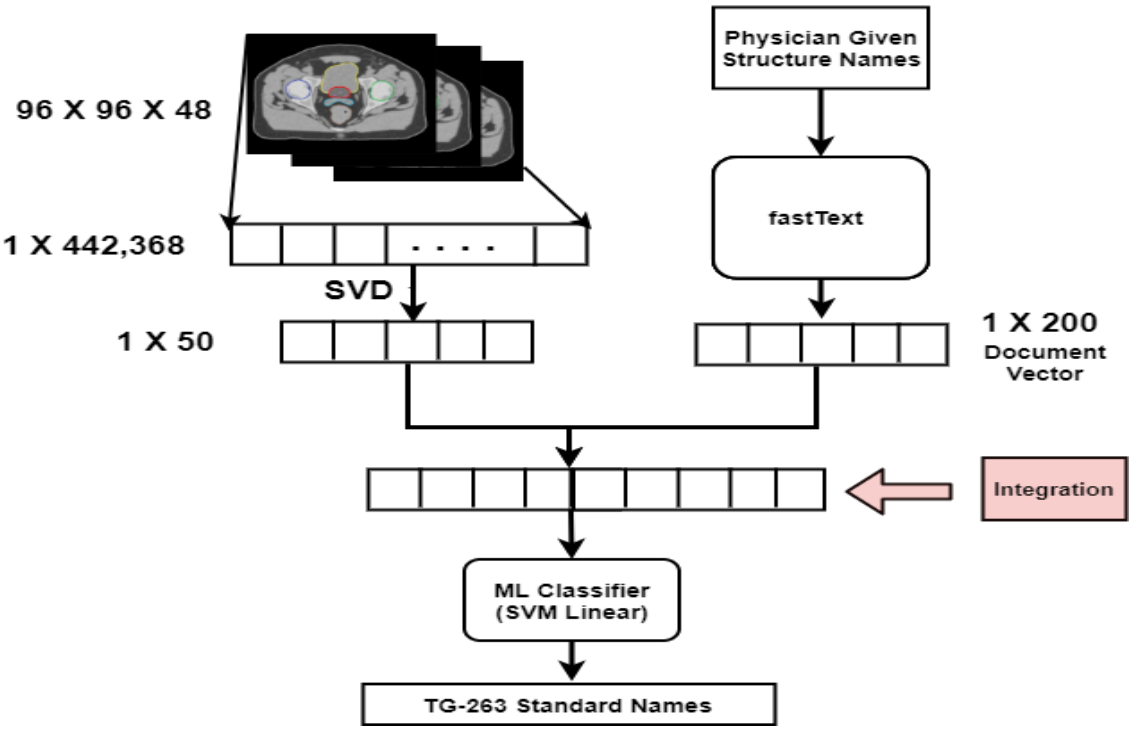


Fig 21: Intermediate stage integration method for structure name standardization.

4.2.4.3 Late Integration

In late integration, each view is analyzed separately and the results are then integrated to generate the final result. It is also known as model based integration. Integrating at a late stage has an advantage over other types of integration; the best algorithm can be selected to build a model based on the data type and each model can be run in parallel. In this work, we used different algorithms to build the models for each view. A prediction probability vector is generated for each sample from each model instead of a class prediction. The size of the prediction probability vector is equal to the number of classes in the dataset; in this scenario, it is eight classes for prostate and six classes for the lung dataset. The result vector from each view is then combined to generate the final prediction probabilities. These prediction probabilities are used for the final class prediction.

We used two techniques to combine the prediction probabilities from each view.

- Average (AVG): We created the final prediction probability vector by adding element-wise from each view and dividing it by the number of views. The final class is selected whose AVG probability is the highest.

- Maximum (MAX): In this technique, we selected the maximum probability from each view such that the resulting vector contains the maximum for each class from all the views. The final class is predicted by selecting the class from this resultant vector with the highest probability.

4.2.5 Model Evaluation

An essential part of building a machine learning system is to demonstrate its quantifiable generalizability. For example, the critical goal of a machine learning classification algorithm is to create a learning model that accurately predicts the

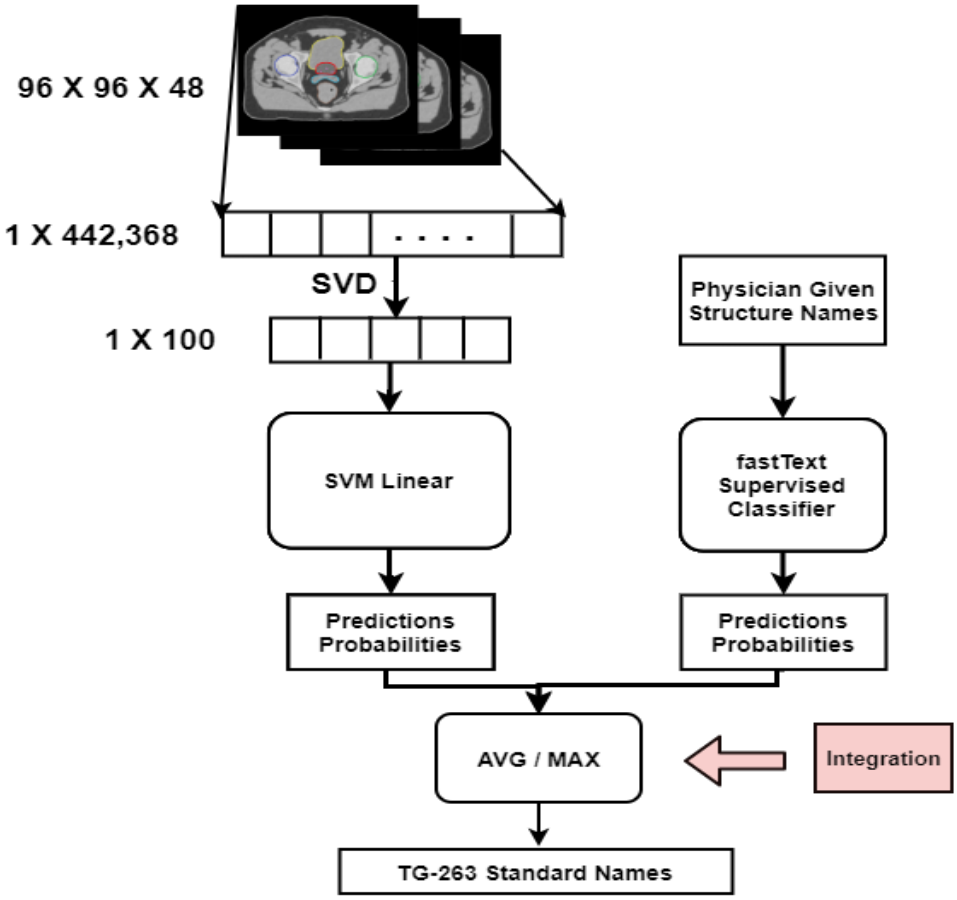


Fig 22: Late integration method for structure name standardization.

class labels of unseen data samples; this ensures that the machine learning model should work well for classifying future data.

Model validation is another important step in the machine learning process as explained before and we again used k-fold cross-validation. We divided the VA-ROQS dataset into training and testing datasets. We randomly selected data from 30 centers for training and remaining 10 centers data are for testing. We further divided the VA-ROQS dataset into 70:30 ratio as training and validation sets. Along with VA-ROQS testing dataset, we tested with the VCU dataset as an external validation dataset as before.

Model Validation

- **Hold-out set validation**: The VA-ROQS dataset was divided into a 70:30 ratio as the Training and Validation sets. The split is stratified by TG-263 standard names, which ensures that an equal percent of data is taken from each standard name for training, validation, and testing, thus avoiding center-based bias in modeling.

- **VA Center Based Cross-validation**: The data from randomly selected 30 VA-ROQS centers is used to validate the data from each center separately. Data from 29 (n-1) centers were used for training, and the remaining one center data for validation. We repeated this process until all centers are validated.

Model Testing

Once the model is thoroughly validated and finalized, we need to test it on entirely new data (unseen by the model during training). We built a final model on the data from 30 VA-ROQS centers and we tested it with two datasets: VA-ROQS test set (data from 10 centers) and the VCU dataset.

- **VA Center Based Test:** The data from randomly selected 30 VA centers is used for training and 10 centers for testing. We tested each center separately and results are reported to show the generalizability of model across multiple centers. We used the data from 10 VA centers as a test set to show that our model is able to predict the labels correctly from multiple centers.

- **VCU Test:** We used data from 30 VA centers for training the model and tested it on the VCU dataset. This model testing with VCU dataset shows the our model's ability to generalize on a completely external dataset.

4.3 Results

In this section, we present the results of our experiments. The results are divided into three subsections: Singe-view, Intermediate Integration, and Late Integration results.

4.3.1 Single-View

In the single-view approach, we built two separate models with physician-given structure names and geometric information of structures. We observed that the model utilizing the structure names consistently out performed the models built utilizing geometric information. Table 15 and 16 shows the model performance for the VCU and VA-ROQS datasets. We observed that the text based model has precision of 0.778 for VCU prostate dataset and 0.855 for VCU lung dataset. Figures 23 and 24 shows the confusion matrix for both VCU and VA-ROQS dataset. VCU prostate dataset has no instances of "large bowel" structures in dataset, but model predicted the "large bowel" for three structures. A macro-averaged metric penalizes equally regardless of number of samples in each individual class. In the VCU lung dataset there are only four "BrachialPlexus" structures but our model predicted the 9 false positives.

4.3.2 Intermediate Integration

In this method, we transformed the structure names and geometric information into similar feature space. We applied truncated SVD and selected top 50 components, and structure name word embedding of size 200 using fastText algorithm. These two features space from different view are then concatenated for training and testing. We trained the SVM with linear kernel to build a classifier with this combined dataset.

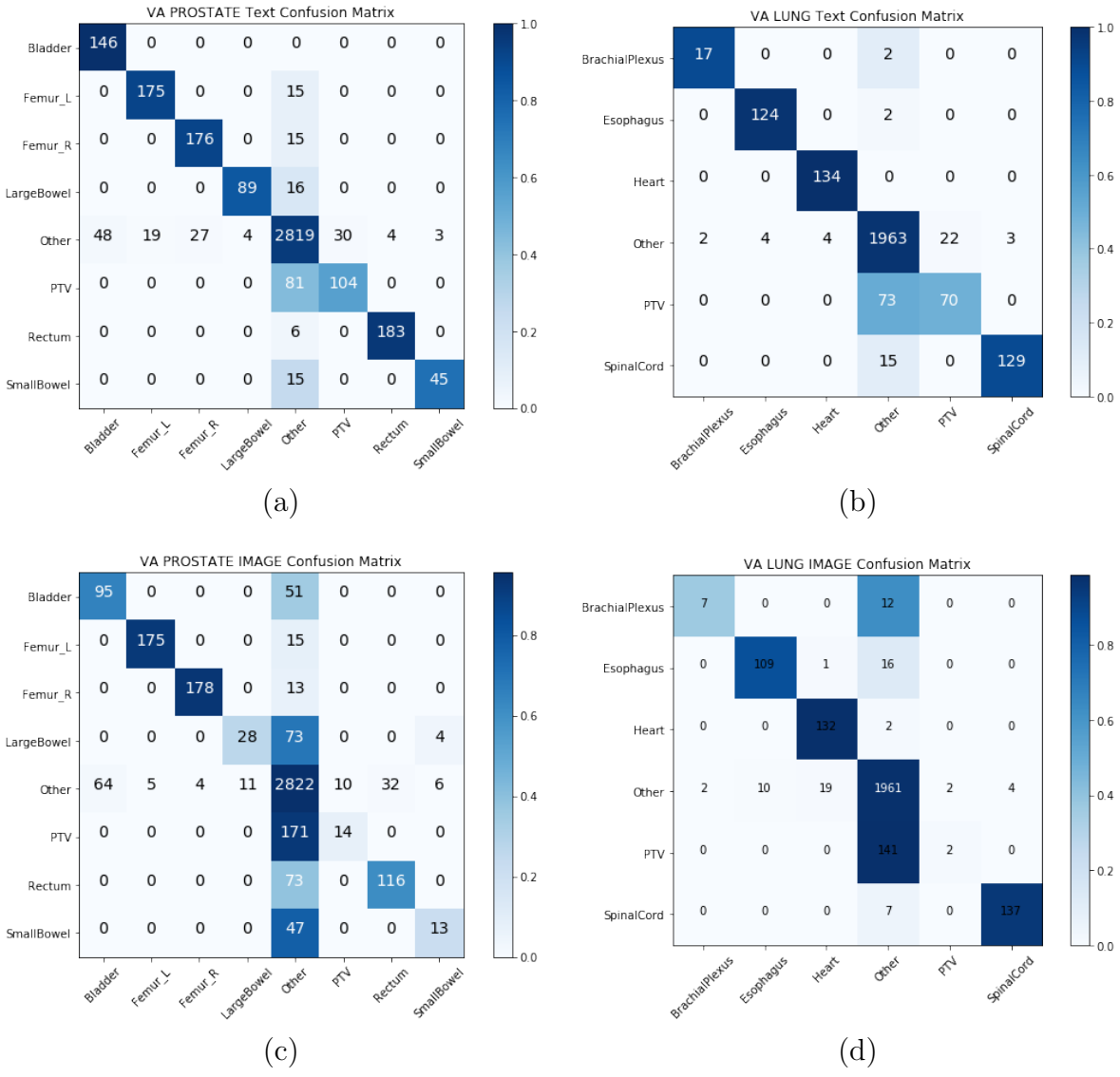


Fig 23: Single View Results: (a) VA-ROQS Prostate Text Based featuress (b) VA-ROQS Lung Text features. (c) VCU Prostate Image feature (d) VCU Lung Image features. Darker color indicates better prediction. Diagonal indicates the correctly predicted labels.

Table 15 shows the macro-averaged precision, recall, and F_1-Score for intermediate integration. We observed that our intermediate integration method performs better

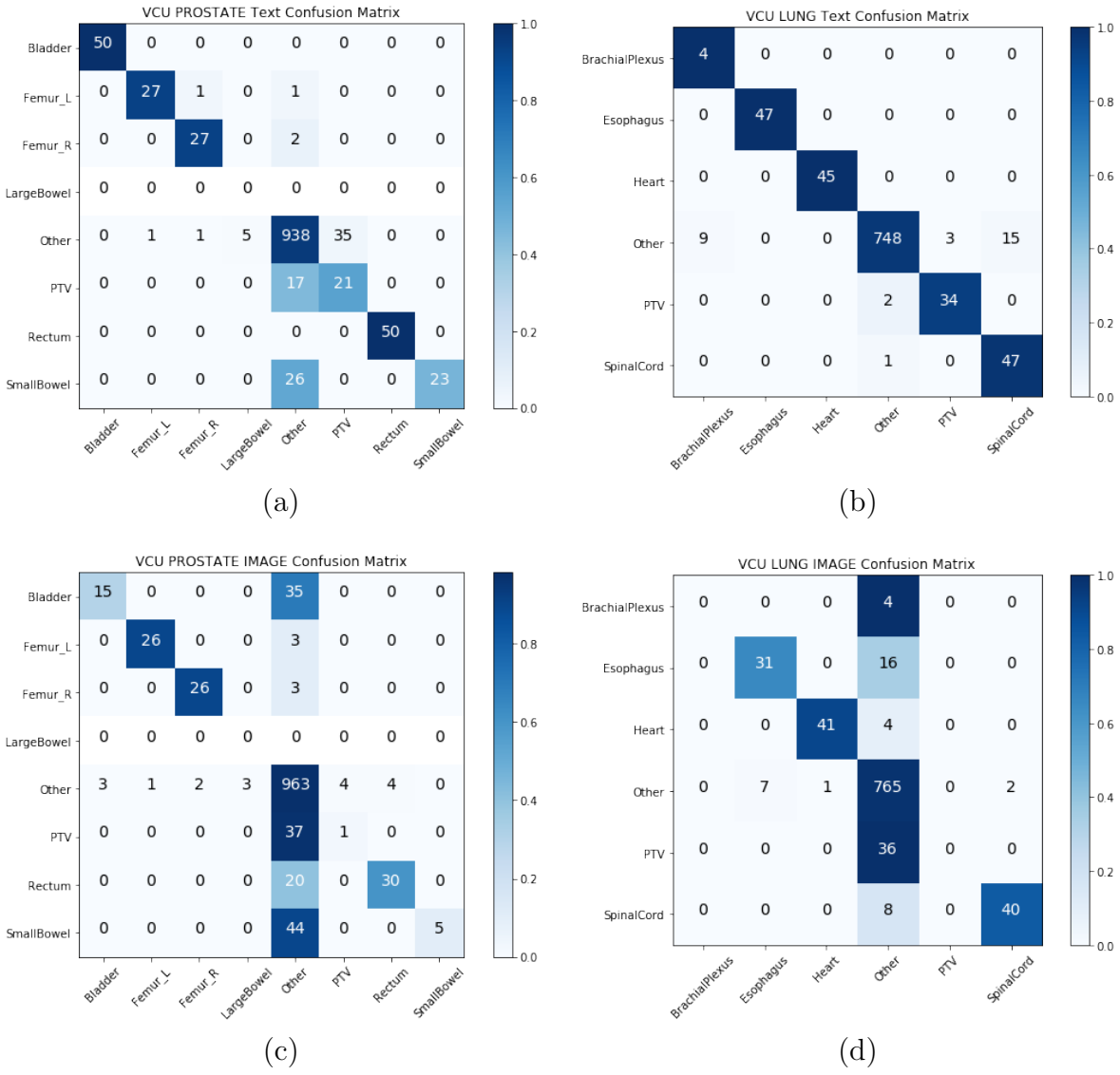


Fig 24: Single View Results: (a) VCU Prostate Text Based features (b) VA-ROQS Lung Text features. (c) VCU Prostate Image feature (d) VCU Lung Image features. Darker color indicates better prediction. Diagonal indicates the correctly predicted labels.

on VA-ROQS and VCU datasets. Precision is higher than the single view models for three out of four datasets; the overall F_1-Score is also higher for the VCU prostate

and lung datasets. Increase in precision indicates that the model can predict fewer number of false positives for OAR and target structures. Figure 27 shows the confusion matrices for prostate and lung structures in VA-ROQS and VCU datasets. We can observe that the Intermediate integration method consistently reduces the false positives for all OAR and target structures and increased the false positives in the other structures.

4.3.3 Late Integration

Table 16 shows the macro-averaged precision, recall, and F_1-Score for the proposed late integration method. We noticed that in the late integration method with MAX probability selection, the precision is better than the single view models for both prostate and lung VCU dataset. However, the recall and F_1-Score dropped. We also observed that using the MAX scores on the VA-ROQS prostate dataset, precision is increased by 0.07 but recall and F_1-Score are negatively affected. Overall, the late integration with MAX scores exhibited a negative affect on the VA-ROQS dataset. Figure 26 and 27 shows the confusion matrices for the lung and prostate datasets respectively.

4.4 Discussion

4.4.1 Strengths and Limitations

In this chapter, we proposed novel approaches to standardize the radiotherapy structure names using the heterogeneous prostate and lung radiotherapy structures. We demonstrated that our multi-view integration approach improves the standardization process. Structure delineation generates significantly imbalanced datasets, but our approach can overcome the data imbalance issues thereby demonstrating that

Datasete	Disease	Data Type	Precision	Recall	F_1-Score	Acc
Test (VCU)	Prostate	MLB	0.110	0.140	0.130	0.800
		Text	0.778	0.730	0.740	0.927
		Image	0.710	0.476	0.519	0.870
		combined	**0.778**	**0.792**	**0.782**	**0.941**
	Lung	MLB	0.140	0.170	0.150	0.810
		Text	0.830	**0.981**	0.873	0.969
		Image	0.610	0.565	0.585	0.918
		combined	**0.855**	0.895	**0.873**	**0.971**
Test (VA-ROQS)	Prostate	MLB	0.09	0.120	0.110	0.730
		Text	0.**890**	0.866	**0.872**	0.930
		Image	0.758	0.579	0.619	0.856
		combined	0.848	**0.897**	0.864	**0.932**
	Lung	MLB	0.130	0.170	0.150	0.780
		Text	0.921	**0.874**	**0.893**	**0.950**
		Image	0.825	0.694	0.708	0.916
		combined	**0.939**	0.741	0.778	0.943

Table 15: Intermediate Integration - Disease specific macro-averaged Precision, Recall, F_1-Score, and Overall Accuracy. MLB: Majority Label Baseline.

the proposed approaches can work on real-world datasets. However, our proposed approach has a few limitations, which are divided into clinical and methodological limitations.

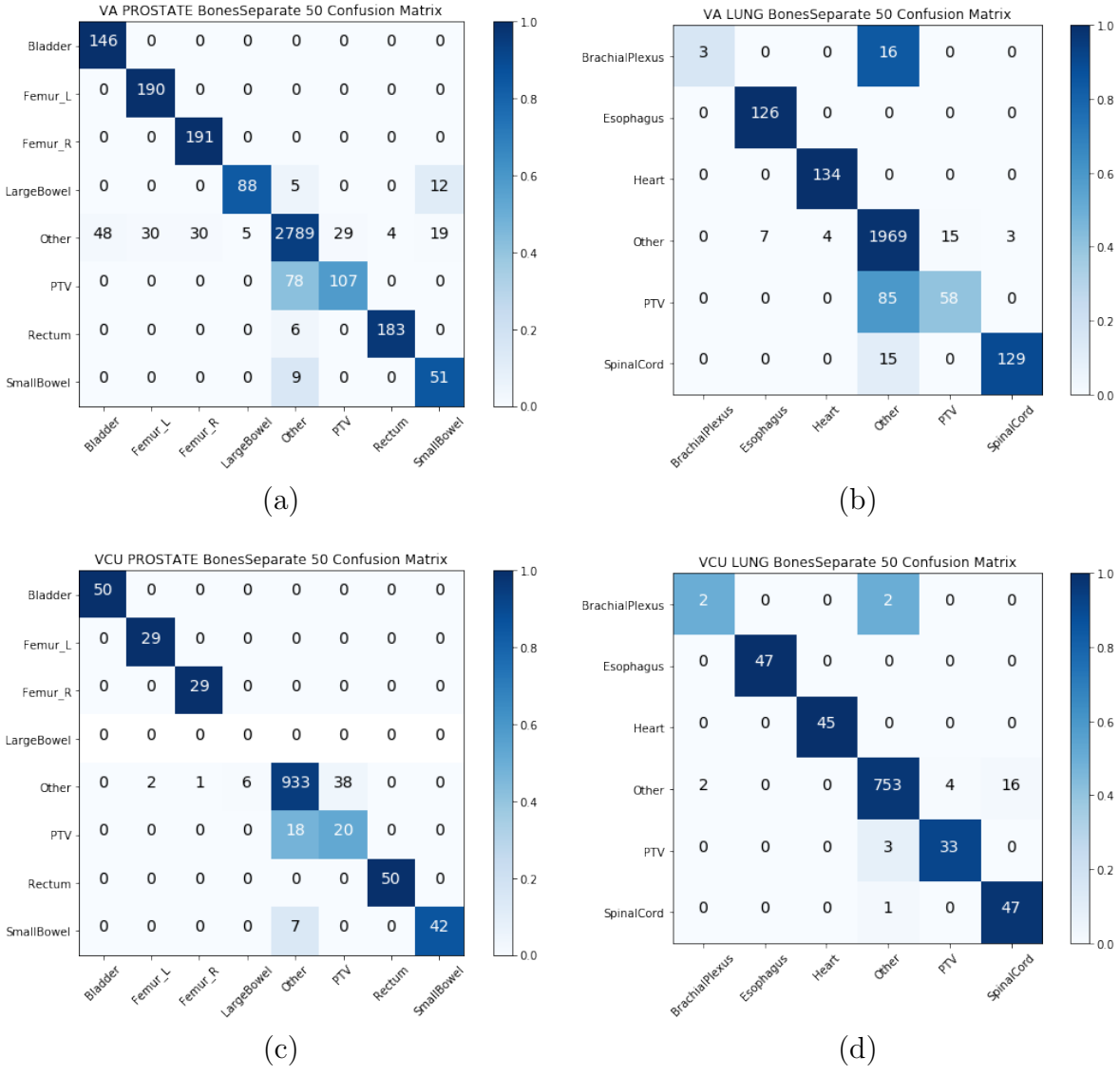


Fig 25: Intermediate Integration for VA-ROQS and VCU Lung Dataset Confusion Matrix. (a) Text Based features (b) Image features. (c) AVG of predictions. (d) MAX of two prediction. Darker color indicates better prediction. Diagonal indicates the correctly predicted labels.

Clinical Limitation

- So far, we were able to identify only OARs and Target (PTV) structures. Although these are critical structures, radiotherapy treatment involves other types

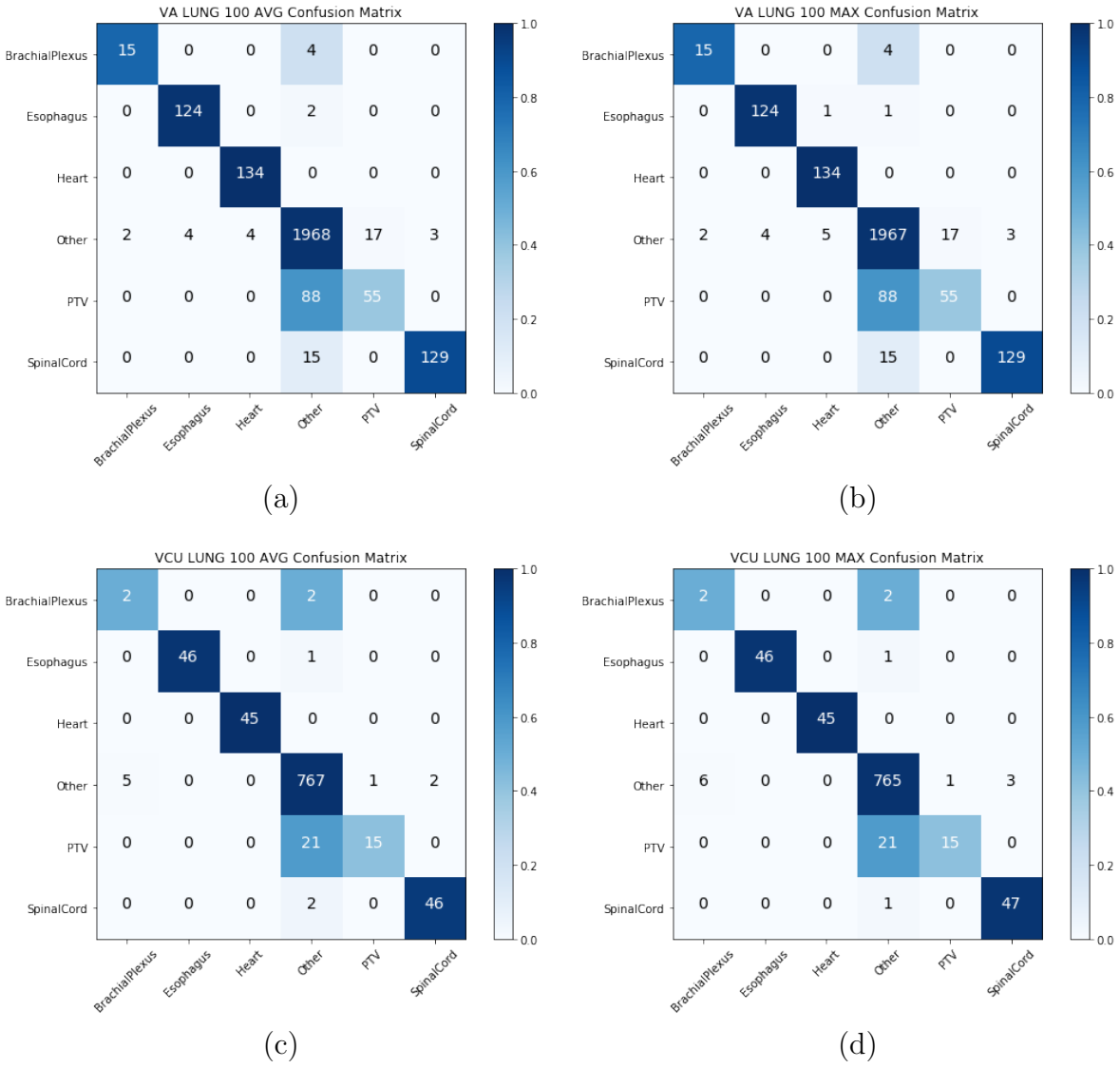


Fig 26: Late Integration Confusion Matrix for VA-ROQS and VCU Lung Datasets: (a) VA-ROQS Lung AVG Integration.(b) VA-ROQS Lung MAX Integration. (c) VCU Lung AVG Integration. (d) VCU Lung MAX Integration. Darker color indicates better prediction. Diagonal indicates the correctly predicted labels.

of structures, such as PRV and derived structures. To fully standardize the data, we need to standardize all structures, and not just the OARs and PTV.

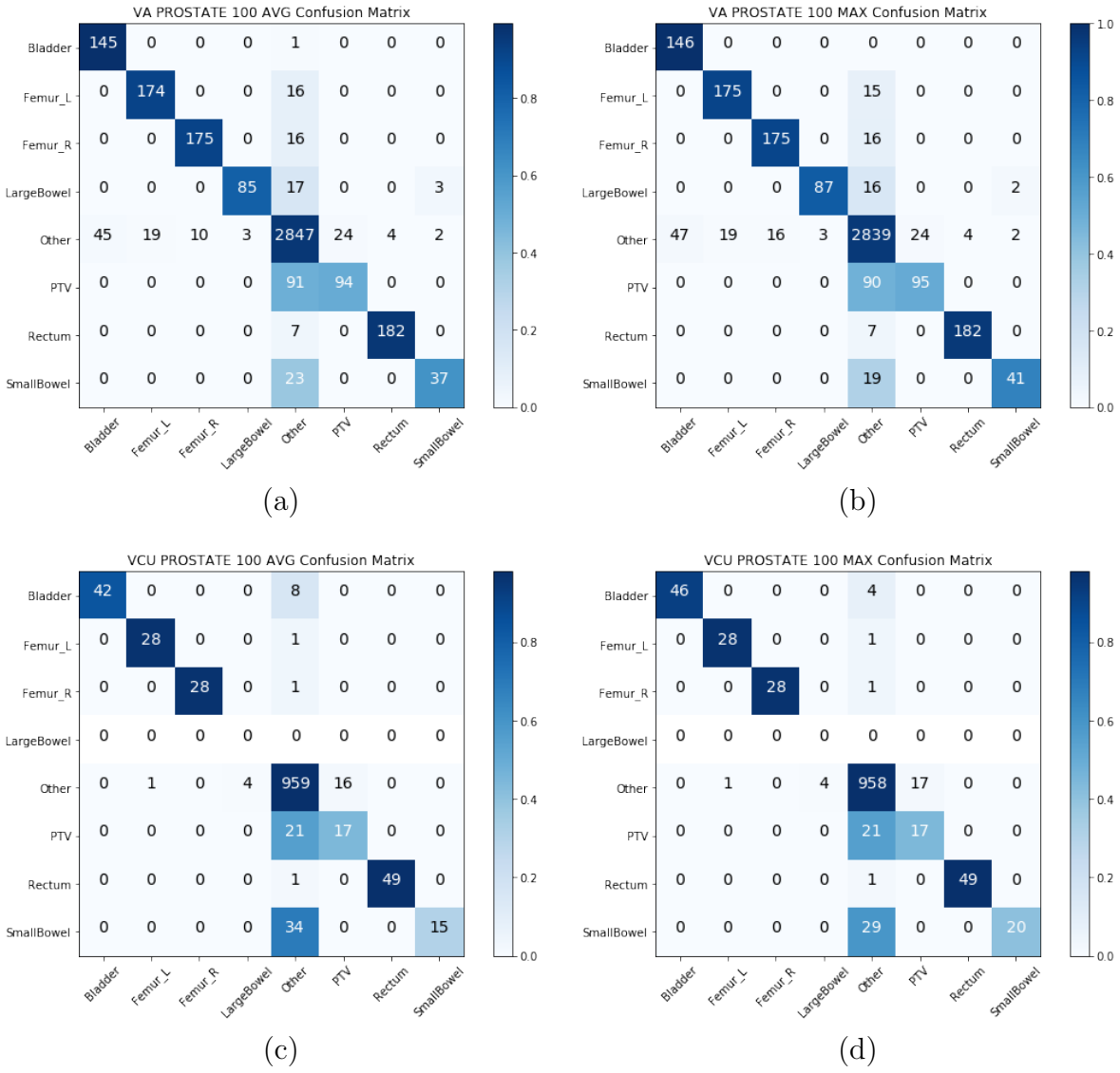


Fig 27: Late Integration Confusion Matrix for VA-ROQS and VCU Prostate Datasets: (a) VA-ROQS Prostate AVG Integration. (b) VA-ROQS Prostate MAX Integration. (c) VCU Prostate AVG Integration. (d) VCU Prostate MAX Integration. Darker color indicates better prediction. Diagonal indicates the correctly predicted labels.

- The OARs were selected based on the requirements of the VA-ROQS project whose primary focus was treatment quality assessment based on the specific

Dataset	Disease	DataType	Precision	Recall	F_1-Score	Acc
Test (VCU)	Prostate	MLB	0.110	0.140	0.130	0.800
		Text	0.778	**0.730**	**0.740**	0.927
		Image	0.710	0.476	0.519	0.870
		Avg	0.802	0.685	0.719	0.929
		Max	**0.801**	0.708	0.739	**0.936**
	Lung	MLB	0.140	0.170	0.150	0.810
		Text	0.830	**0.981**	**0.873**	**0.969**
		Image	0.610	0.565	0.585	0.918
		Avg	0.858	0.807	0.811	0.964
		Max	**0.849**	0.810	0.806	0.963
Test (VA-ROQS)	Prostate	MLB	0.090	0.120	0.110	0.730
		Text	0.890	**0.866**	**0.872**	0.930
		Image	0.758	0.579	0.619	0.856
		Avg	0.897	0.836	0.856	0.930
		Max	**0.897**	0.848	0.864	**0.930**
	Lung	MLB	0.130	0.170	0.150	0.780
		Text	**0.921**	**0.874**	**0.893**	**0.950**
		Image	0.825	0.694	0.708	0.916
		Avg	0.918	0.840	0.868	0.964
		Max	0.916	0.840	0.867	0.945

Table 16: Late Integration - Disease specific macro-averaged Precision, Recall, F_1-Score, and Overall Accuracy. MLB: Majority Label Baseline.

quality metrics; analysis of our datasets have shown that radiation oncologists have delineated many other OAR structures e.g., Kidney and Liver structure in prostate cancer. To truly build the generalized system that can identify all possible structures, the dataset needs to identify all correctly labeled OAR structures, and not just the significant OAR structures.

Methodological limitations

- Extraction of 3D volumes of structures requires selecting the bounding box size to make sure it covers the biggest possible structure in any given disease. Although, it is a one time step needed at the beginning of the dataset preparation, it does add an overhead in the complete automation of the pipeline.

- In recent years, deep learning algorithms such as Convolutional Neural Networks (CNN) have worked best for image based data classification. We plan to extend our pipeline to integrate CNN based image classification methods in the multi-view integration approach.

- It is difficult to capture the image semantics by turning images into a single vector and taking the top 50 components from it.

- We have extracted the structures fitted inside the bounding boxes. Using just structures information and discarding the other surrounding structures and anatomical information negatively affects the model performance.

- In late integration, we have tested only AVG and MAX for combining the data. This gives equal importance to both the Text and Geometric data. As we have seen from the single-view results, the geometric information model performed poorly when compared to the text based single view model. Hence we surmise

that a weighted average technique to integrate the results from the different views might produce better results.

4.5 Conclusion

In this chapter, we presented two types of multi-view integration methods: intermediate and late integration methods for structure name standardization. We utilized the physician-given names and geometric information of structures. We observed that the intermediate integration methods improves the overall performance of the models while late integraiton helps in reducing the false negatives. We tested our approach by training it on data from 30 VA RT centers and testing it on 10 VA RT centers and the VCU dataset. We demonstrated that our method works well on multiple disease sites and is also generalizable. We believe that the multi-view integration methods are best suited for structure name standardization, as they make the best use of different information to avoid the confusion. High VA-ROQS test set performance indicates that our approach was able to generalize very well within the VA system. Whereas excellent performance on VCU dataset suggests the model's ability to generalize well on the data from outside the VA systems

Contribution summary: In this chapter we address the limitations of structure name standardization using solely physician-given names and present an approach to combine the physician-given names with the geometric information of structures for structure name standardization. Specific contributions of this chapter are as follows.

1. We demonstrate that the use of bony anatomy information along with structures helps in the standardization process using geometric information.

2. We show that even target structure can be identified along with the Organs-at-Risk (OARs) with the physician-given names.

3. We demonstrate that it is still challenging to predict the standard name with just geometric information in real-world clinical datasets.

4. We finally demonstrate that integrating physician-given structure names with geometric information of structures improves the overall structure name standardization process.

CHAPTER 5

AUTOMATIC INCIDENT TRIAGE IN RADIATION ONCOLOGY - INCIDENT LEARNING SYSTEM

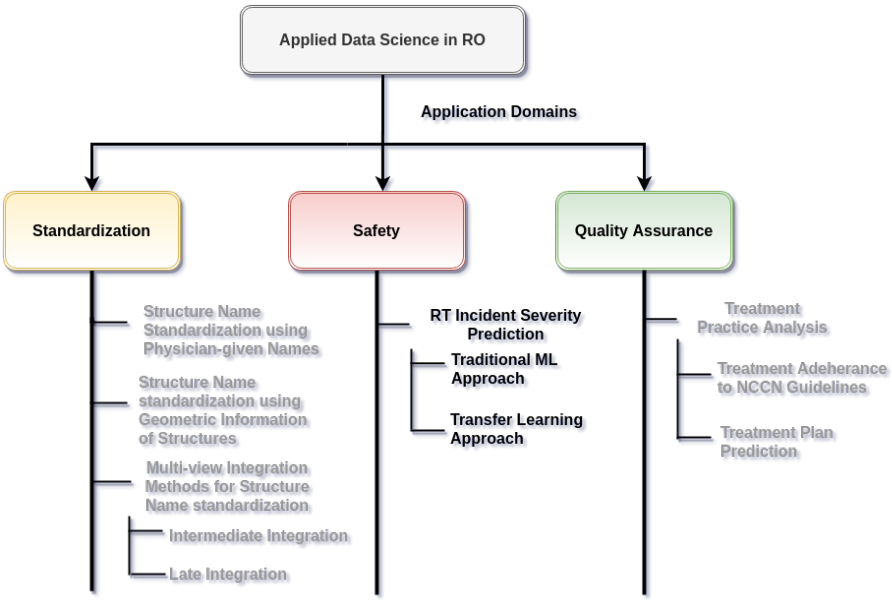


Fig 28: Thesis contribution, Chapter 5 contribution are highlighted.

5.1 Introduction

The radiation therapy (RT) cancer treatment speciality involves coordinated interactions between various clinical staff such as, dosimetrists, physicists, radiation therapists, nurses, and physicians. However, misadministration of RT can lead to potentially severe consequences [49, 50]. High-risk industries, such as the aviation and nuclear power industries [51], have demonstrated that the incident learning system can prevent such errors. The American Society for Radiation Oncology (ASTRO)

and American Association of Physicist in Medicine (AAPM) are professional societies that oversee the accuracy, safety, and quality of RT treatments. In March 2014, these societies started the Radiation Oncology Incident Learning System (RO-ILS) to enable documentation and analyses of incident reports in the radiation oncology domain.

In the wake of RO-ILS, the Veterans Health Administration (VHA) has deployed the Radiotherapy Incident Reporting and Analysis System (RIRAS). The system is being used by the 40 VHA radiation therapy centers as well as the Virginia Commonwealth University (VCU) Health center. RIRAS is a web-based Incident Learning System (ILS) developed by TSG Innovations Inc. and is accessible via the intranet, where any member within the department can submit incident/good catch reports. The taxonomy, data dictionary, and radiotherapy process of care incorporated in the design of RIRAS are based on the AAPM report on "Error Reporting" [52]. Furthermore, RIRAS is fully compliant with the Patient Safety and Quality Improvement Final Rule [53]. RIRAS is built to report all types of workflow events, that includes even minor errors in documentation and processes; such errors may decrease the efficiency of treatments and cause delays besides having other downstream effects.

Figure 29 shows the typical schematic representation of the RIRAS system. All events reported are reviewed by the ILS committee on a call or face to face interaction; typically such ILS team comprises of medical physicists, dosimetrists, therapists, nurses and physicians. The ILS team completes the analysis form section where event summary titles, error type, causes based on a standard dictionary and safety barriers or quality control measures affecting the event are entered. The event is reported to the chief of the appropriate clinical group if the severity is determined to be high or the ILS team determines that further review is necessary. Otherwise, the ILS committee reviews and codes the events by consensus at their weekly review meeting. Severe

incidents require immediate action and root cause analysis (RCA). Understanding the cause of severe incidents helps in preparing an appropriate plan of action. Even the less severe incidents are further analyzed and tracked to avoid similar events. An appropriate action plan and feedback is sent to the incident reporter and professional group so that policy and process can be improved.

Natural language processing (NLP) is a popular technique for analyzing large quantities of clinical texts, notably in medical specialties such as radiation oncology and radiology [54, 55]). According to Meystre and Pons [55], the five major categories of application of NLP in radiation oncology are (1) diagnostic surveillance, (2) cohort building for epidemiological studies, (3) query-based case retrieval, (4) quality assessment of radiologic practice, and (5) clinical support services. In this chapter, we introduce a sixth category for the application of NLP in radiation oncology: analysis of radiotherapy incident reports. Specifically, we present the use of NLP to automate the prediction of severity from the incident description. As shown in Figure 29, the bottleneck step in the RIRAS system is triaging. We propose a machine learning method to automate the triage process which can thereby reduce the manual efforts needed by the subject matter expert (SME) to determine the severity; providing an initial prediction of low and high severity with confidence also helps to augment the incident analysis process.

The rest of the chapter is structured as follows. In Section 5.3, we present the methods used and details of the dataset. Section 5.4 describes the results and in Section 5.5, we present the discussion and conclusion. In the final section, we present the limitations in our approach that can motivate future work.

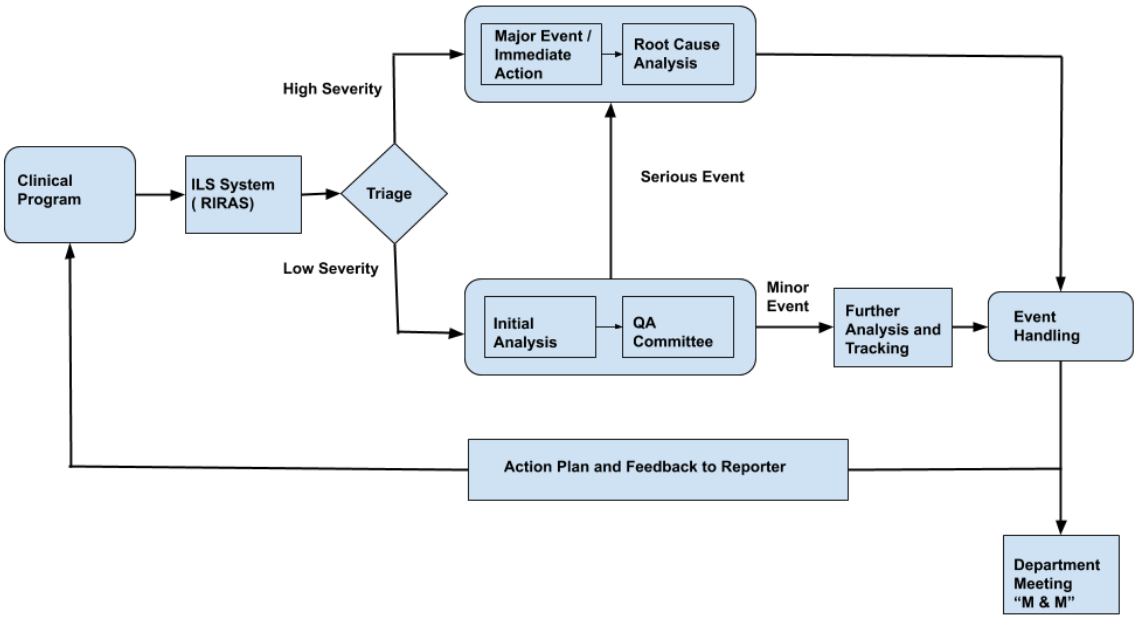


Fig 29: Schematic Representation of Radiation Oncology - Incident Learning System (RIRAS).

5.2 Background

Healthcare incident reports, including the radiotherapy incidents submitted into the RIRAS software, are similar to the safety reports of various industrial environments in that their narratives are reported in an unstructured free-text format. Free text, while convenient for the reporter, presents a challenge for data aggregation and requires suitably-qualified personnel to read and analyze. However, due to the lack of dedicated incident-analysis personnel, minor incident reports in healthcare often accumulate, as resources are used to deal with front-line issues that are typically considered more urgent.

To the best of our knowledge, there is no work reported in the field of radiotherapy to identify the severity of the incidents reported using incident description. However there have been well reported research in other industries such as aviation, and nuclear

[56, 57, 58, 59, 60] to classify the incidents reported in the respective fields. In healthcare there has been successful work done in classifying the verbal autopsies [61]. A team in Canada has done a study on identifying the incident types from Canadian medication incident report [62]. Another team in Australia performed more extensive study predicting the two types of patient safety incidents: incorrect patient identification and inadequate clinical handover [63]. Hence, there is an urgent need for creating an actionable learning-based incident reporting system in healthcare [64].

5.3 Methods and Materials

5.3.1 Dataset

RIRAS is a web-based ILS deployed on the VHA radiation oncology centers intranet and VCU intranet in early 2014. It was designed to collect good catch data and adverse events, besides analyzing their causes and contributing factors, and finally, to prevent possible occurrences in the future. This system provided a platform to report the adverse events across 40 VHA radiotherapy treatment centers. We collected data from both sources, which consisted of incidents that were triaged into four levels of severity, namely, A through D, where A is most severe, and D is least. From here on, the dataset collected from VHA centers and VCU radiotherapy center will be referred to as VHA data and VCU data, respectively. Table 17 shows the sample examples of incident descriptions reported and their respective severities assigned.

VHA Dataset: The VHA clinical reporters entered incidents into RIRAS since 2014. For the time period between 2005 and 2014, the incident reports were collected for only high severity (level: A) incidents. These reports were collected by mostly

emailing the VHA's National Health Physics Program office who logged the reports in excel spreadsheets. These reports (46 reports) were entered into RIRAS in 2015. For the purposes of this analysis we used the data collected till 2017. A total of 530 incidents were reported across the VHA centers at the time when this data was collected, in which 345 incidents were analyzed by the subject matter experts and the incident analysis reports were assigned severities. The incidents distributed based on the severity in VHA dataset is as shown in Figure 30a, where the incidents are distributed as A (62), B (52), C (162), and D (67). A total of 185 incidents were not analyzed and hence were missing the severities; such non-analyzed incidents cannot be used in our classification task.

VCU Dataset: The incidents collected at VCU were between 2014 to 2019. A total of of 540 incidents were reported, among which 7 were not analyzed by the subject matter experts. The incidents were distributed based on their severity as shown in Figure 30d, where the incidents were distributed as A (9), B (40), C (165), and D (318). A total of 7 incidents were missing severities.

5.3.2 Incident Severity Types

The AAPM (professional society of Medical Physicist in the US) formed a working group on Prevention of Errors in Radiation Oncology where a panel of experts developed consensus recommendations considering five key areas: data elements, definitions, severity scales, process maps, and causality taxonomy [52]. RIRAS was implemented following these recommendations. Following are the important terminologies related to ILS:

- **Incident**: refers to events that are unintended or unexpected in the realm of

Incident Description	Severity
The patient on the EMR screen was not the patient called for treatment. During set up the radiation therapist noticed that the patient on the table is not the patient selected on EMR. Introduced new policy of double checking the patient ID by therapists.	High (A or B)
Spinal cord and Brainstem max doses were incorrectly recorded in dose summary spreadsheet and in paper chart and Aria printouts. Aria dose recording paper chart and Aria PDF were corrected.	Low (C or D)

Table 17: Examples of Incident description and respective Severity assigned by Subject Matter Experts.

standard clinical operations. Such events may cause adverse effects on equipment, healthcare providers or patients.

- **Near Miss or Good Catch:** refers to unplanned events that could potentially cause a damage, illness or injury, but did not actually do so. However, such near misses were only averted due to good fortune. Such events are mostly labeled by "human error", while faulty systems or processes may exaggerate the harm, and needs to be studied better. Other terms used for such are "close call", and for moving objects, "near collision".

- **Unsafe Condition:** refers to hazardous work environments, circumstances, or physical conditions that increase the probability of an incident.

In the VHA, the National Radiation Oncology Program (NROP) consists of 40

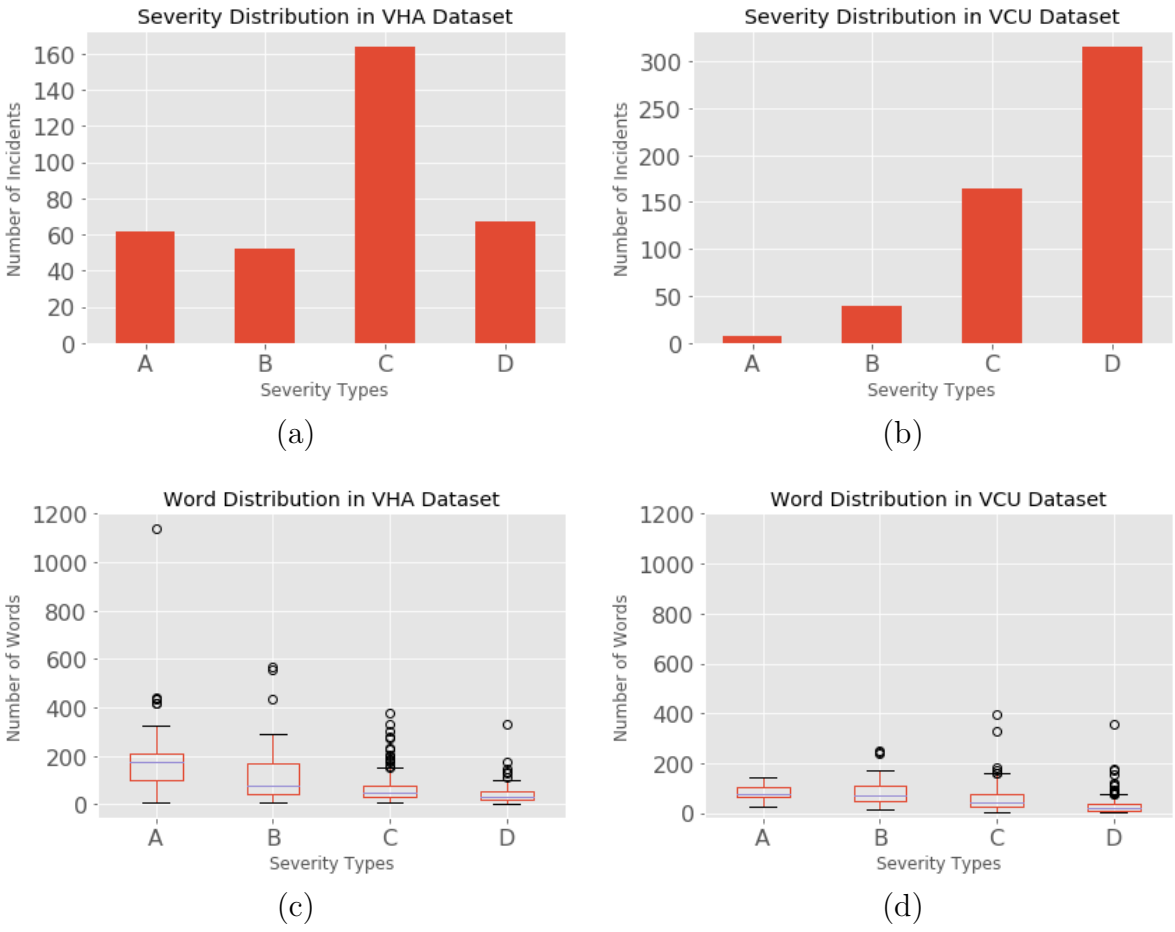


Fig 30: Dataset Distributions: (a) Severity Distribution in VHA dataset. (b) Severity Distribution in VCU dataset. (c) Word Distribution in VHA dataset. (d) Word Distribution in VCU dataset.

facilities treating over 12,000 patients annually within the system, and an additional 14,000 outside of the system. As the rate of errors has been estimated to occur as frequently as 1 per 600 patients [65], the utilization of ILS can provide a means of gathering and analyzing incident data so that patient safety and workflow process improvements can be implemented and the effects of such changes tracked over time. For multi-institutional programs such as the NROP, aggregating incident reports

from all facilities into a single database increases the effectiveness of incident learning and allows for the assessment of systematic errors and trends as well as national standardization of policies and procedures. Based on the recommendation of AAPM, NROP defined the reasoning behind the severity categorization and explained what constitutes of low to high severity. Reports were subsequently categorized based on four levels of severity: A through D. Explanations for these incident severity categories are shown below:

- **Level A**: It is a significant event or near miss with a potential for a medical event or serious patient injury, as well as a repeat of a Level B event. The problem has an urgent need for correction and may impact multiple patients or Radiation Oncology processes. Level A incidents require a full Root Cause Analysis. The Lead Responder for a level A incident will typically be a medical physicist. Very few ($< 2\%$) incidents should fall into this category.

  Example: A patient is treated at the wrong site. The Lead Responder would be a medical physicist appointed by the Director of Clinical Physics.

- **Level B:** It is a significant event or near miss that did or could result in a dose deviation $> 5\%$, a significantly larger than intended dose outside the treatment field, a treatment delay of greater than one day, or a similar scenario that is neither a Medical Event nor poses a risk of serious patient/staff injury. The problem should be confined to a single process step and could likely be promptly addressed with an Apparent Cause Analysis. The Lead Responder for a level B incident will either be a medical physicist or a department lead. Few ($< 5\%$) incidents will fall into this category.

  Example: A case is planned and treated for five fractions (out of 20) with an improperly expanded contour that is 5 mm larger than intended by the

physician. The Lead Responder would be the Director of Dosimetry.

- **Level C**: A minor incident, near miss, or condition that warrants an appropriate response from a department lead, who is typically the Lead Responder. The level of the response will be up to the department lead, but the response must be reported back to the Quality Assurance (QA) committee. Many incidents will fall into this category.

  Example: A case is planned and prepared for treatment assuming 5 mm bolus. The physician opts not to use the bolus, and only the monitor units are not recalculated before treatment approval. The Lead Responder could be the Director of Clinical Physics.

- **Level D**: A very minor incident, near miss, or condition that warrants awareness by the department lead. The level of the response will be up to the department lead, and there is no mandate for them to report back to the QA committee. The incident will be logged within RIRAS for trend tracking purposes.

  Example: A field is mislabeled in a plan. The Director of Dosimetry is informed.

5.3.3 Model Selection

In this section, we describe the model selection techniques using traditional machine learning and deep learning approaches with model fine tuning and transfer learning.

5.3.4 Traditional Machine Learning

We first pre-processed the textual data from the incident reports. Next, we identified the features from the text to build the corresponding feature vectors necessary for any supervised machine learning model. The next step was to select the appro-

priate machine learning algorithm for which we tested different types of algorithms to predict the severity of the incidents.

Since machine learning algorithms require numerical data, we next converted the textual data into numerical features. This involves the following major steps [30]: 1) tokenization, 2) feature set generation, and 3) vectorizing the features with different feature weight calculation techniques. To this end, we applied the following steps in developing the proposed traditional machine learning pipeline (as shown in Figure 31).

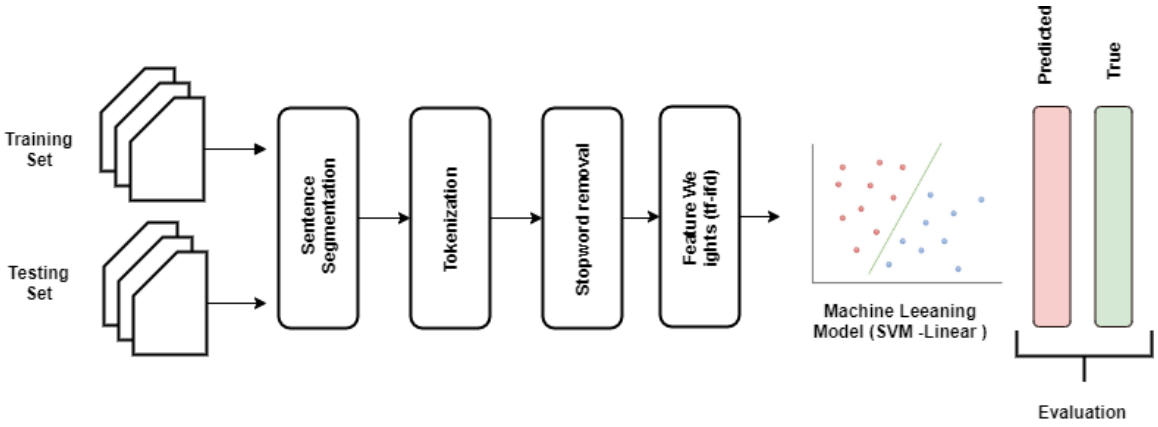


Fig 31: Triage Process: Pictorial representation of the traditional machine learning severity classification pipeline.

5.3.4.1 Data Splits:

As before we built a model by splitting the data into three sets: the training set, validation set and test set. Using the separate data for evaluation not seen during training lets us test if the trained model is not over trained. Once the final model is prepared, the test dataset is used to test the model with unseen data (not seen during training and not used as validation).

5.3.4.2 Data Preprocessing

All incident descriptions were first processed using NLTK (python library for text processing) [66]. The following procedures were applied:

- ***Data Cleaning:*** Removing the unnecessary parts of text. In our dataset, we removed the characters "&quot", "&&", which were added to the text when collecting the data from XML files.

- ***Tokenization:*** It is the process of splitting the long string of text (sentences) into tokens (words). These tokens are used as features. We used NGram tokenization to produce uni-gram, bi-grams, and tri-grams [67]. Uni-grams are also known as bag-of-words representing individual terms that occur in a document (e.g., "surgery", "prostate", "dosimetry"). bi-grams and tri-grams represent the consecutively occurring two or three terms in a document (e.g., patient scheduled, patient rescanned, patient planned radiation therapy), which help capture the semantics of text; one such example is negation (e.g., no pain).

- ***Text Normalization:*** It is the process of converting terms occurring in text into one form. We used lower case normalization to ensure that all the words occurring in different forms are represented as one (e.g. Patient, PATIENT, patient, and pAtient are converted to "patient") [68].

- ***Stopword Removal:*** It is the process of identifying and removing more frequently occurring words from the text. We considered removing commonly occurring English language words (e.g. a, the, it, what, why, she, etc.), which hold no classification value [67]. We used general English language stop words provided in the NLTK Package. This technique is commonly used in information retrieval and NLP document classification implementations [68].

- ***Term frequency filtering***: It is the process of identifying the infrequently appearing words in the corpus [69], which helps with reducing the feature vector size. We have used a minimum term frequency of 5 as cutoff.

- ***Feature Weighting Techniques***: We used three types of feature weighting methods as shown below. Term presence (tp), Term Frequency (tf), and Term Frequency-Inverse Document Frequency (tf-idf). We have explained each of these weighting techniques in Section 3.3.4

- ***Vectorization*** – It involves using the above steps to extract features and weights to generate uniform vector representations of each report. Each feature weighting technique (shown above) was used to create three types of feature vectors. One such feature weighting technique is *tf-idf*; *tf-idf* assigns the weight to the term based on its frequency in a document, and its appearance in all documents in the corpus. The assigned weight indicates the relevancy of that term to the document when classifying the documents into different classes [68, 70, 71]. The higher value of the term indicates its higher importance. The term frequencies are normalized so that longer documents do not skew the results [72].

5.3.4.3 Classification Algorithms

We next tested the classification algorithms explained in Section 2.3 to select the best algorithm for the traditional machine learning pipeline.

5.3.4.4 Evaluation Metrics

To evaluate our model we considered macro-averaged precision, recall, and F_1-Score that can better capture how well a classifier can identify cases that it does not

see often as explained before. Results are also presented using a confusion matrix which shows the number of correct and incorrect predictions as summarized with prediction counts between each class. It provides insight not only into the errors being made by the classifier but more importantly, the types of errors that are being made.

5.3.4.5 Initial Model Selection

The extracted incident reports were used to train machine learning classifiers with Python's scikit-learn (version 0.21.3) [39]. The labeled incident report corpus was stratified as 80:20 as training and test split. A total of 276 (80%) incident reports were used for model training and 69 (20%) for model testing to characterize the model performance.

In our initial work, to test the viability of predicting all four severities, we built four different models by combining severities as below [7]:

- Model-1: We considered incidents with severities A and C.
- Model-2: We combined $A\&B$ as high and $C\&D$ as low severities.
- Model-3: We considered only B and D severity.
- Model-4: All 4 severities, A, B, C, and D are considered as separate.

These models provide insight into our methods' ability to find patterns when incidents with different severities are considered. We built above mentioned four models with SVM-linear classification algorithms, and NGram features with tf-idf feature weights. Table 18 shows the results of these four models. We observed that Model-1 and Model-3 achieved an F_1-Score of 0.87 and 0.78 respectively; we can infer that incidents A & C (Model-1) and B & D (Model-3) have better patterns to classify

incidents. The poor performance of Model-4 indicates that there is a lot of similarities between the A & B and C & D severities in the confusion matrix. Model-2 achieves the F_1-Score of 0.81. It is clear from the results that predicting all the four categories is difficult based on our current datasets. However, categorizing incidents into high (A&B) and low (C&D) severity (Model-2) is viable.

Models	**Severities**	**Precision**	**Recall**	**F_1-Score**
Model-1	A and C	0.86	0.87	0.87
Model-2	A&B and C&D	0.83	0.80	0.81
Model-3	B and D	0.80	0.77	0.78
Model-4	A, B, C, and D	0.53	0.56	0.53

Table 18: Results from the severity categorization model for different combinations of severities. Results reported are macro-averaged precision, recall and F_1-Score for SVM with linear kernel model.

Hence, we used Model-2 for building the automated triage system. To select the best classification algorithm to build the final model, we applied the above explained steps to build the severity prediction model. Figure 31 shows the pictorial representation of the classification pipeline. Five different classification algorithms were used: k-Nearest Neighbors (KNN) [18], SVM-Linear [15], SVM-RBF [16], Random Forests [20], and Logistic Regression [38] with feature extraction and weighting methods. Standard macro-averaged precision, recall, and F_1-Score are used as evaluation metrics for discrimination on the training and test sets. Table 19 and 20 shows the initial model selection results for VHA and VCU datasets. We observed that SVM with linear kernel consistently performed well with all feature vector generation methods.

In all combinations of algorithms and features, SVM with linear kernel algorithm and tf-idf features performed the best with an F_1-Score of 0.808. With this observation, we chose the tf-idf and SVM-linear to build our final model.

Dataset	Features Weights	Algorithm	Precision	Recall	F_1-Score
VHA	tp	SVM_RBF	0.809	0.519	0.418
		SVM_Linear	**0.792**	**0.698**	**0.705**
		Random_Forest	0.856	0.685	0.686
		Logistic_Regression	0.792	0.698	0.705
		KNeighbors	0.304	0.500	0.378
	tf	SVM_RBF	0.797	0.655	0.649
		SVM_Linear	**0.815**	**0.735**	**0.747**
		Random_Forest	0.837	0.729	0.740
		Logistic_Regression	0.815	0.735	0.747
		KNeighbors	0.813	0.537	0.454
	tf-idf	SVM_RBF	0.720	0.562	0.512
		SVM_Linear	**0.835**	**0.798**	**0.808**
		Random_Forest	0.818	0.692	0.696
		Logistic_Regression	0.759	0.599	0.571
		KNeighbors	0.680	0.664	0.668

Table 19: Model-2 selection results for severity categorization for VHA dataset. Results reported are macro-averaged.

Dataset	Features Weights	Algorithm	Precision	Recall	F_1-Score
VCU	tp	SVM_RBF	0.458	0.500	0.478
		SVM_Linear	0.458	0.500	0.478
		Random_Forest	0.458	0.500	0.478
		Logistic_Regression	0.458	0.500	0.478
		KNeighbors	0.458	0.500	0.478
	tf	SVM_RBF	0.458	0.500	0.478
		SVM_Linear	0.460	0.500	0.475
		Random_Forest	0.458	0.478	0.478
		Logistic_Regression	0.460	0.490	0.473
		KNeighbors	0.458	0.500	0.478
	tf-idf	SVM_RBF	0.458	0.500	0.478
		SVM_Linear	0.460	0.495	0.475
		Random_Forest	0.458	0.500	0.478
		Logistic_Regression	0.458	0.500	0.478
		KNeighbors	0.457	0.490	0.473

Table 20: Model-2 selection results for severity categorization for VCU dataset. Results reported are macro-averaged.

5.3.5 Traditional Machine Learning Vs. Transfer Learning:

Traditional machine learning refers to training a model on a particular task (say, text classification) from one domain and expecting it to perform well on unseen data from the same domain. Whereas, transfer learning refers to the use of a model

that has been trained to solve one task (e.g., language modeling: predict next word in a sentence) as the basis to solve some other or somewhat similar problem (text classification) [73]. It also refers to the training of a model with a large-scale dataset and next using this pre-trained model for the same task with different dataset and labels. The computer vision domain popularized transfer learning with the ImageNet dataset [74].

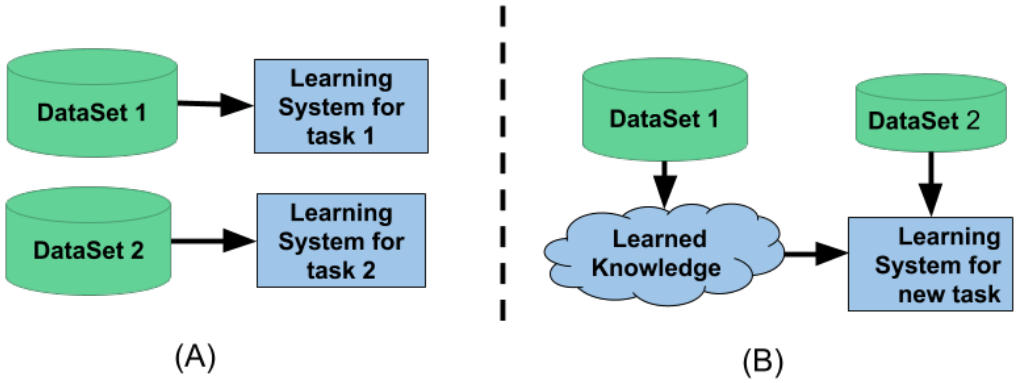


Fig 32: (A) Traditional machine learning system (B) Transfer Learning system.

Figure 32 (A) shows the traditional machine learning setup. This method is isolated and performs single-task learning. It is not possible to use the knowledge from one task to learn the new task. Traditional machine learning also needs a lot of data to learn the given task. Whereas, Figure 32 (B) shows the transfer learning setup. This setup utilizes the knowledge learned from one task to learn a new task; because of the knowledge transfer, it requires less data and computation time to learn a new task.

5.3.6 Transfer Learning

Transfer learning is the process of training a model on a large-scale dataset and then using the pre-trained model to conduct learning for another downstream

task. One such simple transfer learning technique is to use the word2vec embeddings, which uses a single layer of weights from the trained model. However, full neural networks in practice contain many layers, and using transfer learning for a single layer is only scratching the surface of what is possible. From the immediate past, one such technique that fine-tunes the full network for transfer learning on textual data is the universal language model fine-tuning (ULMFiT) [75].

Universal Language Modeling and Fine Tuning

The ULMFiT is one of the revolutionary algorithms in the field of NLP for knowledge transfer used for text classification. It uses all the layers of a neural network for transfer learning. Figure 33 shows the high-level pictorial representation of ULMFiT.

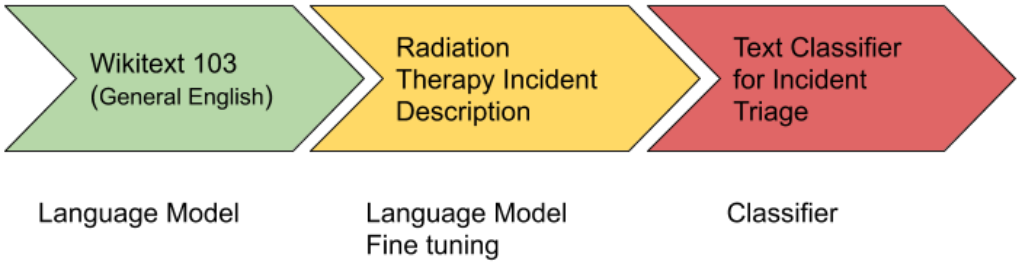


Fig 33: Pictorial representation of high level Universal Language Model Fine-tuning (ULMFiT) approach used for incident triage.

The ULMFiT has three main steps.

1. General Domain Language Modeling: In the first step, an unsupervised language model is trained on a large corpus to generate a general-domain language model. For this, a pre-trained general-domain English language model was used [75], which is trained with language model ASGD Weight-Dropped Long

Short-Term Memory (AWD-LSTM) on Wikitext-103 [76].

2. Target Task Language Model Fine Tuning: In the second step, the general domain language model is fine-tuned with the domain/target specific dataset. A pre-trained general-domain language model allows the target task language model to converge faster and results in a robust language model even for small target datasets. A pre-training provides a robust representation for uncommon words in the target training dataset.

3. Target Classifier Fine Tuning: In the third and final step, it adds two additional linear blocks to the pre-trained language model. The first linear layer takes the pooled last layer of the language model as input on which it applies ReLU activation. The last layer is a fully connected layer having softmax activation that provides the target classes' prediction probability.

5.4 Results

In this research, our goal was to augment the triage process in RIRAS by predicting the severity of the incident using the textual description of the incidents reported. We used two different approaches to predict the severity of the reported incidents: a traditional ML and transfer learning approach with the more advanced algorithm called ULMFiT. Below we describe the results from each of these approaches.

Traditional ML Results

From the initial model selection results, we observed that SVM-linear performed best in comparison with others. Hence, we used the SVM-linear to build the final model. We built separate models for VHA and VCU datasets. Table 21 shows the traditional ML results. We compared the results with the majority label baseline

(MLB Baseline) model. In the MLB baseline, all the predictions are done as a label that occurs the majority of the time. The metrics are calculated based on the majority label. In a balanced binary classification model, the random probability of predicting a correct class is 50%, but both the datasets used in this work are imbalanced. Hence, we compared the results with the Random and MLB baseline. The VHA dataset model achieved 0.80, 0.77, and 0.78 of precision, recall, and F_1-Score, respectively. When compared to the MLB baseline, it achieved much better results. Whereas for VCU, we noticed that SVM-Linear results are the same as the MLB baseline, indicating that the model was not able to learn the classification patterns from the training data. Figure 34 shows the confusion matrix of traditional ML results for both VHA and VCU. We noticed that for the VCU dataset, the ML model assigned the Low severity (majority label in the training set) to all test set instances.

Models	**DataSource**	**Precision**	**Recall**	**F_1-Score**
Random	–	0.50	0.50	0.50
MLB Baseline	VHA	0.33	0.50	0.40
	VCU	0.458	0.500	0.478
SVM-linear	VHA	0.80	0.77	0.78
	VCU	0.458	0.500	0.478

Table 21: Traditional Machine Learning Results for Model-2. Reported results are macro averaged precision, recall, and F_1-Score for SVM with linear model. MLB: majority label baseline.

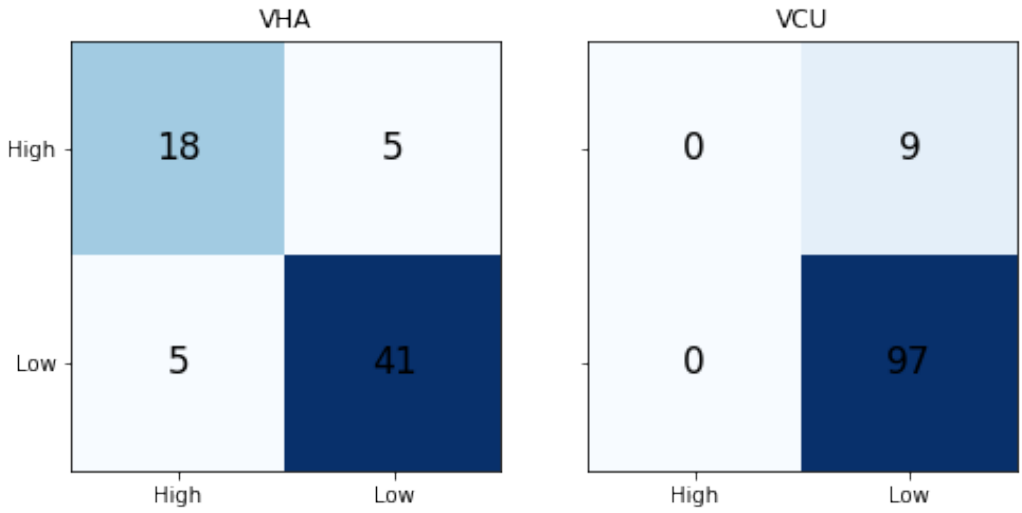


Fig 34: Traditional ML Results Confusion Matrix. Left confusion matrix is for VHA test set and right is for VCU test set. Diagonal indicates the correctly predicted class count.

Transfer Learning Results

Table 22 shows the results for different models built with ULMFiT. As explained in Section 5.3.6, transfer learning is a way to utilize the knowledge learned from one task into another task. In this research, we used ULMFiT to build the transfer learning based approach to predict the severity of incident reports in radiation oncology. ULMFiT involves building the language model (LM) and use it in the classification model.

In order to test the effects of data source on the models' ability to predict the severity of the incident reported using the descriptions, we built three different LM models based on the data source: VHA, VCU, VHA_VCU; the VHA_VCU dataset combines both the VHA and VCU datasets. Next, we trained the separate classification models with VHA and VCU datasets by taking knowledge from the LM models. This provided us with (3 X LM model) X (2 X Classifiers) = 6 pipelines to test for

LM	Train	Test	Precision	Recall	F_1-Score	Support
VHA	VHA	VHA	0.77	0.78	0.78	69
VHA	VCU	VHA	0.68	0.61	0.61	69
VCU	VHA	VHA	**0.80**	**0.83**	**0.81**	69
VCU	VCU	VHA	0.33	0.49	0.39	69
VHA_VCU	VHA	VHA	0.76	0.79	0.75	69
VHA_VCU	VCU	VHA	0.54	0.51	0.46	69
VHA	VHA	VCU	0.56	0.68	0.48	106
VHA	VCU	VCU	**0.67**	**0.69**	**0.68**	106
VCU	VHA	VCU	0.55	0.64	0.53	106
VCU	VCU	VCU	0.46	0.49	0.47	106
VHA_VCU	VHA	VCU	0.55	0.61	0.54	106
VHA_VCU	VCU	VCU	0.59	0.54	0.55	106

Table 22: Transfer Learning Results for Model-2. First six rows for VHA test set models and last six rows are for VCU test set. Results reported are macro-averaged. Support indicates the total number of samples in test sets. LM: Language Model.

each data source, and a total of 12 models for VHA and VCU. Table 22 shows the transfer learning results. The results reported are macro-averaged precision, recall, and F_1-Score.

We observed that transfer learning results are comparably better than traditional ML learning results. For the VHA test set, we noticed that the pipeline with VCU

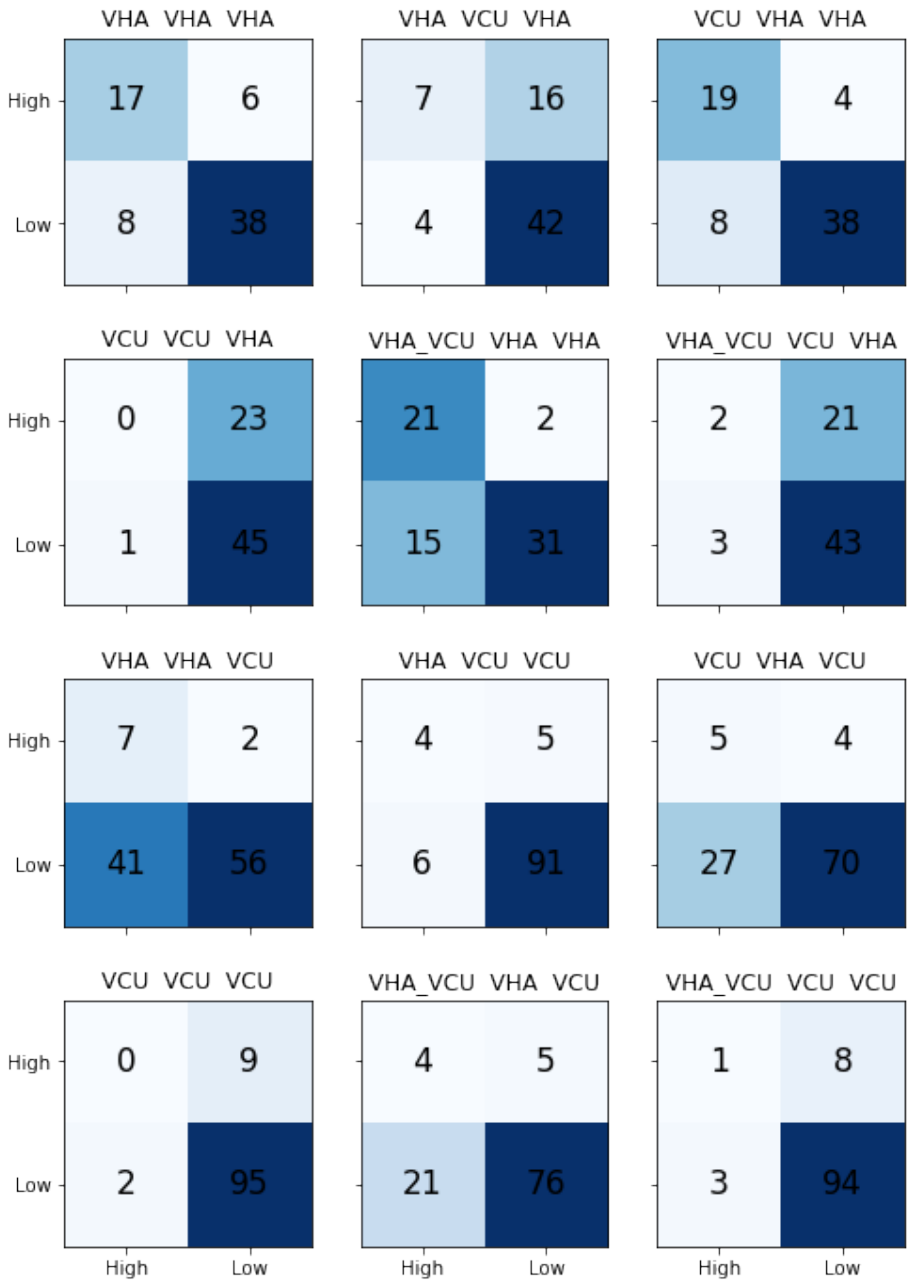


Fig 35: Transfer Learning Results: Confusion Matrix for each model in test dataset. Title in each confusion matrix indicates the respective model. Top two rows (six models) is for VHA test set and bottom two rows (six models) for VCU test set. Diagonal indicates the correctly predicted class count.

LM model and classification model trained with VHA achieved the best results. LM models trained separately with VHA, VCU, and VHA_VCU performed similarly for the VHA test set. It is clear from the results that the classification model needs to be trained with VHA data to predict the VHA test set. Transfer learning models performed well for the VCU dataset with precision 0.67, recall 0.69, and F_1-Score of 0.68 compared to the traditional ML model. Figure 34 shows the confusion matrices for all the models. The model with LM trained on VHA data and classifier trained on VCU data performed better on the VCU test set.

5.5 Discussion

In this chapter, we presented an approach to predict the severity of the radiation oncology incidents. The purpose of this work is not to replace the manual triage process, but rather, augment it by predicting the severity of the incident with reported description and provide the recommendation to the subject matter experts on the likelihood of an incident being of low or high severity. To do that, we used NLP techniques and ML algorithms to build the automated triage pipeline. We used traditional ML and transfer learning approaches.

The datasets used in this work come from two different sources; they are similar, yet have different characteristics. We noticed that the distribution of incidents based on the severity type is different in VHA and VCU datasets; there are fewer High severity incidents in the VCU dataset compared to the VHA dataset even though the total number of incidents in VCU are higher than VHA. We noticed that the descriptions of the incidents reported in the VHA dataset are longer on average compared to the incident descriptions reported in the VCU dataset. The length of the incidents also correlates with the severity of the incidents. The High severity (A & B) incidents, on average, have long descriptions compared to the Low severity (C & D)

incidents. It does not mean that the length of the description of the incident indicates the severity of the incidents. However, we believe it may be because the incident reporters tend to describe incidents in detail if they deem the incident is severe. The difference in length of descriptions may be due to the institution type and practice at those institutes. VHA incidents are coming from 40 VHA treatment centers, whereas VCU is a single institute. NLP makes use of the words in the description to find the patterns of the specific severity. Hence, a well-explained description is always better than a short one. Talking to SMEs, we have learned that some times just incident description provided is not enough to infer the severity; they always reach out to incident reporters for more information before analyzing the incident and assign severity. Thus, we believe that there is a need and opportunity to build guidelines on reporting practices. All the staff who use the RIRAS system to report incidents needs to be aware of guidelines and follow the instructions while reporting an incident.

Comparison with previous work

While ML and NLP based methods have been widely used to analyze incident reports from other domains, such as aviation [77], they have only been scarcely used in the healthcare domain before [63]. Straightforward comparison of our work with others is not possible because of the following two reasons. First, there has been no prior work related to the radiation oncology incident severity prediction using ML and NLP. Second, related work in healthcare incident analysis is more focused on other types of incident reports, where such incidents were recorded as free text. For example, Wong and Akiyama [62] analyzed 227 medication incident reports using a logistic regression based classifier to categorize the incident types based on adverse drug effects. Similarly, Wang et al. [78] used an integrated ML and NLP based pipeline to categorize incident reports related to patient safety; however, their method performed

poorly in properly classifying the severity levels. Finally, another related work in the healthcare domain considered verbal autopsies for text-based classification [61] with good accuracy; such autopsies bear some resemblance to incidence reports. However, none of these works are directly comparable to our proposed method which considers incident reports from the radiation oncology domain for automatic classification of severity levels and hence precludes any direct comparison with prior work.

Limitations

The work presented in this chapter for automatic incident triage in radiation oncology - incident learning system has the following limitations.

First, the method proposed was only able to predict the severity into only High or Low categories, not four as required in the incident learning system.

Second, with this approach, we are unable to type incident, which is significant for making an effective change in the ILS system.

5.6 Conclusion

Incident reports in the radiation oncology domain provide very useful information to analysts and subject matter experts to decide on the right course of action for incidents. With the current trends in digitization of medical data (such as, incident reports) and automation of operations and logistics (such as our proposed automated incident triage and prioritization module), artificial intelligence related methods have become a necessity. In this chapter, we presented a deep learning based ULMFiT model that can effectively identify the incidents based on the initial report and narrative. We demonstrated that this transfer learning based approach outperforms standard supervised machine learning based approaches for classifying narratives. Our work provides encouraging results towards the end goal of a fully

automated incident triage and prioritization system in the future. Additional data from the national safety registry RO-ILS should help to improve the accuracy of our proposed model and provide human-level fidelity and performance. Our models can also work on retrospective data on incident reports to automatically classify the incident severity and provide rapid summarization of past events for subsequent data driven research studies in the future.

Contribution summary: In this chapter we focused on the safety aspects of radiation oncology. We specifically looked at the triage process in incident learning system. Specific contributions of this chapter are as follows.

1. We present an approach to automatically identify the severity of the radiation oncology incidents using the textual incident description.

2. We demonstrate that identifying the severity is a challenging problem when it comes to classifying the incidents into the four possible categories using just the incident description. However, merging severity types into two categories (High and Low severities) results in much better classification results considering the incident report data from multiple VHA radiation oncology centers as well as the VCU medical center datasets.

3. We next demonstrated that transfer learning does help in the severity prediction process specifically considering multi-institution data that may each follow a different protocol for recording the incident reports.

4. We show that incident reports are correlated with institutional practices and there is a need for standardized incident reporting guidelines to reduce the subjective incident analysis practices.

CHAPTER 6

ANALYSIS OF TREATMENT SELECTION PRACTICES FOR INTERMEDIATE OR HIGH RISK PROSTATE CANCER

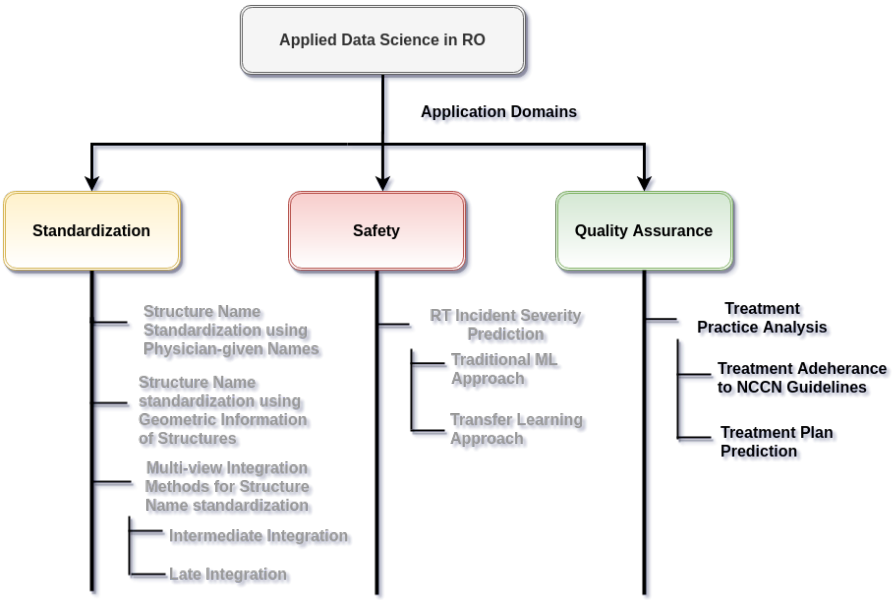


Fig 36: Thesis contribution, Chapter 6 contributions are highlighted.

6.1 Introduction

Prostate cancer (PCa) is the most commonly diagnosed type of cancer after breast and lung cancer. In 2018 alone, over 160,000 new prostate cancer cases and over 29,000 prostate cancer-related deaths were estimated in the United States [79]. PCa is also one of the most heterogeneous type of cancer specifically with respect to intermediate or high-risk PCa [80]. The non-invasive prostate-specific antigen (PSA) test that has led to an increase in early detection of PCa leading to more localized

PCa diagnosis in recent years [81].

The National Comprehensive Cancer Network (NCCN) provides clinical practice guidelines that are created by physicians to determine the best way of treating PCa patients (besides other types of cancers), depending on their diagnosis, disease stage, age and other factors. PCa is also treated with monotherapy or polytherapy. Physicians select the treatment modality based on four major criteria - age, race, life expectancy, and NCCN Risk. Factors such as patient preferences, survivorship goals along with tumor biology also play a crucial role in optimizing the treatment modality.

A major consideration during the treatment options for PCa is to check whether the cancer is contained within the prostate gland (localized), or has spread outside the prostate (locally advanced) or has spread to other parts of the body (metastasized). Radical prostatectomy (RP), external beam radiotherapy (EBRT) and brachytherapy (BT) are the common primary treatment options for localized PCa. Hormonal therapeutics such as androgen deprivation therapy (ADT) is also used as neoadjuvant/adjuvant therapy. However, ADT as monotherapy is not recommended for intermediate and high-risk cancer patients by NCCN. Ideally, a treatment option recommendation would be based on the randomized controlled trials (RCT) that compare efficacy and morbidity of alternative treatment methods. There are no randomized trials showing that one treatment is better than the other for the above-mentioned treatment options. Hence, physicians use their personal experience and expertise to predict the outcome of these treatment methods. Physicians also tend to have difficulty weighing the relative importance of each of these factors and inherently possess biases when predicting the treatment outcomes.

Based on the aforementioned considerations, determining an optimal treatment plan for the patient can be a challenging task for the physician. In order to assist

the physicians with more accurate prognosis, subsequent treatment outcome prediction, and to make informed decisions, numerous predictive tools have been developed [82]. These include probabilistic models, lookup and propensity scoring tables, risk-stratification tools, classification, and regression tree analysis, nomograms, and artificial neural networks[83, 84, 85, 86, 87, 88, 89, 90, 91, 92]. However, to the best of our knowledge, no models have been reported that can identify why a prescribed (or administered) treatment plan do not adhere to NCCN guidelines.

The predictive models for treatment plan (or outcome) prediction have a major disadvantage. Such models do not consider the impact of non-clinical factors associated with the treatment center. The factors associated with the treatment center have shown to play a determining role in the physicians' treatment prescription practices. Non-clinical factors can be patient-related, physician-related or practice-related. These factors include patient's preference/availability, patients' adherence, physician's availability, cost, geographical proximity, treatment centers' equipment condition/availability, treatment centers' cultural aspects, type of practice (private vs. public), availability of health resources, etc.[93, 94, 95, 96]. However, there have not been many studies which have investigated the extent of the contribution of these factors in the treatment selection process itself. Thus the motivation of this study is two-fold:

1. To use both clinical and non-clinical features for localized and locally advanced PCa patients from multiple Veterans Health Administration (VHA) centers and use machine learning methods to predict the treatment prescribed; such methods provide a statistical approach for calculating the weight (impact) of these clinical/non-clinical features from an empirical and retrospective point-of-view.

2. To perform quality assurance assessments across the different centers and verify

if the prescribed treatments were in concordance with NCCN guidelines.

This study presents a comparative analysis of treatment prescription consistency across multiple VHA centers.

6.2 Materials and Methods

6.2.1 Dataset

The VHA has 40 centers treating cancer patients with radiation therapy (RT) across the US. But for this study, a maximum of 20 patients from 34 VHA RT centers are selected based on the whose treatment was completed below criteria.

- Patients should have been treated between 2010 to 2017.
- Patients must have been treated for intermediate or high-risk PCa.
- Patients must not have previous malignancy, M1 disease, or lymph node involvement.

A total of 552 patients from the 34 centers were selected. Subject matter expert (radiation oncology nurse) gone through all health records to manually extract the related clinical information. Hence, we consider this dataset as a gold dataset. Table 24 show the dataset details.

The dataset was split 80 : 20 ratio into training and test sets. *One hot encoding technique* was used to binarize the categorical features, this technique simply creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature. Continuous features were scaled to a min-max scale (for normalization). We used random forest algorithm for building predictive models. Models are evaluated with macro-average precision, recall, and F_1-Score.

6.2.2 Definitions of Variables

Definitions of variables used in our study are as follows.

Clinical variables: We considered pre-treatment PSA count, Gleason score (GS) [primary grade, secondary grade], Gleason Grade, Tumor staging [TNM-stage], NCCN risk group, performance status, and quality of life (QoL) measures. The values for these clinical variables were manually extracted from the consult notes. :

Non-Clinical Variable: We defined Center-ID as a non-clinical variable. It designates a unique ID to identify the VA radiation treatment center.

ADT Duration: NCCN guidelines define ADT duration as short term (ST) or long term (LT). ST duration is 4-6 months, and LT duration is 2-3 years. We further differentiated ADT duration based on intended and administered duration. The intended duration signifies whether it was mentioned in consult notes during treatment planning, whereas ADT administered duration is calculated based on the dates of ADT injection. Table 23 shows the ADT injection type and their effective period in months depending on the dose. Table 25 shows the distribution of ADT intended and administered duration. A third category of not otherwise specified (NOS) was used to indicate cases where ADT duration was not mentioned in consult as a treatment plan.

Treatment Prescribed: During the consultation, the radiation oncologist discusses with the patient all possible treatment by explaining the side effects of each of the treatment plan. The decision is taken with the patient, and this intended treatment is recorded in consult notes. Hence, we call this treatment intended at the time of consult as treatment prescribed.

Treatment Administered: At the end of the treatment, radiation oncologists make a note of treatment details. We call this treatment as treatment administered.

We have used two different terminologies because, for some patients, there is a change in from intended to administration treatment. This change in treatment was mainly observed in the ADT duration.

NCCN Concordance: We defined the treatment prescribed or administered is concordant with NCCN guidelines if they were as per the NCCN guidelines [97].

ADT Injection	Dose	Effective Period
Leuprolide	3.75 mg	1 month
	7.50 mg	1 month
	22.50 mg	3 months
	30.00 mg	4 month
	45.00 mg	6 months
Goserelin/Zoladex	3.60 mg	1 month
	10.80 mg	3 months

Table 23: ADT Injection Effective period based on the injection type and dose.

6.2.3 Model Selection

In this section we present the details of feature-set selection, predictive models, machine learning algorithms, and model evaluation metrics.

We used machine learning algorithms as a statistical tool to find the association between the treatments and clinical and non-clinical features. We used a supervised machine learning algorithm called random forests (RF) [20], to find these associations. The RF algorithm takes the features (clinical and non-clinical variables) and target (treatments) to builds multiple decision trees and merges them together to get a more accurate and stable prediction. It also provides the significance of features in

Data Element	Count	Percentage
Total Patients	552	-
Centers	34	-
Gleason Score		
Primary + Secondary	549	99.50
3 + 3	17	3.00
3 + 4	219	39.67
4 + 3	128	23.18
3 + 5	18	3.26
4 + 4	79	14.31
5 + 3	2	0.36
4 + 5	61	11.05
5 + 4	19	3.44
5 + 5	3	0.54
NOS + NOS	2	0.36
PSA	549	99.50
T Stage	549	99.50
T1a - T2a	457	82.79
T2b - T2c	64	11.59
T3a -T3b	20	3.63
TX	1	0.18
NOS	7	1.26
Risk	545	98.73
Intermediate	304	55.60
High	241	44.40
Performance Status	523	94.75
Quality of Life	400	72.46
Treatment Prescribed	552	100.0
BT	24	3.07
BT-ADT	1	0.13
EBRT	132	20.23
EBRT-ADT	382	59.28
EBRT-BT	2	0.27
EBRT-BT-ADT	11	2.00

Table 24: Details of the clinical factors in the VHA ROPA dataset and their distribution, NOS: Not Otherwise Specified.

classifying the targets. The significance of all features sums to 1, where higher the significance of a feature stronger is its association with the target class, and lower significance indicates the weaker or no association.

NCCN Risk	Treatment	ADT Duration	Intended	Administered	Concordance with NCCN
Intermediate	ADT-BT	NS	1	-	No
		LT	-	1	Yes
	BT		24	24	Yes
	EBRT		115	115	Yes
	EBRT-ADT	LT	8	15	No
		NS	11	-	No
		ST	142	146	Yes
	EBRT-ADT-BT	ST	1	1	Yes
	EBRT-BT		2	2	Yes
High	EBRT		17	17	No
	EBRT-ADT-BT	LT	9	4	Yes
		ST	1	6	Yes
	EBRT-ADT	LT	185	145	Yes
		NS	18	-	No
		ST	12	70	No

Table 25: Treatment concordance with NCCN guidelines. ST :Short Term, LT: Long Term, and NS: Not Specified.

6.2.3.1 Features and Labels

We created two feature sets using the clinical and non-clinical features to highlight the contribution of non-clinical features. The feature sets (FS) are as below

1. FS-1: Clinical features only. (PSA, Risk, Total GS, Primary GS, Secondary GS, T_Stage)

2. FS-2: Clinical and Non-clinical (Center-ID) features. (PSA, Risk, Total GS,

Primary GS, Secondary GS, T_Stage, Center-ID)

Above two feature sets used in two models with target labels as below:

1. Model-1 Labels: EBRT-ADT, ADT
2. Model-1 Labels: EBRT-ADT-ST, EBRT-ADT-LT

In below section we explain the two models we have built with combinations of feature sets nd labels.

6.2.3.2 Statistical Models

VA-ROPA dataset has patients treated with six different treatment methods (Table 24): BT, BT-ADT, EBRT, EBRT-ADT, EBRT-BT, and EBRT-BT-ADT. Based on the available treatment plans, we built the following two models.

1. Model-1: Initial Treatment (EBRT-ADT vs EBRT): This model predicts whether the patients will be treated with EBRT and ADT (EBRT-ADT), or EBRT alone. A total of 514 patients were treated with these two techniques, among which 382 patients were treated with EBRT-ADT, and 132 patients were treated with EBRT alone.
2. Model-2: ADT Duration (EBRT-ADT-ST vs EBRT-ADT-LT): This model predicts whether the ADT duration is *short term* or *long term*. Model-2 is further divided into 2A and 2B. Where 2A is EBRT with ADT intended duration and 2B is EBRT with ADT administered duration. 382 patients were treated with EBRT and ADT. Table 25 shows the treatment with intended and administered ADT duration.

These models use machine learning techniques to serve the dual purpose of (i) creating a predictive model of initial treatment selection or ADT duration based on the clinical

and non-clinical features and (ii) showing the statistical correlation of the individual features in terms of impacting the treatment selection or ADT duration process.

6.3 Results

In this section, we present our results. Table 26 shows the Precision, Recall, F_1-Score for model-1 (EBRT-ADT vs EBRT). The goal in this model was to classify patients with treatment intent being either EBRT or a combination of EBRT and ADT (EBRT-ADT). Model 1 with FS-2 performed better in all metrics when compared to FS-1. We observed that model-1 has F_1-Score of 74% with FS-1 and 82% with FS-2. These results clearly demonstrate the significance of non-clinical feature (Center-ID) in improving the overall classification performance.

Model	**ADT Duration**	**F-Set**	**Precision**	**Recall**	**F_1-Score**
Model 1	-	FS-1	0.75	0.73	**0.74**
		FS-2	0.82	0.82	0.82
Model 2A	Intended	FS-1	0.95	0.94	**0.94**
		FS-2	0.92	0.92	0.92
Model 2B	Administered	FS-1	0.74	0.73	**0.73**
		FS-2	0.72	0.71	0.71

Table 26: Macro-averaged Precision, Recall, F_1-Score, for Model-1:(EBRT-ADT vs EBRT), Model-2: (EBRT-ADT-ST vs EBRT-ADT-LT) 2A:ADT Intended Duration, 2B:ADT Administered Duration.

Table 26 also shows the results of model-2 (EBRT-ADT-ST vs EBRT-ADT-LT). Interestingly, in this case, FS-1 and FS-2 perform quite similarly with 94% F_1-Score

FS	Features	Model 1	Model 2A ADT Intent	Model 2B ADT Administered
	PSA	**0.52**	0.14	**0.39**
	Risk	0.25	**0.79**	0.30
FS-1	Total GS	0.03	0.04	0.14
	T_stage	0.09	0.02	0.07
	Primary GS	0.06	0.01	0.05
	Secondary GS	0.05	0.01	0.05
	PSA	0.23	0.08	0.24
	Risk	0.28	**0.79**	**0.27**
	Total GS	0.02	0.03	0.19
FS-2	T_stage	0.07	0.02	0.05
	Primary GS	0.13	0.02	0.04
	Secondary GS	0.04	0.02	0.04
	Center ID	**0.29**	0.06	0.17

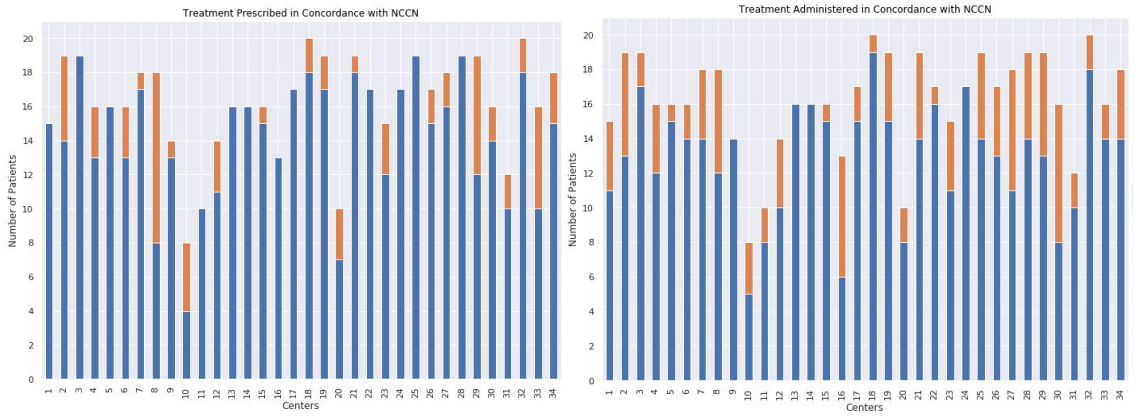
Table 27: Feature importance in each model. Model 1:EBRT-ADT vs EBRT, Model 2A: ADT course intended, Model 2B: ADT course Administered. FS:Feature Set.

for models with ADT intent labels (with FS-1), while F_1-Score is decreased when the ADT administered labels were used. This may mean that some external factors (not considered in our feature sets) play a role for causing the alteration from treatment from the prescribed to administered. Also, non-clinical feature (Center-ID) found to have no affect on predicting the ADT duration type as opposed to Model-1 (EBRT-ADT vs EBRT). Based on these observations, we hypothesize that while centers do play a role in determining whether to prescribe ADT or not, they do not impact the

actual ADT duration, in case it was administered; in other words, all centers follow similar practice in administering ADT for localized intermediate or high-risk PCa treatment.

We next evaluated the individual significance (i.e., contributions) of each of the features from FS-1 and FS-2 in our models; the feature significance were generated using the RF algorithm. Table 27 shows the feature importance of all features in all models. For both FS-1 and FS-2, PSA and Risk consistently ranked as significant features in all the models. Specifically, for FS-1, PSA was ranked as the top feature for Models 1, 2B . For Model-2A (ADT duration intended), Risk was ranked as the top feature. This suggests that decisions on ST or LT ADT duration depend primarily on the Risk with PSA being a secondary feature of importance; these two features are primarily responsible in deciding the ADT course at the initial treatment level; however, decisions in altering the treatment intent (as captured in Model-2B with treatment administered) are impacted by the PSA and Total Gleason score (which is the third ranked feature in this model). For Model-1, PSA was ranked as the top feature with Risk as the secondary feature and T_stage as the third significant feature suggesting that decisions on treating the patients with EBRT alone or a combination of EBRT and ADT depend primarily on the Risk, PSA, and T_stage values.

When we considered FS-2, PSA and Risk show similar significance. In this case however, Center-ID plays a crucial role and shows up specifically as the top ranked feature in Model-1 (EBRT-ADT vs EBRT); this reconfirms our earlier hypothesis that nonclinical factors like the center play a significant role in determining whether patients undergo ADT treatment or not. However, it's significance is much lower in Model-2A (EBRT-ADT-ST vs EBRT-ADT-LT) with ADT intended duration. Center-ID also shows up as the fourth ranked feature in Model-2B (EBRT-ADT-ST vs EBRT-ADT-LT) for ADT duration administered; thus we can hypothesize that



(A) (B)

Fig 37: Treatments in concordance with NCCN when all treatments are considered at each center. Blue: treatments in concordance, Orange: not in concordance. (A): Treatments prescribed at each center when ADT intent course is considered along with all other treatments; (B): Treatments administered at each center when ADT administered course is considered along with all other treatments.

nonclinical factors may have a role to play in altering the treatment intent.

We observed that treatment non-concordance with NCCN guidelines can be due to the following two reasons:

- Firstly, overall treatment may not be in concordance with NCCN guidelines. For example, high-risk cancer patients treated with EBRT alone are not in concordance with NCCN. Figure 37 (A) & (B) shows the center wise all non-concordant treatment counts based on ADT intended duration (i.e., prescribed ADT) and ADT administered duration treatments respectively.

- Secondly, overall treatment is in concordance with NCCN however the treatment guidelines may be partially not followed. For example, a high-risk cancer patient is treated with EBRT and ADT, but ADT duration is for short-term instead

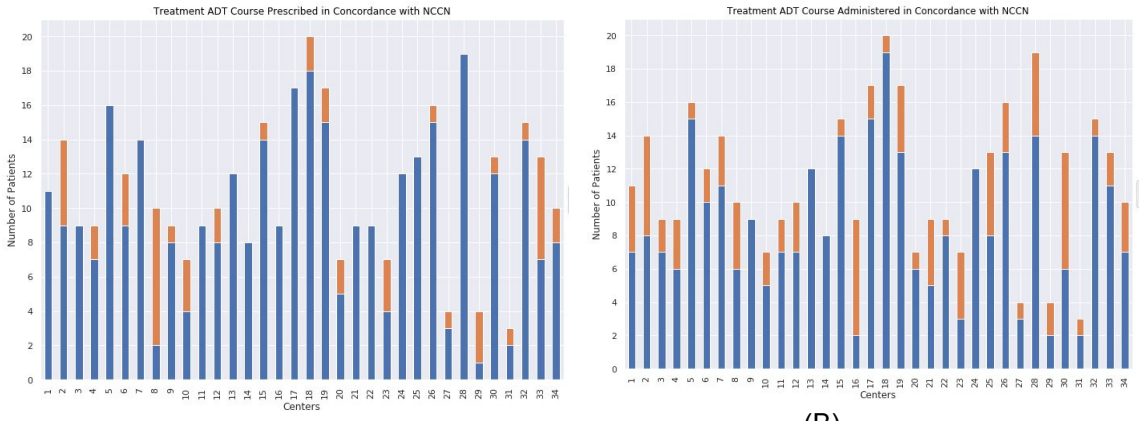


Fig 38: Patients treated with EBRT and ADT (Short Term or Long Term). Blue: number of patients whose treatments are in concordance with NCCN, Orange: number of patients whose treatments are partially not in concordance with NCCN (A): Treatments prescribed at each center when ADT intent course is considered (B): Treatments administered at each center when ADT administered course is considered.

of long-term. Figure 38 (A) & (B) shows the partially non-concordant patient count of each center when patients are treated with EBRT and ADT; the counts are again based on the ADT intended and administered duration respectively.

6.4 Discussion

In this study, we present an exploratory analysis of localized or locally advanced PCa patients from 34 different VHA treatment centers. We compared the treatments prescribed against the NCCN guideline recommendations and observed that most of the treatment plans (prescribed or administered) matched with the NCCN guidelines. We built machine learning based models to predict the treatment plans for patients and also the likelihood of NCCN concordance of their treatment plans. We observed

that PSA and Risk were the top-ranked features in determining the treatment plans for PCa patients.

Center-ID improved the performance of the model's that predicts if the selected treatment plan has ADT or not; however, it did not impact the models that predict if the prescribed ADT duration was ST or LT. We also observed some variability in ADT treatments prescribed versus actual ADT treatments administered; the Center-ID, however, had a negligible role to play in such alterations and instead PSA and total Gleason score had significant roles to play in such decisions. We also noticed that the performance status measure had a negative effect on model predictability and hence we dropped it from our feature set. We feel that performance status will be a critical feature in treatment outcome predictions in the future, currently which is outside the scope of this work. Additionally, Risk showed up as the primary feature in predicting ST vs. LT ADT duration. We also observed that the primary reason for treatment plans to be non-concordant with NCCN is due to the ADT course duration not following the guidelines.

To better understand the impact of non-clinical features like Center-ID in predicting whether the treatment plans were concordant with NCCN guidelines or not, we computed the Pearson correlation between center-specific details (such as staffing details) and the number of non-concordant patients undergoing EBRT-ADT or EBRT-only treatments (either prescribed or administered). Figure 39 shows a small negative correlation between staff details and non-concordance; specifically fewer number of radiation oncologists or radiation therapists led to higher number of non-concordant patients in all cases; while the number of radiation physicists or other staff members did not show any worthwhile correlation. This can be potentially attributed to higher workloads and scheduling conflicts for radiation oncologists/therapists leading to non-adherence to ADT treatment duration requirements from NCCN.

Figure 39 also shows the impact of Center-ID in predicting whether a patient will undergo EBRT-only or EBRT-ADT treatment. We can observe a strong positive correlation between EBRT-only treatment selection and the number of radiation therapists and a less pronounced positive correlation between EBRT-ADT treatment selection and the number of radiation oncologists. While this positive correlation was expected as more radiation oncologists or therapists will lead to more patients being treated with EBRT-ADT or EBRT-only respectively, it is however not clear why the number of radiation physicists or other staff members correlates poorly with these treatment types. It can arise from the bias of the selected patient cohort.

Our findings corroborate previous studies showing the impact of non-clinical factors on prostate cancer treatment patterns. For example, a recent study done on SEERs data reported that prostate cancer treatment patterns were not strictly influenced by outcomes data and varied significantly by patient age, insurance status, financial model, regional bias and socioeconomic factors [98]. An earlier survey on factors influencing treatment selection for localized prostate cancer suggests that recognizing the beliefs that patients hold about their cancer and its treatment could guide the counseling of patients about the treatments available to them and ultimately, help patients make more informed decisions about both their treatments and subsequent adjustments [99]. Prior work on NCCN non-concordance was conducted on elderly patients with high-risk prostate cancer from SEERs was reported that NCCN concordance in elderly patients with aggressive prostate cancer is low [100]. These findings underline the importance of non-clinical factors in treatment decisions, however, reported results were based on single center data; hence they could not identify the center-specific bias. However, such non-clinical factors can vary appreciably between multiple centers and result in the bias; our future work will include such non-clinical features from the VHA centers to identify the proper reasons behind

such center-specific bias.

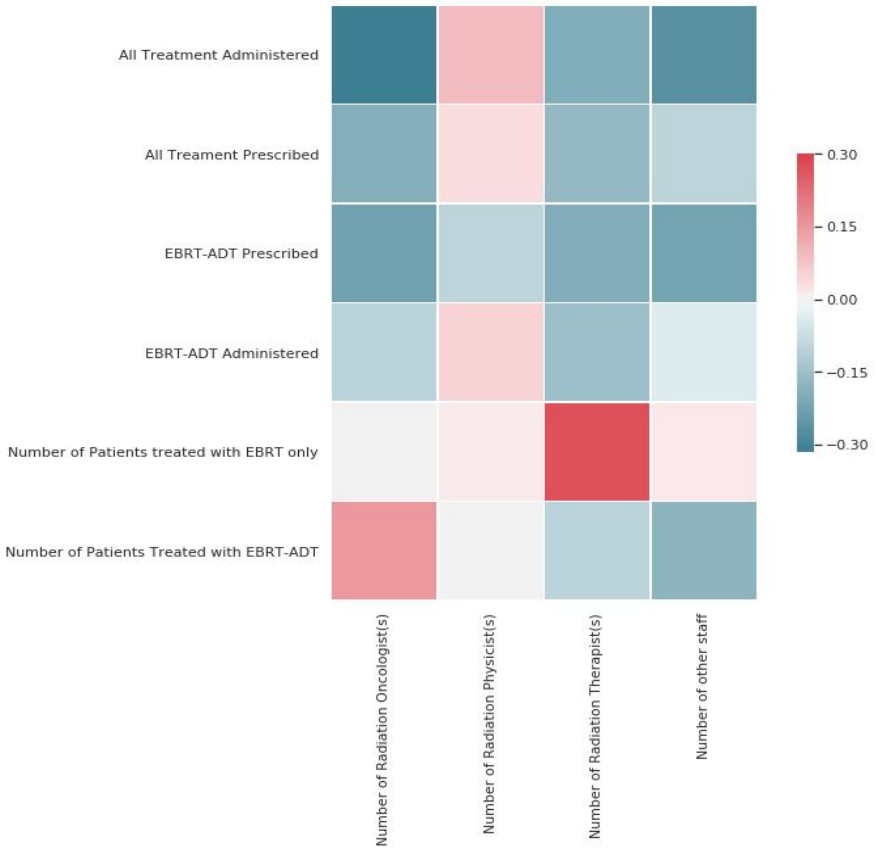


Fig 39: Pearson correlation between center details (Number of radiation oncologists, radiation physicists, radiation therapists and Other staff), and (i) treatment non-concordance (number of non-concordant patients considering all treatments prescribed, all treatments administered, EBRT-ADT prescribed, and EBRT-ADT administered), and (ii) treatment selections (number of patients treated with EBRT-only or with EBRT-ADT).

Limitations

This work has following limitations. First, in data collected for this work includes patients treated with EBRT only or EBRT with ADT. There are other modalities such as Brachytherapy and Surgery. Patients treated with all modalities will provide a better understanding treatment selection practices. Second, we have analyzed overall treatment selected not the sequence of treatment given in multi-modality treatments. Third, maximum of only 20 patients were considered from each of 34 RT treatment centers, which is small number of patients as representative for analysis of treatment selection practices.

6.5 Conclusion

The VHA ROPA dataset was extracted from recently treated patients having very little to no follow-up data for oncological outcome analysis. Similar predictive models will be built in the future for treatment outcome analysis considering a patient cohort that was treated at earlier dates. Additionally, the ADT duration is generally dependent on the type of drugs used. In this study, we calculated ADT administered duration based on the ADT injection dates; the calculated ADT duration may slightly change considering the ADT injection types. Finally, our study depicts the importance of non-clinical factors, such as Center-ID, in predictive models for treatment selection or concordance to NCCN guidelines. In the future, we will investigate the effects of other types of non-clinical factors (not limited to staffing) pertinent to the specific VHA centers considered here.

Contribution summary: In this chapter, we considered the treatment quality component of the radiation therapy process and our specific contributions are as below.

1. We present feature engineering methods to analyze the treatment selection practices for High or Intermediate risk prostate cancer patients across 34 different VHA radiation therapy centers.

2. We demonstrate that there is an inherent bias in the treatment selection process at the VHA treatment centers. The selected treatments deviate from the NCCN guidelines and there is little to no correlation for this deviation with specific treatment center attributes such as, number of radiation oncologists, radiation therapists, other staff or treatment resources.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

In this dissertation, we have investigated different data science approaches for standardization, safety, and quality assurance in radiation oncology.

For data standardization, in Chapters 3 and 4, we have presented a novel multi-view machine learning approach to standardize the radiotherapy structure names. We considered two views of RT structure data individually, namely, the physician-given structure names and the imaging based geometric features. For the text classification problem, we observed that considering only the fastText algorithm works best when compared to other feature weighting and classification algorithms. Our method was evaluated with the data from 40 VA radiotherapy centers and tested on an external dataset from VCU. We demonstrated that our text classification method works well on multiple disease sites and is also generalizable. To the best of our knowledge, this is the first and the only model using the physician-given name to predict the TG-263 standard name using NLP and ML based methods. We also observed that our approach fails in certain conditions, when enough information is not available for the model to infer the correct label. This text-classification approach was next augmented with imaging information, such as geometric information of structures to build a multi-view pipeline for structure name standardization which improved the overall accuracy of our methods. We believe that the proposed structure names standardization methods can help with big data analytics in the radiation therapy domain using population-derived datasets, including standardization of the treatment

planning process, clinical decision support systems, treatment quality improvement programs, and hypothesis-driven clinical research.

For patient safety, in Chapter 5, we analyzed the incident reports from the radiation oncology domain that provide beneficial information to analysts and subject matter experts to decide on the right course of action. The current trends in digitization health care (such as incident reports) and automation of operations and logistics (such as our proposed automated incident triage and prioritization module), machine learning methods have become necessary. In this chapter, we compared the traditional machine learning and transfer learning approaches to automatically identify the severity of the RT incident based on the incident description. We demonstrated that this transfer learning using the ULMFiT algorithm outperforms a standard supervised machine learning-based approach. With the limited data, our approach provided encouraging results towards the end goal of a fully automated incident triage and prioritization system in the future. Additional data from the national safety registry RO-ILS should help improve our proposed model's performance. Our models can also work on retrospective data on incident reports to automatically classify the incident severity and provide rapid summarization of past events for subsequent data-driven research studies. There are no specific guidelines on incident reporting practices, specifically the structure of the incident description. Hence, the length of the incident descriptions varied depending on the severity types and across institutions (VCU and VHA). Thus, we believe there is a need and opportunity to build guidelines on incident reporting practices.

For quality assurance, in Chapter 6, we presented a machine learning pipeline to assess the treatment quality for prostate cancer patients considering clinical datasets from both VHA and VCU. The goal of this work was to build a predictive model for assessing whether radiation therapy treatment plans adhere to the NCCN guidelines

or not. We additionally observed that non-adherence to NCCN standards did not exhibit any correlation with the radiation therapy center-specific features, such as the number of radiation oncologists, therapists, physicists, and other staff. However, the treatment plan prediction models exhibited a center-specific bias demonstrating that individual RT-centers exercise their own preference in choosing the treatment plans. However, the identification of exact features that affect these preferences is part of our future work.

7.2 Future Work

In Chapter 3, we presented the structure name standardization pipeline while in Chapter 4, we presented different methods to integrate the heterogeneous radiotherapy structure data for structure name standardization. We next outline the following future works for the structure name standardization problem.

- In the Late integration approach, we have used the top 100 SVD features with an RF classification algorithm. However, there are more suitable algorithms for image data such as 2D CNN algorithm, ResNet [101], and VoxNet and a 3D CNN supervised classification algorithm [102]. The radiotherapy structure set is 3D in nature, making it more suitable to solve using 3D algorithms.

- Our structure name standardization ML pipeline, from data preprocessing to prediction, works as a standalone system. We plan to create a seamless enterprise informatics platform that automatically collects data from the treatment planning systems and performs automatic structure name standardization on retrospective data and stores the standardized names back in databases.

- The current list of OARs identified for both lung and prostate datasets is per the VA-ROQS project requirement, which has selected these OARs in consensus

with a team expert. Radiation oncologists also delineate other types of OARs for each patient, such as Kidney (left and right) and Liver, in prostate cancer patients. Although these are not critical OARs in prostate cancer treatment, we believe building a system to identify and standardize all structures delineated according to the TG-263 guideline provides the radiation therapy healthcare institutes with an opportunity to produce a robust dataset for downstream analysis projects.

- Other future works using the standardized structure sets include dose outlier detection, toxicity prediction, treatment outcome analysis, treatment planning, automated structure delineations.

In Chapter 5, we presented an approach to automatically identify the severity of the radiotherapy incident reports based on the textual description provided in the radiotherapy incident reporting and analysis system (RIRAS). For Chapter 5, we outline the following future work.

- We have used ULMFiT in our current work for the transfer learning method. Ther are other contextual word embedding algorithms, ELMo [103], OpenAI GPT [104], and BERT [105]. In the biomedical domain, researchers have fine-tuned BERT LM models (SciBERT, clinicalBERT [106], and BioBERT [107]) and reported better performance on downstream tasks over the standard BERT model. We can integrate similar approaches into the transfer learning model. Although there are pre-trained biomedical domain-specific BERT based language models, which are closely related to radiation oncology, we still believe that training a radiation oncology-specific BERT model is needed. The national registry of the radiation oncology incident - learning system (RO-ILS) collects the incidents and analysis reports submitted from radiation oncology

institutes across the USA. We believe that fine-tuning the SciBERT, bioBERT, clinicalBERT, and BERT-base separately, and comparing the performance of downstream tasks provide the understanding of the model's dependency on domain knowledge.

- Incident analysis involves many other steps along with the severity assessment, such as identifying the incident process step and providing the short and appropriate title to the analyzed report. The title of the analysis report needs to represent the issue reported. We believe that a fine-tuned BERT model will give better results for this task. Another vital work will be to identify similar reports in the incident database and recommend the solution based on the previously analyzed reports.

- Understanding why incidents occur may be more critical for effecting change than understanding what events have occurred. Further studies exploring NLP's ability to classify incident reports by contributory factors could offer more learning opportunities. We believe contextual topic modeling would be beneficial for determining the contributory factors.

- In the current RIRAS dataset, one SME assigns severity to the incident reported based on the incident description. Incident analysis is a highly subjective task; to reduce the subjectiveness and make it more objective, we believe each report must be analyzed by two or more SMEs independently. The inter-annotator agreement score needs to be calculated to understand the subjective biases in reviewers. Addressing these biases will generate more consistent incident analysis reports and offer more appropriate severity labels for automated severity assignment models.

Finally, we discuss the future work for Chapter 6.

- We explored the multi-center treatment selection practices. In current work, we have analyzed the treatment selection practices. However, the treatments selected were multi-modality treatments. A plausible future work is to analyze the treatment selection paths and their association with patient pre-treatment attributes and outcome analysis based on the treatment path selection.

Appendix A

ABBREVIATIONS

AAPM	Association of Physicists in Medicine
AJCC	American Joint Committee on Cancer API Application Programing Interface
ASTRO	American Society for Radiation Oncology
DICOM	Digital Imaging and Communications in Medicine
ESTRO	European Society for Therapeutic Radiation Oncology
EHR	Electronic Health Record
HIPAA	Health Insurance Portability and Accountability Act
JSON	JavaScript Object Notation
NCCN	National Comprehensive Cancer Network
PACS	Picture Archive and Communication Systems
RIRAS	Radiotherapy Incident Reporting and Analysis System
RCT	Randomized Controlled Trial
RO-ILS	Radiation Oncology Incident Learning System
RT	Radiation Therapy
TPS	Treatment Planning System

Appendix B

STRUCTURE NAME STANDARDIZATION WITH PHYSICIAN-GIVEN NAMES

Fig 40: Radiotherapy Structure name distribution per center for Prostate cancer patients in the VA-ROQS dataset.

FIGURE Page

Fig 41: Radiotherapy Structure names distribution per center for Lung cancer patients in the VA-ROQS dataset.

FIGURE Page

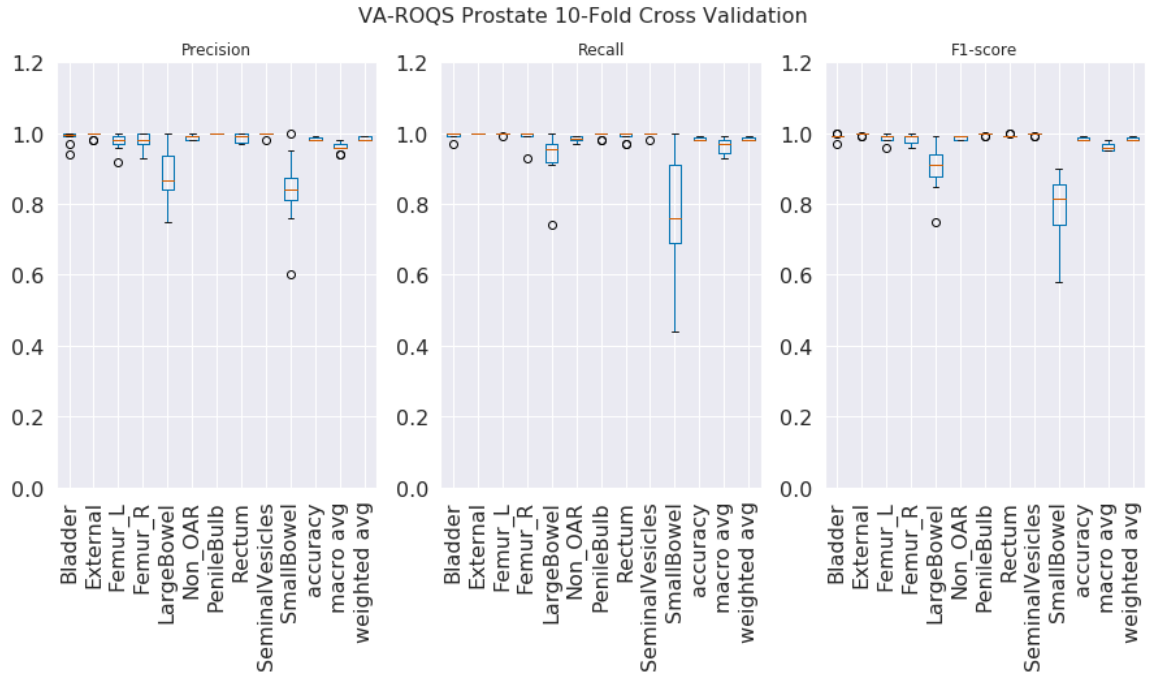


Fig 42: VA-ROQS Prostate 10 fold cross-validation results

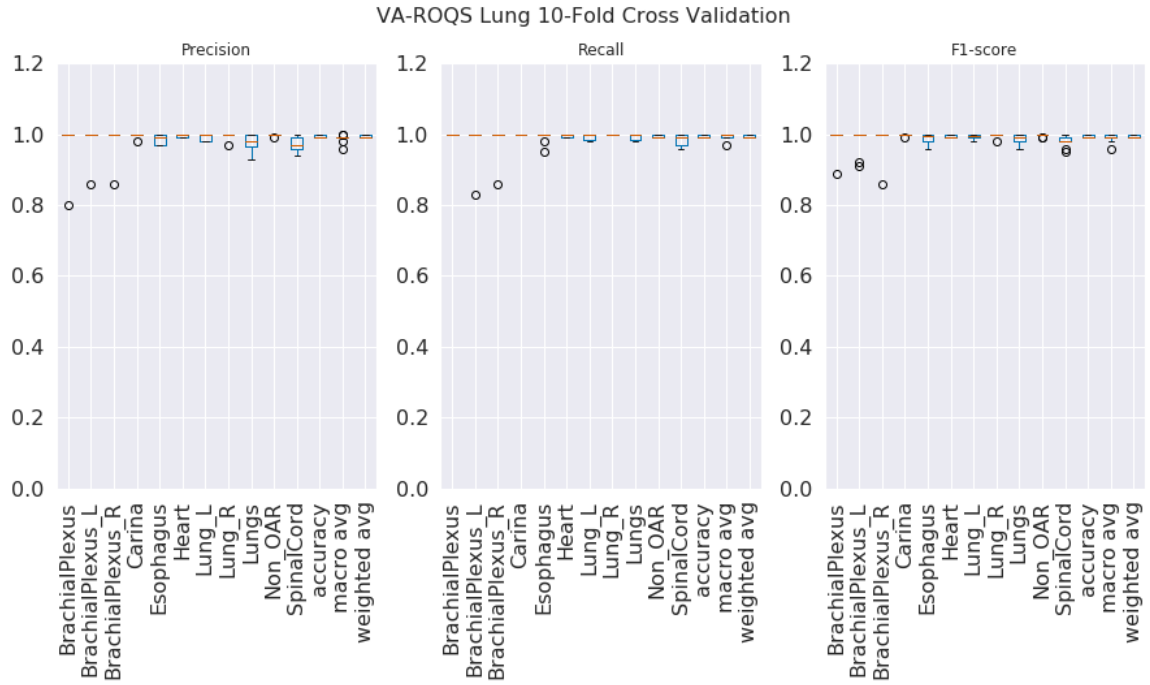


Fig 43: VA-ROQS Lung 10 fold cross-validation results.

FIGURE Page

Structure Name	Precision	Recall	F_1-Score	Support
Bladder	0.99	0.97	0.98	152
External	1.0	1.0	1.0	119
Femur_L	0.99	1.0	1.0	141
Femur_R	1.0	0.99	0.99	145
LargeBowel	0.9	0.87	0.88	70
Non_OAR	0.99	0.99	0.99	1970
PenileBulb	0.99	1.0	1.0	117
Rectum	0.99	0.99	0.99	148
SeminalVesicles	1.0	0.99	1.0	103
SmallBowel	0.82	0.85	0.84	48
accuracy	0.99	0.99	0.99	3013
macro avg	0.97	0.97	0.97	3013
weighted avg	0.99	0.99	0.99	3013

Table 28: VA-ROQS Prostate 70:30 validation results.

Structure Name	Precision	Recall	F_1-Score	Support
Bladder	0.96	0.99	0.98	738
External	1.0	1.0	1.0	597
Femur_L	0.95	0.98	0.97	711
Femur_R	0.95	0.95	0.95	717
LargeBowel	0.86	0.89	0.87	341
Non_OAR	0.98	0.98	0.98	9869
PenileBulb	1.0	1.0	1.0	590
Rectum	0.97	0.99	0.98	742
SeminalVesicles	1.0	1.0	1.0	510
SmallBowel	0.77	0.64	0.7	250
accuracy	0.97	0.97	0.97	15065
macro avg	0.94	0.94	0.94	15065
weighted avg	0.97	0.97	0.97	15065

Table 29: VA-ROQS Prostate Center validation results.

FIGURE Page

Structure Name	Precision	Recall	F_1-Score	Support
Bladder	0.99	0.99	0.99	738
External	1.0	1.0	1.0	597
Femur_L	0.97	1.0	0.99	711
Femur_R	0.98	0.99	0.98	717
LargeBowel	0.87	0.93	0.9	341
Non_OAR	0.99	0.99	0.99	9869
PenileBulb	1.0	1.0	1.0	590
Rectum	0.99	0.99	0.99	742
SeminalVesicles	1.0	1.0	1.0	510
SmallBowel	0.85	0.73	0.79	250
accuracy	0.98	0.98	0.98	15065
macro avg	0.96	0.96	0.96	15065
weighted avg	0.98	0.98	0.98	15065

Table 30: VA-ROQS Prostate dataset 5 fold validation results.

Structure Name	Precision	Recall	F_1-Score	Support
Bladder	0.99	0.99	0.99	738
External	1.0	1.0	1.0	597
Femur_L	0.97	1.0	0.99	711
Femur_R	0.98	0.99	0.98	717
LargeBowel	0.87	0.93	0.9	341
Non_OAR	0.99	0.99	0.99	9869
PenileBulb	1.0	1.0	1.0	590
Rectum	0.99	0.99	0.99	742
SeminalVesicles	1.0	1.0	1.0	510
SmallBowel	0.81	0.76	0.79	250
accuracy	0.98	0.98	0.98	15065
macro avg	0.96	0.97	0.96	15065
weighted avg	0.98	0.98	0.98	15065

Table 31: VA-ROQS Prostate dataset 10 fold validation results.

FIGURE Page

Structure Name	Precision	Recall	F_1-Score	Support
BrachialPlexus	1.0	1.0	1.0	9
BrachialPlexus_L	1.0	1.0	1.0	12
BrachialPlexus_R	1.0	1.0	1.0	14
Carina	1.0	1.0	1.0	99
Esophagus	1.0	0.99	1.0	128
Heart	1.0	0.99	0.99	141
Lung_L	0.99	1.0	1.0	110
Lung_R	1.0	0.99	1.0	113
Lungs	1.0	0.96	0.98	92
Non_OAR	0.99	1.0	1.0	1750
SpinalCord	0.99	0.97	0.98	141
accuracy	1.0	1.0	1.0	2609
macro avg	1.0	0.99	0.99	2609
weighted avg	1.0	1.0	0.99	2609

Table 32: VA-ROQS Lung dataset 70:30 validation results.

Structure Name	Precision	Recall	F_1-Score	Support
BrachialPlexus	0.57	0.86	0.68	44
BrachialPlexus_L	0.97	0.56	0.71	59
BrachialPlexus_R	0.9	0.94	0.92	69
Carina	1.0	1.0	1.0	497
Esophagus	0.98	0.99	0.99	636
Heart	0.98	0.99	0.99	693
Lung_L	0.99	0.97	0.98	555
Lung_R	0.98	0.98	0.98	563
Lungs	0.97	0.98	0.97	439
Non_OAR	0.99	0.99	0.99	8800
SpinalCord	0.96	0.97	0.96	689
accuracy	0.99	0.99	0.99	13044
macro avg	0.94	0.93	0.93	13044
weighted avg	0.99	0.99	0.99	13044

Table 33: VA-ROQS Lung dataset Center validation results.

FIGURE Page

Structure Name	Precision	Recall	F_1-Score	Support
BrachialPlexus	0.98	0.91	0.94	44
BrachialPlexus_L	0.92	0.98	0.95	59
BrachialPlexus_R	0.99	0.99	0.99	69
Carina	1.0	1.0	1.0	497
Esophagus	0.99	1.0	0.99	636
Heart	0.99	1.0	1.0	693
Lung_L	0.99	0.99	0.99	555
Lung_R	0.99	1.0	1.0	563
Lungs	0.98	0.99	0.99	439
Non_OAR	1.0	0.99	1.0	8800
SpinalCord	0.97	0.98	0.98	689
accuracy	0.99	0.99	0.99	13044
macro avg	0.98	0.98	0.98	13044
weighted avg	0.99	0.99	0.99	13044

Table 34: VA-ROQS Lung dataset 5 fold validation results.

Structure Name	Precision	Recall	F_1-Score	Support
BrachialPlexus	0.98	1.0	0.99	44
BrachialPlexus_L	0.98	0.98	0.98	59
BrachialPlexus_R	0.99	0.99	0.99	69
Carina	1.0	1.0	1.0	497
Esophagus	0.99	0.99	0.99	636
Heart	0.99	1.0	0.99	693
Lung_L	0.99	0.99	0.99	555
Lung_R	1.0	1.0	1.0	563
Lungs	0.98	0.99	0.99	439
Non_OAR	1.0	0.99	1.0	8800
SpinalCord	0.97	0.98	0.98	689
accuracy	0.99	0.99	0.99	13044
macro avg	0.99	0.99	0.99	13044
weighted avg	0.99	0.99	0.99	13044

Table 35: VA-ROQS Lung dataset 10 fold Validation results.

REFERENCES

[1] AAPM. *TG263 AAPM. TG-263 Structure Spreadsheet.* `https://www.aapm.org/pubs/reports/RPT_263_Supplemental/TG263_Nomenclature_Worksheet_20170815.xls`.

[2] Joseph J Nalluri et al. "A smart healthcare portal for clinical decision making and precision medicine". In: *Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking*. 2018, pp. 1–6.

[3] Khajamoinuddin Syed et al. "Integrated Natural Language Processing and Machine Learning Models for Standardizing Radiotherapy Structure Names". In: *Healthcare*. Vol. 8. 2. Multidisciplinary Digital Publishing Institute. 2020, p. 120.

[4] W Sleeman et al. "Machine Learning Method to Automate Structure Name Mapping". In: *MEDICAL PHYSICS*. Vol. 46. 6. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA. 2019, E276–E276.

[5] William C. Sleeman IV et al. "A Machine Learning Method for Relabeling Arbitrary DICOM Structure Sets to TG-263 Defined Labels". In: Elsevier, 2020.

[6] Khajamoinuddin Syed et al. "Integrated Natural Language Processing and Machine Learning Models for Standardizing Radiotherapy Structure Names". In: *Healthcare - Under Review*. Multidisciplinary Digital Publishing Institute. 2020.

[7] K Syed et al. "Machine Learning Method to Automate Incident Triage in Radiotherapy Incident Reporting and Analysis System (RIRAS)". In: *MEDICAL PHYSICS*. Vol. 46. 6. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA. 2019, E258–E259.

[8] Khajamoinuddin Syed et al. "Treatment Practice Analysis of Intermediate or High Risk Localized Prostate Cancer: A Multi-center Study with Veterans Health Administration Data". In: *International Conference on Computational Advances in Bio and Medical Sciences*. Springer, Cham. 2019, pp. 134–146.

[9] Khajamoinuddin Syed et al. "Machine Learning Models for Predicting Intermediate or High-Risk Localized Prostate Cancer Treatment Plans across Multiple Veterans Health Administration Centers". In: *Journal of Computational Biology* (2020).

[10] Giorgio Ruffa. *Towards unification of organ labeling in radiation therapy using a machine learning approach based on 3D geometries*. 2019.

[11] NCBO BioPortal. *TG263 AAPM. TG-263 Structure Spreadsheet*. `http://bioportal.bioontology`.

[12] *American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology*. Mar. 2018. DOI: `10.1016/j.ijrobp.2017.12.013`.

[13] D Ruan W Shao J Wong D Veruttipong M Steinberg D Low P Kupelian. "Standardizing naming conventions in radiation oncology". In: *International Journal of Radiation Oncology* Biology* Physics* 83.4 (2012), pp. 1344–1349.

[14] JJ Kim et al. "A standardized nomenclature system for head and neck (H&N) IMRT contouring, planning and quality assurance". In: *International Journal of Radiation Oncology• Biology• Physics* 69.3 (2007), S473.

[15] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[16] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, pp. 144–152.

[17] Huw Prosser Evans et al. "Automated classification of primary care patient safety incident report content and severity using supervised machine learning (ML) approaches". In: *Health informatics journal* (2019), p. 1460458219833102.

[18] Thomas Cover and Peter Hart. "Nearest neighbor pattern classification". In: *IEEE transactions on information theory* 13.1 (1967), pp. 21–27.

[19] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[20] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[21] Jean L. Wright et al. "Standardizing Normal Tissue Contouring for Radiation Therapy Treatment Planning: An ASTRO Consensus Paper". In: *Practical Radiation Oncology* 9.2 (Mar. 2019), pp. 65–72. ISSN: 18798500. DOI: 10.1016/j.prro.2018.12.003.

[22] Stanley H Benedict et al. "Overview of the American Society for Radiation Oncology–National Institutes of Health–American Association of Physicists in Medicine Workshop 2015: Exploring opportunities for radiation oncology

in the era of big data". In: *International Journal of Radiation Oncology• Biology• Physics* 95.3 (2016), pp. 873–879.

[23] Thilo Schuler et al. "Big Data Readiness in Radiation Oncology: An Efficient Approach for Relabeling Radiation Therapy Structures With Their TG-263 Standard Name in Real-World Data Sets". In: *Advances in radiation oncology* 4.1 (2019), pp. 191–200.

[24] Tim Lustberg et al. "Radiation Oncology Terminology Linker: A Step Towards a Linked Data Knowledge Base." In: *MIE*. 2018, pp. 855–859.

[25] Desheng Ruan et al. "SU-F-T-102: Automatic Curation for a Scalable Registry Using Machine Learning". In: *Medical Physics* 43 (June 2016), pp. 3485–3485. DOI: 10.1118/1.4956238.

[26] D Rhee et al. "TG263-Net: A Deep Learning Model for Organs-At-Risk Nomenclature Standardization". In: *MEDICAL PHYSICS*. Vol. 46. 6. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA. 2019, E263–E263.

[27] Qiming Yang et al. "A Novel Deep Learning Framework for Standardizing the Label of OARs in CT". In: *Workshop on Artificial Intelligence in Radiation Therapy*. Springer. 2019, pp. 52–60.

[28] Michael Hagan et al. "VA-Radiation Oncology Quality Surveillance Program". In: *International Journal of Radiation Oncology* Biology* Physics* (2020).

[29] Austin McCartney, Svetlana Hensman, and Luca Longo. ""How short is a piece of string?" The Impact of Text Length and Text Augmentation on Short-text Classification Accuracy". In: (2017).

[30] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[31] Zellig S Harris. "Distributional structure". In: *Word* 10.2-3 (1954), pp. 146–162.

[32] Ronen Feldman, James Sanger, et al. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[33] Thomas Hill, Pawel Lewicki, and Paweł Lewicki. *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. StatSoft, Inc., 2006.

[34] Fabrizio Sebastiani. "Text categorization". In: *Encyclopedia of Database Technologies and Applications*. IGI Global, 2005, pp. 683–687.

[35] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[36] Jeffrey Pennington, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[37] Armand Joulin et al. "Bag of Tricks for Efficient Text Classification". In: *arXiv preprint arXiv:1607.01759* (2016).

[38] David G Kleinbaum et al. *Logistic regression*. Springer, 2002.

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

[40] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Vol. 26. Springer, 2013.

[41] Stuart Russell and Peter Norvig. "Artificial intelligence: a modern approach (global 3rd edition)". In: *Essex: Pearson* (2016).

[42] Xiang Zhang and Yann LeCun. "Text understanding from scratch". In: *arXiv preprint arXiv:1502.01710* (2015).

[43] Henry Lieberman. "How to color in a coloring book". In: *ACM SIGGRAPH Computer Graphics*. Vol. 12. 3. ACM. 1978, pp. 111–116.

[44] Mirek Fatyga, Baoshe Zhang, and William C Sleeman. "Designing and implementing a computing framework for image-guided radiation therapy research". In: *Computing in Science & Engineering* 14.4 (2011), pp. 57–68.

[45] Uwe Schneider, Eros Pedroni, and Antony Lomax. "The calibration of CT Hounsfield units for radiotherapy treatment planning". In: *Physics in Medicine & Biology* 41.1 (1996), p. 111.

[46] R. Bellman, Rand Corporation, and Karreman Mathematics Research Collection. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957. ISBN: 9780691079516.

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

[48] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. "Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions". In: *SIAM Review* 53.2 (2011), pp. 217–288.

[49] Benedick A Fraass. "Errors in radiotherapy: motivation for development of new radiotherapy quality assurance paradigms". In: *International Journal of Radiation Oncology* Biology* Physics* 71.1 (2008), S162–S165.

[50] Walt Bogdanich. "Radiation offers new cures, and ways to do harm". In: *The New York Times* 23 (2010).

[51] Paul Barach and Stephen D Small. "Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems". In: *Bmj* 320.7237 (2000), pp. 759–763.

[52] EC Ford et al. "Consensus recommendations for incident learning database structures in radiation oncology". In: *Medical physics* 39.12 (2012), pp. 7272–7290.

[53] *Patient Safety Rule*. `https://www.pso.ahrq.gov/legislation/rule`. 2008.

[54] Ewoud Pons et al. "Natural language processing in radiology: a systematic review". In: *Radiology* 279.2 (2016), pp. 329–343.

[55] Stéphane M Meystre et al. "Extracting information from textual documents in the electronic health record: a review of recent research". In: *Yearbook of medical informatics* 17.01 (2008), pp. 128–144.

[56] Andrés Felipe Torres Cano. "Automatic aviation safety reports classification". MA thesis. University of Twente, 2019.

[57] Kosio Marev and Krasin Georgiev. "Automated Aviation Occurrences Categorization". In: *2019 International Conference on Military Technologies (ICMT)*. IEEE. 2019, pp. 1–5.

[58] C Morais, K Yung, and E Patelli. "Machine-learning tool for human factors evaluation-application to lion air Boeing 737-8 max accident". In: (2019).

[59] Saul D Robinson. "Multi-label classification of contributing causal factors in self-reported safety narratives". In: *Safety* 4.3 (2018), p. 30.

[60] Christophe Pimm et al. "Natural Language Processing (NLP) tools for the analysis of incident and accident reports". In: 2012.

[61] Samuel Danso, Eric Atwell, and Owen Johnson. "A comparative study of machine learning methods for verbal autopsy text classification". In: *arXiv preprint arXiv:1402.4380* (2014).

[62] Zoie Shui-Yee Wong and Masanori Akiyama. "Statistical text classifier to detect specific type of medical incidents." In: *Medinfo*. 2013, p. 1053.

[63] Mei-Sing Ong, Farah Magrabi, and Enrico Coiera. "Automated categorisation of clinical incident reports using statistical text classification". In: *Quality and Safety in Health Care* 19.6 (2010), e55–e55.

[64] Charitini Stavropoulou, Carole Doherty, and Paul Tosey. "How effective are incident-reporting systems for improving patient safety? A systematic literature review". In: *The Milbank Quarterly* 93.4 (2015), pp. 826–866.

[65] Brenda G Clark et al. "The management of radiation treatment error through incident learning". In: *Radiotherapy and Oncology* 95.3 (2010), pp. 344–349.

[66] Edward Loper and Steven Bird. "NLTK: the natural language toolkit". In: *arXiv preprint cs/0205028* (2002).

[67] Amir A Kimia et al. "An introduction to natural language processing: how you can get more from those electronic notes you are generating". In: *Pediatric emergency care* 31.7 (2015), pp. 536–541.

[68] Anita Alicante et al. "A study on textual features for medical records classification". In: *Innovation in Medicine and Healthcare 2014* 207 (2015), p. 370.

[69] Guergana K Savova et al. "Mayo clinic NLP system for patient smoking status identification". In: *Journal of the American Medical Informatics Association* 15.1 (2008), pp. 25–28.

[70] Erdem Alparslan, Adem Karahoca, and Hayretdin Bahşi. "Classification of confidential documents by using adaptive neurofuzzy inference systems". In: *Procedia Computer Science* 3 (2011), pp. 1412–1417.

[71] Ashwin Ittoo, Antal van den Bosch, et al. "Text analytics in industry: Challenges, desiderata and trends". In: *Computers in Industry* 78 (2016), pp. 96–107.

[72] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

[73] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[74] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[75] Jeremy Howard and Sebastian Ruder. "Universal language model fine-tuning for text classification". In: *arXiv preprint arXiv:1801.06146* (2018).

[76] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. "Regularizing and optimizing LSTM language models". In: *arXiv preprint arXiv:1708.02182* (2017).

[77] Ludovic Tanguy et al. "Natural language processing for aviation safety reports: from classification to interactive analysis". In: *Computers in Industry* 78 (2016), pp. 80–95.

[78] Ying Wang et al. "Using multiclass classification to automate the identification of patient safety incident reports by type and severity". In: *BMC medical informatics and decision making* 17.1 (2017), p. 84.

[79] Rebecca L Siegel. "colleagues. Cancer statistics, 2018. 2018; 68: 7-30". In: *CA Cancer J Clin* 68 (2018), pp. 7–30.

[80] Carvell T Nguyen and Michael W Kattan. "Nomograms in Prostate Cancer". In: *Prostate Cancer: A Comprehensive Perspective*. Springer, 2013, pp. 581–592.

[81] H Gilbert Welch and Peter C Albertsen. "Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986–2005". In: *Journal of the National Cancer Institute* 101.19 (2009), pp. 1325–1329.

[82] Phillip L Ross et al. "Comparisons of nomograms and urologists' predictions in prostate cancer." In: *Seminars in urologic oncology*. Vol. 20. 2. 2002, pp. 82–88.

[83] Anthony V Dámico et al. "Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer". In: *Jama* 280.11 (1998), pp. 969–974.

[84] Michael W Kattan et al. "A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer". In: *JNCI: Journal of the National Cancer Institute* 90.10 (1998), pp. 766–771.

[85] Michael W Kattan, Thomas M Wheeler, and Peter T Scardino. "Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer". In: *Journal of Clinical Oncology* 17.5 (1999), pp. 1499–1499.

[86] Anthony V D'amico et al. "The combination of preoperative prostate specific antigen and postoperative pathological findings to predict prostate specific antigen outcome in clinically localized prostate cancer". In: *The Journal of urology* 160.6 (1998), pp. 2096–2101.

[87] Anthony V D'Amico et al. "Pretreatment nomogram for prostate-specific antigen recurrence after radical prostatectomy or external-beam radiation therapy for clinically localized prostate cancer". In: *Journal of Clinical Oncology* 17.1 (1999), pp. 168–168.

[88] Anthony V D'Amico et al. "Combination of the preoperative PSA level, biopsy gleason score, percentage of positive biopsies, and MRI T-stage to predict early PSA failure in men with clinically localized prostate cancer". In: *Urology* 55.4 (2000), pp. 572–577.

[89] Peter B Snow, Deborah S Smith, and William J Catalona. "Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study". In: *The Journal of urology* 152.5 (1994), pp. 1923–1926.

[90] Stefan Conrad et al. "Prospective validation of an algorithm with systematic sextant biopsy to predict pelvic lymph node metastasis in patients with clinically localized prostatic carcinoma". In: *The Journal of urology* 167.2 (2002), pp. 521–525.

[91] Markus Graefen et al. "A validated strategy for side specific prediction of organ confined prostate cancer: a tool to select for nerve sparing radical prostatectomy". In: *The Journal of urology* 165.3 (2001), pp. 857–863.

[92] Anthony V D'Amico et al. "Clinical utility of the percentage of positive prostate biopsies in predicting prostate cancer-specific and overall survival after radiotherapy for patients with localized prostate cancer". In: *Interna-*

tional Journal of Radiation Oncology Biology Physics 53.3 (2002), pp. 581–587.

[93] Erica S Spatz, Harlan M Krumholz, and Benjamin W Moulton. "Prime time for shared decision making". In: *Jama* 317.13 (2017), pp. 1309–1310.

[94] William F Athas et al. "Travel distance to radiation therapy and receipt of radiotherapy following breast-conserving surgery". In: *Journal of the National Cancer Institute* 92.3 (2000), pp. 269–271.

[95] JS Blumenthal-Barby et al. "The neglected topic: presentation of cost information in patient decision AIDS". In: *Medical Decision Making* 35.4 (2015), pp. 412–418.

[96] Cheryl K Altice et al. "Financial hardships experienced by cancer survivors: a systematic review". In: *JNCI: Journal of the National Cancer Institute* 109.2 (2017).

[97] NCCN. *NCCN Guidlines for Prostate Cancer 2016*. `https://www.cancer.gov/about-cancer/understanding/statistics`.

[98] Lindsay M Burt, Dennis C Shrieve, and Jonathan D Tward. "Factors influencing prostate cancer patterns of care: An analysis of treatment variation using the SEER database". In: *Advances in radiation oncology* 3.2 (2018), p. 170.

[99] LA Robles et al. "Factors influencing patients' treatment selection for localised prostate cancer: A systematic review". In: *British Journal of Medical and Surgical Urology* 5.5 (2012), pp. 207–215.

[100] Ronald C Chen et al. *Receipt of guideline-concordant treatment in elderly African American and Caucasian patients with prostate cancer.* 2012.

[101] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[102] Daniel Maturana and Sebastian Scherer. "Voxnet: A 3d convolutional neural network for real-time object recognition". In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2015, pp. 922–928.

[103] Matthew E Peters et al. "Deep contextualized word representations". In: *arXiv preprint arXiv:1802.05365* (2018).

[104] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI Blog* 1.8 (2019).

[105] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[106] Emily Alsentzer et al. "Publicly available clinical BERT embeddings". In: *arXiv preprint arXiv:1904.03323* (2019).

[107] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

VITA

Khajamoinuddin Syed received his MS in computer science degree from State University of New York at Albany and BS information science degree from Visvesvaraya Technological University in 2014 and 2009, respectively. He worked for Cognizant Global Services from 2009 to 2013 as a Programmer Analyst in Bangalore, India. His research interests include Bio-inspired computational techniques, genomics, health informatics, machine learning, and natural language processing in the biomedical domain.