UNIVERSITY OF LEEDS

Machine learning for outcome prediction after pelvic radiotherapy

Behnaz Elhaminia

University of Leeds

School of Computing

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

June 2023

This thesis is dedicated to:

all brave-heart patients fighting against cancer

Intellectual Property Statement

The candidate confirms that the work submitted is her own except where work which has formed part of jointly authored publications has been included. The candidate confirms that appropriate credit has been given where reference has been made to the work of others. The contribution of the candidate and the other authors to this work has been explicitly indicated as follows.

I am the main author of all the following publications. I led the design of the study and wrote the manuscript, including the analysis and discussion of the results. I also wrote the analysis code and carried out the data processing, statistical analysis, and experiments. The other authors' contribution was in helping me designing the study, providing the data annotations, discussion of the results, and review of the manuscript. They also took full responsibility for the design and data collection of the original clinical study, thereby providing all the raw data for the research.

- Chapter 5:

    1) B. Elhaminia, et al., "Toxicity Prediction in Pelvic Radiotherapy Using Multiple Instance Learning and Cascaded Attention Layers", IEEE Journal of Biomedical and Health Informatics, 27.4 (2023), pp. 1958–196.

- Chapter 6:

    1) B. Elhaminia, et al., "Deep learning combining imaging, dose and clinical data for predicting bowel toxicity after pelvic radiotherapy", Medical Physics, under review.

    2) B. Elhaminia, et al., "Deep learning with visual explanation for radiotherapy-induced toxicity prediction", SPIE Medical Imaging 2023, P. 124651V.

I am also the joint-first author of the following publication. This paper is a collaborative effort; I conducted research for specific sections of the paper, while Dr. Ane Appelt designed the study, wrote the paper, and conducted research for the other sections. The other authors provided comments and feedback on the discussion of the results, and reviewed the manuscript.

- Chapter 2:

    1) A. Appelt, B. Elhaminia, et al., "Deep learning for radiotherapy outcome prediction using dose data–a review", Clinical Oncology 34.2 (2022), e87–e96

Signed 

Acknowledgements

I have always dreamed of being a scientist, and now I feel like I am almost there! I wanted to take a moment to thank all those individuals who have helped me along the way.

First, I want to thank my main supervisor, Dr Ali Gooya. He provided me with invaluable technical advice and guidance, and he gave me the freedom to explore my research interests in my own way. I really appreciate his expertise and supervision. I also want to thank my second supervisor, Dr Ane Appelt. She was incredibly supportive, both technically and non-technically. She played a pivotal role in enhancing multiple skills, such as presentation and writing, and consistently provided unwavering support. She was an inspiration for me, and she helped me learn more about the academic world. I am immensely grateful for her remarkable mentorship throughout every stage of my PhD. I am also very grateful to Dr Alexandra Gilbert, another one of my supervisors. She provided me with highly valuable and helpful technical support and guidance on the dataset, as well as in preparing my manuscripts and documents. Finally, I want to thank my other supervisor Professor Andy Scarbrook, for his amazing support, guidance and supervision. He was an inspiration to me of how to be a cool professional scientist.

I am so grateful for the feedback and comments that all of my supervisors provided on this thesis. Their insights helped me to develop a more scientific view for preparing this document. I was so lucky to have their support on my journey.

Besides my advisors, I would like to thank the School of Computing for their scholarship, which helped me to complete my research. I'd also like to thank the RADNET group in Leeds for their support. I also want to thank my friends in the office for all the fun times we had together at conferences, parties, nights out and in the lab.

Last but not least, I want to thank my Johny♡, for always being there for me. I could not have gotten through the scary days of the COVID-19 quarantine without his endless love and support. I am so grateful for his wise advice and his willingness to listen to me whenever I was struggling with my PhD. I am also very thankful to my family and friends in Iran, who always made me feel loved and supported. To my sister, for our daily random funny/political/scientific discussions on the phone that made me forget that I live so far away from my family. Finally, I want to express my deepest gratitude to my parents. Thank you for believing in me, and for your endless support and encouragement.

Abstract

Radiation therapy plays a crucial role in the treatment of cancer and is widely used in curative and palliative care. While this therapy effectively targets tumour tissue, it can inadvertently harm cells in nearby organs, resulting in toxicity. These toxicity effects significantly impact patients' well-being, causing physical and mental challenges. In pelvic radiotherapy, the majority of patients may develop three common toxicities related to irradiation of the bowel including diarrhoea, faecal incontinence, and bowel urgency.

The severity of toxicity depends on various factors, including cellular characteristics, radiation dosage to non-cancerous tissues, patient attributes, and oncologic treatment variables. However, the relationship between these factors and the risk of toxicity remains unclear. To minimise the impact of toxicity, it is crucial to accurately assess potential risks during treatment planning.

This research delves into the application of machine learning, particularly deep learning methods, to identify the correlation between late toxicity and various features after radiotherapy. Specifically, the research aims to establish a framework for predicting radiotherapy-induced toxicity, as well as detecting and analysing potential risk factors for patients with pelvic cancers.

The main framework of this research includes two convolutional blocks for analysis of computerized tomography (CT) scans and dose distribution data, and a fully-connected path for analysis of clinical variables, including demographics, comorbidities, medications, and treatment features. An attention mechanism was employed to determine possible risk factors and critical anatomical regions. In the study, a dataset of 315 patients treated at Leeds Cancer Centre in the United Kingdom between 2009 and 2014 was utilized.

The summary of the results indicates that the attention weights for bowel urgency were primarily concentrated on the right iliac fossa, and the attention weights for faecal incontinence were focused on the postero-inferior region (i.e., corresponding to the anorectum). However, no specific anatomical region could be identified from the attention weights for predicting diarrhoea. The analysis of clinical data, in conjunction with CT and dose, led to an improvement in prediction performance, resulting in an area under the receiver operating characteristic curve (AUC) of 88% for bowel urgency and 78% for faecal incontinence. In contrast, the best performance for predicting diarrhoea was achieved when analyzing clinical features alone, resulting in a 68% AUC. The proposed frameworks and the outcomes of this study can assist clinicians in gaining a better understanding of toxicity and its intricate relationship with different factors.

List of Abbreviations

AE	Autoencoder
AI	Artificial Intelligence
ANN	Artificial Neural Network
BMI	Body Mass Index
CNN	Convolutional Neural Network
CRT	Conformal Radiotherapy
CT	Computed Tomography
DVH	Dose Volume Histograms
EBRT	External Beam Radiotherapy
EORTC	European Organisation for Research and Treatment of Cancer
FcNN	Fully Connected Neural Network
FPR	False Positive Rate
GAN	Generative Adversarial Neural Network
GI	Gastrointestinal
GPU	Graphical Processing Units
Grad-CAM	Gradient-weighted Class Activation Mapping
IMRT	Intensity Modulated Radiation Therapy
KNN	K-nearest Neighbour
LR	Logistic Regression
MIL	Multiple Instance Learning
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
NTCP	Normal Tissue Complication Probability
OAR	Organ At Risk
PCA	Principal Component Analysis
PET	Positron Emission Tomography
QOL	Quality Of Life
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
RF	Random Forest
ROC	Receiver Operating Characteristics
RT	Radiation Therapy
RTOG	Radiation Therapy Oncology Group
SBRT	Stereotactic Body Radiotherapy
SMOTE	Synthetic Minority Oversampling Technique
STD	Standard Deviation
SVM	Support Vector Machine
TCP	Tumour Control Probability
TPR	True Positive Rate
VMAT	Volumetric Modulated Arc Therapy

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The objective of this chapter is to provide a succinct overview of the issue of radiotherapy-associated toxicity and the ways in which current techniques assist in reducing this. The contribution of this doctoral studies to this area and the overall structure of the thesis are also included.

1.1 Problem definition

Radiation therapy (RT) is currently utilised in over 50% of patients with cancer, either as a curative treatment or for palliative care [35]. It can be considered as the main treatment or in combination with chemotherapy, surgery and immunotherapy to manage cancerous tissues and it is sometimes used to treat non-cancerous (benign) tumours and other diseases, such as thyroid eye disease and blood disorders [58]. Broadly, there are two different methods for RT: external beam radiotherapy (irradiation from outside the body) and internal radiotherapy (irradiation from inside the body). The former involves focusing high-energy radiation beams typically using a linear accelerator into the tumour area, while the latter can be performed by injecting/swallowing radioactive liquids or implanting radioactive metal inside the body near

the cancerous tissue either temporarily (brachytherapy) or permanently (seed brachytherapy). Depending on the type, stage, and location of the tumour, the number (or fraction) of treatments administered varies. The use of concurrent chemotherapy or other concurrent agent (used as a radiosensitizer) again depends on the type of cancer and staging and patient fitness.

Using ionising radiation to eliminate tumour tissue can also affect normal tissues in surrounding organs, referred to as "organs at risk" (OAR), potentially resulting in damage to that tissue, leading to toxicity; RT generates highly reactive molecules within the cellular structures to disrupt DNA strands and other vital cellular components, leads to irreversible cellular damage and death to healthy cells.

The majority of patients report various side effects after RT, which affect their well-being and quality of life [35]. Side effects are related to the anatomical position of RT and the adjacent tissues. Hair loss, nausea, sore mouth, diarrhoea, sore or red skin, mucosal injury, and moderate fatigue are all common acute side effects of radiation therapy ([35], [35]).

These toxicities may be associated with short-term and long-term consequences. In short-term (acute) toxicity, adverse effects appear during radiotherapy or within three months after the treatment, and they may resolve within months. Long-term (late) toxicities are observed later than three months, and are usually considered progressive and irreversible over time and affect the patient's quality of life.

The damage caused by RT to normal cells depends on various factors, including cellular characteristics, organs' physiology and anatomy, radiation dose to normal tissues, patient characteristics, and oncological treatment features. However, the relationship between these factors and the risk of late toxicity for many organs remains unclear. To reduce the damage to normal tissues and minimise toxicity risk, RT is precisely planned for each patient. Clinicians must have a clear understanding of the relationship between different risk factors for late toxicity affecting each patient in order to develop the most efficient treatment plan. As a result, an accurate assessment of potential toxicity risks is a crucial component of radiotherapy treatment

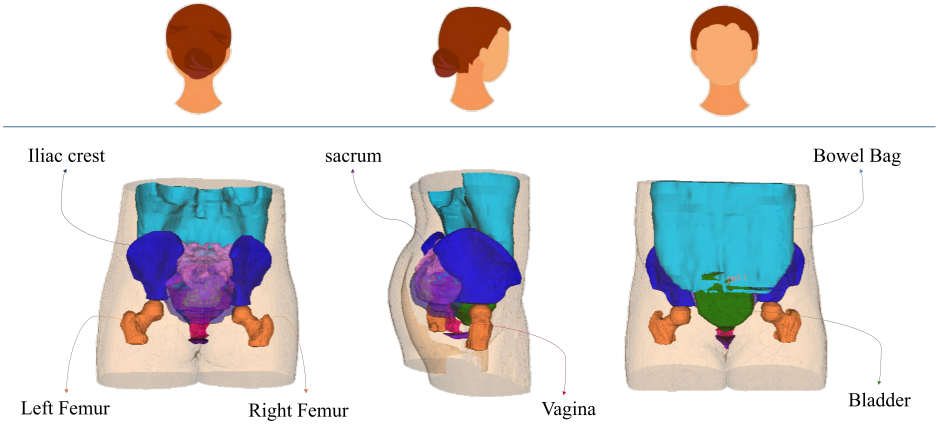


Figure 1.1: Example of a female pelvis in three different views. The "peritoneal cavity" structure is considered as the bowel bag (organ at risk) for this thesis.

planning.

1.1.1 Long-term toxicity from pelvic radiation therapy

The pelvis is the lower part of the torso consisting of the intestines, urinary bladder and reproductive organs. Fig. 1.1 illustrates important anatomical structures in the female pelvis which can be affected by RT. Bilateral femurs (hip bone) are connected in front, at the pubic symphysis, along with iliac bones at the side and the sacrum and coccyx behind form the major pelvic bones, and provide support and balance for the torso. Throughout the entire thesis, the "peritoneal cavity" structure is identified as the "bowel bag", which is the organ at risk being studied for potential toxicities.

Pelvic cancers develop in organs and structures such as the reproductive organs(uterus/cervix in females or prostate in males), the bladder, rectum, and the anus. A variety of pelvic cancers, including anal cancer, rectal cancer and bladder cancer, affect both men and women. Other cancers are dependent on patient sex; cervical, uterine/ endometrial, ovarian, vaginal, and vulval cancers being female-specific, whilst prostate and testicular cancer are specific to male. The most prevalent side effect or toxicity after pelvic RT are gastrointestinal (GI) disorders.

It is estimated that 50% of patients treated with RT develop GI toxicities over a period of 10 years after their treatment [3], [32]. Diarrhoea, faecal incontinence, bleeding, bowel urgency, abdominal discomfort, cramping, mucous in the stool, and tenesmus are the most common. The severity of these side effects depends on the irradiated area and treatment regimen, but majority of the patients reported that it decreased the quality of their life([32]).

1.2 Reducing radiotherapy toxicity

In modern radiotherapy techniques, three-dimensional (3D) images of the body, usually CT scans, are utilised for the purpose of planning and optimising individual patient treatment plans. This process involves creating organ segmentation (structure sets) and ultimately generates a 3D dose distribution or matrix, which accurately represents the radiation dose delivered to the specific anatomical structures involved. Efforts to minimise adverse effects of RT have mirrored advances in radiation physics and treatment regimens. Conformal radiotherapy (CRT) and intensity-modulated radiation therapy (IMRT) are techniques where radiation beams are tightly shaped to closely fit the area of the cancer and avoid healthy tissues as much as possible.

CRT involves shaping radiation beams to match the shape of the tumour as closely as possible. This is achieved by using multiple radiation beams and customised blocking devices to shape the radiation field. The goal of CRT is to deliver a higher dose to the tumour while minimising radiation exposure to surrounding healthy tissues. On the other hand, IMRT takes the concept of conformal radiotherapy a step further. IMRT uses advanced computer algorithms to divide the radiation beam into many smaller beamlets, each with individually adjustable intensity. This allows for more precise modulation of radiation dose within the tumour, enabling higher doses to be delivered to specific regions while sparing nearby healthy tissues. IMRT offers greater flexibility in delivering varying radiation doses to different areas within the tumour.

Although CRT and IMRT reduce the risk of toxicity (by exposing less normal tissues to radiation), the higher doses used still affect non-target cells around tumours which can lead to

toxicity.

To address this issue, clinicians personalise treatment plans for each patient based on assumptions about the relationship between radiation dose, tumour control, and potential treatment toxicity. These relationships have been explored with traditional outcome modelling techniques, which rely on volume segmentation information and 3D dose information to create treatment dose volume histograms (DVH) for different organs at risk. In DVH-based models, the radiation dose is simplified into a one-dimensional form, for instance mean absorbed dose, for each anatomical structure. Then this dose representation is related to outcome, e.g., through generalised linear modelling, potentially taking patient characteristics or other clinical factors into account.

However, collapsing a three-dimensional dose distribution into a one-dimensional dose vector, limits the modelling power and may result in a low accuracy prediction. In particular, spatial dose distribution information and heterogeneous radio-sensitivity of organs can improve the prediction power, which are not considered by current outcome modelling methods. Details about analysing dose spatial information and its impact on the prediction performance are discussed in the following chapters.

1.3 Treatment improvement with artificial intelligence

The main disadvantage of DVH-based methods is that they do not consider spatial information within the dose treatment plan. Furthermore, coping with unstructured visual information such as 3D CT scans or magnetic resonance imaging (MRI) data is challenging, and manual methods of processing and analyzing such quantities of data is labour intensive. On the other hand, it has been demonstrated that deep learning models can reliably perform image analysis tasks such as classification, segmentation, and registration [101] due to their ability to efficiently identify spatial patterns. This has inspired researchers to apply machine learning techniques to medical image analysis tasks.

In recent years, medical image analysis has been leveraging artificial intelligence (AI) and particularly deep learning techniques [74], [132]. AI as a broad field encompassing the development of intelligent machines that can perform tasks that typically require human intelligence, wide range of techniques and approaches, such as machine learning (ML), computer vision, and robotics. Deep learning (DL) is a subfield of ML that utilizes artificial neural networks (ANNs) to learn from data. Using ML and deep learning models to analyse medical data has resulted in image processing models with higher accuracy and lower time complexity. Furthermore, deep networks can help uncover hidden patterns and recognise complex structures of medical data autonomously within a couple of minutes, which is humanly impracticable to perform. Automatic dose treatment planning for head and neck cancer [50], [115], prediction and prognostication of treatment outcome [71], image registration for RT [147] and RT quality assurance [80] are some examples of recently developed ML methods for clinical tasks in radiation oncology.

Despite the numerous potential advantages offered by AI in the medical field, there are still several challenges; Firstly, AI technology is relatively new and consequently inaccuracies are possible. Due to the critical nature of medical tasks, there is always a need for human surveillance. Furthermore, the implementation of AI in clinical practice encounters licensing challenges, necessitating a validation process for AI systems to be utilised as medical devices; to ensure patient safety and regulatory compliance, a thorough validation process is required. This process involves rigorous evaluation, testing, and certification of AI systems as medical devices. Moreover, processing dimensionally-huge data requires powerful resources (computers with fast graphical processing units (GPUs) and large memories) that can be expensive. Lack of adequate datasets and being susceptible to security risks are other challenges AI faces in medical tasks.

1.4 Thesis contribution and overview

Although modern treatments for cancer have improved survival rates, and state-of-the-art radiotherapy techniques offer much less toxic treatment than even just a few decades ago, radiation-induced late toxicity still affects a significant number of people. Adverse effects of cancer treatment often impact patients' quality of life and cause physical and psychological problems for them. Knowledge of the treatment outcome leads to a more precise approach for controlling RT toxicity and an improvement in symptom control. Reducing long-term toxicities has emerged as a key component of radiotherapy treatment optimisation, and this needs to be addressed in oncology research and practice.

There are several studies focused on the outcome prediction with the help of ML-based methods to handle dose data (the related works and their challenges are thoroughly reviewed in chapter 2). What is less clear and is a major limitation in many studies is that the response (output) of the model is not fully understood; the complexity of ML models makes it challenging to identify risk factors and explain their behaviour, leading to a significant amount of uncertainty regarding the relationship between input data and the classification outcome.

The focus of this research is to study the use of ML, and in particular deep learning methods, with the aim of exploring the correlation between late toxicity and various features after pelvic radiotherapy. A specific objective is to provide a framework for predicting bowel-related toxicities (as the most common toxicity following pelvic radiotherapy), including urgency, diarrhoea, and faecal incontinence, and identifying possible risk factors in patients with pelvic cancers. The presented framework is a novel approach that leverages multiple instance learning and an attention mechanism to identify correlations between input data and bowel-related toxicities, simultaneously assessing the impact of each data on the toxicity. This thesis analysed data collected from a research project (Ethics Committee approval code: 13-YH-0156; project name: improving assessment and recording of pelvic cancer treatment effects; funder: National Institute for Health and Care Research; start date: 01/04/2014; end date: 19/03/2015)

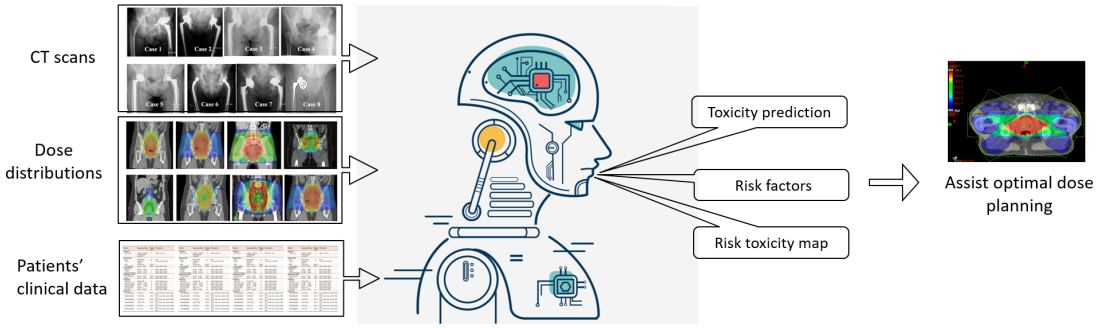


Figure 1.2: Overview of the study. Developing a machine learning framework to analyse CT scans, dose distribution, and clinical data, with the aim of predicting toxicity after pelvic radiotherapy; and detecting correlations between toxicity and various factors.

which included patient reported data (from validated questionnaires) on bowel toxicity, alongside clinical and treatment data, radiotherapy dose and imaging data from patients treated with curative intent RT for four pelvic malignancies (rectal, anal, cervical and endometrial cancer) at Leeds Cancer Centre. Fig.1.2 shows the overall structure of this work. this thesis studies sought to address these outcome prediction tasks through the following objectives:

1. Analysis of patient's numerical data including patient's demographic, treatment features and pre-treatment medical information with machine learning to:
   - predict the occurrence of patient-reported bowel-related toxicities.
   - analyse and quantify the impact of each numerical feature on the predicted toxicity.
2. A fully-automated workflow using patient's 3D image data including CT scans and dose distributions to:
   - predict the occurrence of bowel-related toxicities.
   - provide visual explanations for the predicted outcome in order to identify which anatomical structures are involved in the toxicity.
   - quantitatively analyse the importance of CT and dose data and their associations with toxicity.

3. A fully-automated workflow exploring a combination of image and numerical data to:
    - predict the occurrence of bowel-related toxicities.
    - quantify the correlation between inputs and the predicted outcome.
4. Construct an atlas to summarise, visualise and localise the radiotherapy-induced toxicity based on the anatomical structure of the pelvis.

The aim of these studies is to improve the prediction of toxicity using novel prediction methods, which in the future could be used to develop a clinical decision support tool for radiotherapy planning, not only improving patient outcomes but also potentially providing clinicians with time, cost, and efficiency savings.

The overall structure of this thesis takes the form of six chapters, where:

- **Chapter 2** provides a review of recent studies which used machine learning and deep learning models to predict RT-induced toxicity. A modified version of the review originates from my paper "Deep learning for radiotherapy outcome prediction using dose data–a review" published in Clinical Oncology [6].
- **Chapter 3** provides information on the datasets employed in this research; two datasets of image and numerical data with their pre-processing procedures are described in detail. Then, a data analysis plan presenting a roadmap for analysis, interpretation, and organisation of data in the study is provided.
- **Chapter 4** reports the findings of an outcome prediction study focusing on numerical data. In this chapter, employing three different machine learning methods, toxicity was analysed with respect to patients' clinical data. The quantitative assessment of possible risk factors is provided as the outcome of this chapter.
- **Chapter 5** presents a novel deep learning model for toxicity prediction based on multiple instance learning and attention mechanism. The output of the model addresses three

issues: toxicity classification, toxicity risk map and input associations. A toxicity atlas summarising risk maps based on bowel bag structure is also presented as a separate result of this chapter. A shortened version of this chapter originates from my paper "Toxicity Prediction in Pelvic Radiotherapy Using Multiple Instance Learning and Cascaded Attention Layers" published in IEEE Journal of Biomedical and Health Informatics [47].

- **Chapter 6** reports the findings of an outcome prediction study utilising both imaging and numerical data and their associations with RT-induced toxicity. The proposed method employs both types of data in order to analyse and predict toxicity after pelvic radiotherapy. A shortened version of this chapter is submitted to Medical Physics Journal and a section of the analysis from this chapter was published as a scientific abstract entitled "Deep learning with visual explanation for radiotherapy-induced toxicity prediction", presented at the SPIE Medical Imaging, Computer-Aided Diagnosis conference held in San Diego, California, United States in February 2023 [46]

- **Chapter 7** provides an executive summary of the findings of this research, and reflects on the implementation of the research in practice, discussion of limitations and future work.

1.5 Summary and conclusions

This initial chapter briefly introduced the issue of radiotherapy-induced toxicity and explored current methods aimed at mitigating it. Furthermore, an overview was provided on how machine learning can potentially help address these problems. In order to comprehend why machine learning has emerged in this field and identify the specific shortcomings of traditional models that have prompted the adoption of machine learning, it is necessary to review the advantages and disadvantages of existing models. In the next chapter, a more in-depth discussion of the limitations of current models and a comprehensive summary of recent works will be provided. The subsequent chapter will also define the technical background to this work.

Chapter 2

Literature Review and Technical Background

This chapter offers a comprehensive overview of the current models used for radiotherapy outcome prediction. It delves into the details of various machine learning and deep learning models proposed for this purpose. Additionally, the technical background of these reviewed models is thoroughly explained. By exploring the state of research in this field, this chapter provides insights into the current progress, the challenges faced, and the strategies employed to address them.

A modified version of the review originates from my paper "Deep learning for radiotherapy outcome prediction using dose data–a review" published in Clinical Oncology journal [6].

2.1 Introduction

Recent improvements in personalised radiotherapy have involved the use of predictive models to optimise the treatment plan individually for patients [108], [2], [85], [160]. Traditional outcome modelling in radiation oncology uses statistical models to explore the correlation between

input data and toxicity. These statistical models are typically in the form of generalised linear modelling and rely on one-dimensional (1D) input data. Consequently, there is data dimension reduction as a pre-processing step. For instance, in the vast majority of prediction methods, the 3D dose treatment plan is reduced to a one-dimensional dose vector, where each element represents the mean absorbed dose or volume receiving a certain dose for a specific anatomical structure. Representing highly complex and multidimensional RT data with one-dimensional vectors leads to missing correlations and discards spatial information. Sensitivity to outliers and data independency are other challenges that affect traditional prediction approaches.

Methodologies using machine learning have emerged to overcome the shortcomings of classic radiotherapy outcome prediction models. They have been utilised to explore the complex relationships between input and output data. In this chapter, the studies that leverage machine learning models for RT outcome prediction are reviewed. Prior to that, the technical aspects underlying the ML models employed in these reviewed studies are explained.

2.2 Technical background

2.2.1 Machine learning for classification

Machine learning is a field of study concerned with building methods and algorithms that learn from examples. ML methods leverage these examples, known as "training data," to learn to perform some set of tasks, and they are applied to new examples, known as "test data," for their performance evaluation. Classification is a task where the model learns how to assign "labels" to samples from the input data in order to distinguish their categories. Various types of machine learning models have been researched and developed for classification problems; artificial neural networks, decision trees, regression analysis, K-nearest neighbour, Bayesian networks and support vector machines are amongst well-known methods. Different factors can affect choosing one model over the others; including the nature of the problem, size and quality of the data, type of the desired output, urgency of the task, available computational resources,

etc.

Typically, traditional machine learning techniques require a domain expert to identify and extract features in order to simplify the data and increase the visibility of patterns for the learning algorithms to work effectively. However, deep learning algorithms have a significant advantage in that they are capable of learning high-level features directly from the data. As a result, the need for domain expertise and complex feature extraction is significantly reduced. Leveraging the powerful ability to discover inherent patterns in 3D images, the usage of artificial neural networks to address the problem of predicting radiotherapy-related toxicity was explored.

As previously discussed, traditional outcome modelling in radiation oncology typically analyses 1D input data with generalised linear modelling. In order to evaluate the performance of deep models that take spatial information into account compared to methods without spatial information, clinical data using three machine learning models including support vector machine, logistic regression, and random forest was analysed. Each of these models was selected for specific reasons, which are explained in detail in the following sections.

Logistic regression

Logistic regression (LR) is a generalised linear model for classification that computes the posterior probability of class C_1 with a logistic Sigmoid function σ on a linear combination of feature vector ϕ ([17]) as:

$$p(C_1|\phi) = y(\phi) = \sigma(\mathbf{w}^T\phi) \tag{2.1}$$

where $\mathbf{w}$ are parameters of the model and determined with maximum likelihood during the model training. For a dataset $\{\phi_n, t_n\}$, where $\phi_n = \phi(X_n)$ and $t_n \in \{0, 1\}$, with $n = 1, ..., N$, the likelihood can be noted

$$p(\boldsymbol{t}|\mathbf{w}) = \prod_{n=1}^{N} {y_n}^{t_n}\{1 - y_n\}^{1-t_n}, \tag{2.2}$$

where y_n is the output of the LR model (the probability of belonging to class C_1 where the input is x_n) and $\boldsymbol{t}$ is a vector of all the labels ($\boldsymbol{t} = (t_1, ..., t_n)^T$). The negative logarithm of the likelihood function is the error function -cross entropy function- and it is minimised by the proposed model. The minimisation can be performed during the training with respect to the gradients of the error function.

A considerable amount of literature in radiotherapy outcome prediction has used LR-based models (see e.g. ([75], [99]), [112]). This may be due to a number of reasons: First, logistic regression is relatively easy to implement and interpret. Second, for a M_dimensional feature space, the model needs to adjust M parameters, which means the complexity of the model is linear and dependent on M. In problems with large feature space, such as medical data analysis, LR models are quite efficient. Third, it is easy to infer the importance of the feature as it computes a direct association between features and output; in the toxicity prediction issue, understanding the clinical features that impact the outcome of treatment is a crucial aspect. Furthermore, due to the simplicity of the algorithm, the training time can be less than for other complex algorithms, and it can be extended for multi-class classification.

Logistic regression has a linear decision surface that requires linearly separable input data. Therefore, non-linear problems are very challenging to solve. As the relationship between clinical features and toxicity is not necessarily linear, traditional LR may not be the most appropriate model for the data. There are some approaches, such as data transformation, that can help with non-linear problems, but they increase the learning complexity.

Support vector machines

Support vector machine (SVM) is a ML algorithm that aims to find a hyperplane that separates different classes with the maximal margin. In two-dimensional space, the hyperplane is a line that splits the plane into two sections, each attributed to one class (See Fig.2.1). Support vectors are the data points located near the hyperplane, and they influence the location and orientation

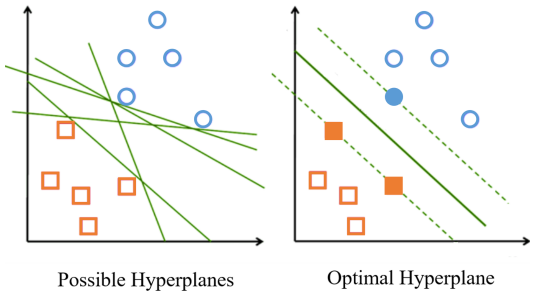


Figure 2.1: SVM is trained to distinguish two different classes by finding the hyperplane with the maximum margin of support vectors. Support vectors are shown with filled-in shapes.

of the final hyperplane. Considering these support vectors, the algorithm finds the optimal hyperplane with the largest margin from the support vectors. With a training dataset containing N input vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$, corresponding labels values $y_1, ..., y_N$ where $y_n \in \{-1, 1\}$, and $f(\boldsymbol{x})$ as the predicted, the loss function helps find the optimal hyperplane is as below:

$$c(\boldsymbol{x}, y, f(\boldsymbol{x})) = \begin{cases} 0, & \text{if } y * f(\boldsymbol{x}) \geq 1 \\ 1 - y * f(\boldsymbol{x}), & \text{otherwise} \end{cases} \tag{2.3}$$

SVMs are widely used for classification and regression in many real-world problems, and there are a number of reasons for that. First, the algorithm can solve both linear and non-linear classification problems by having different kernels, which means it is suitable for toxicity prediction where there are complex relationships between data points. Second, working with support vectors (data points on the margins) makes SVM comparably memory systematic, which is an essential aspect in medical data storage since not all clinical tools can be hosted on systems with large memory resources. Moreover, it can work comparably well in high-dimensional spaces and support multi-class classification ([17]).

However, due to the computational complexity of the method, it does not perform efficiently when the training dataset is large. Additionally, focusing on support vectors, SVM cannot perform well when there are noises or the target classes are overlapped.

Random forest

Random forest (RF) is a popular machine learning algorithm for both classification and regression that is made up of multiple decision trees. In the case of classification, the output of an RF model is the majority vote of all the decision trees, while for regression, it is the average of them. Fig. 2.2 illustrates a simple example of a random forest with three decision trees for a classification problem.

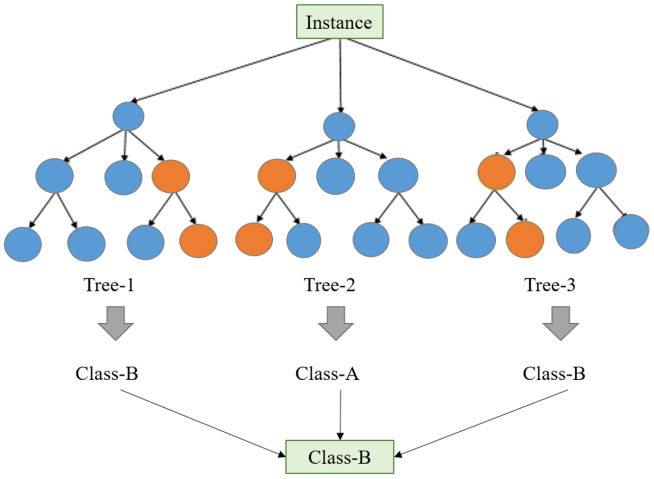


Figure 2.2: An example of a classification problem solved by random forests. The nodes in the leaf layer (last layer of the tree) are possible categories, and the nodes in the intermediate layers are input features. The orange nodes show the decision nodes.

A decision tree is a supervised machine learning algorithm that follows a collection of if-else conditions to generate a classification output. The root and intermediate nodes within the tree represent the features/attributes of the data in the dataset. Leaf nodes are the prediction of a numerical or categorical value for regression or classification, respectively. A typical decision tree is reconstructed by recursively assigning the best attribute to each node, from top to bottom (see Algorithm 1). The word "best attribute" refers to the attribute that best splits the features. This is accomplished by certain evaluation metrics, such as Entropy and Gini index for categorical data and Mean Square Error (MSE) for continuous values ([17]).

Algorithm 1: Decision Tree Induction Algorithm

Data: S = set of classified instances
Result: decision tree

```
while all partitions processed do
    maxGain ← 0
    spliA ← null
    e ← Entropy(Attributes)
    for all Attributes a in S do
        gain ← InformationGain(a,e)
        if gain > maxGain then
            maxGain ← gain
            splitA ← a
        end
    end
    Partition(S, splitA)
end
```

There are several key benefits that random forest presents when used for learning problems. The main advantage of using them is that they reduce the risk of overfitting; when there is an adequate number of decision trees, the averaging of uncorrelated trees remarkably lowers the prediction error and overall variance. Additionally, similar to SVM, decision trees can capture non-linear patterns within the training data, ;therefore, they are suitable for solving non-linear problems, including toxicity prediction. Moreover, in medical datasets missing values are very common and random forests are able to create decision trees to handle the missing feature. The other key feature of RF is the simplicity of the algorithms, which makes it straightforward to identify the important features for the final decision. Therefore, detecting clinical risk factors in toxicity prediction can be easily understood and accomplished.

However, since RF can be trained on large dataset, computing each individual decision tree may slow down the learning process and increase the time and memory complexity. Furthermore, the final output of an RF is a majority vote of all the trees, and this makes them highly sensitive to imbalanced datasets (datasets where one outcome category is much more common than others).

2.2.2 Deep learning and convolutional neural networks

Deep learning is a subset of machine learning methods based on artificial neural networks (ANN). An ANN, inspired by the biological neural model, is a computing system that is developed to mirror the learning procedure of a human brain. The basic structure of a neural network is composed of collections of layers and nodes connected to each other. The first layer receives the input data; mathematics computations are processed through middle (hidden) layers; and the last layer returns the results. In the case of toxicity prediction, the first layer is the patient's data (CT scans, dose, etc.), and the last layer is the predicted toxicity (such as "toxicity yes/no"). Based on the neurons' connectivity and network architecture, there are various types of neural networks. A fully connected neural network (FcNN), where all the neurons (nodes) in one layer are connected to all the neurons of the next layer, is the initial type of proposed neural networks. Fig. 2.3 illustrates the architecture of a FcNN.

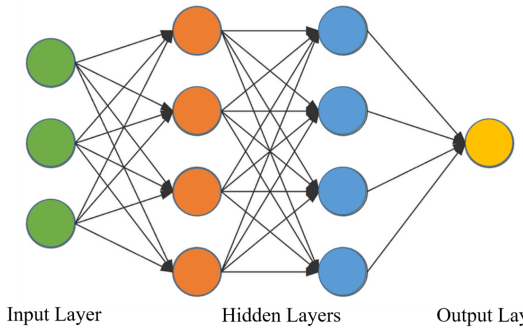


Figure 2.3: The architecture of a fully-connected neural network

The most common type of ANN is the convolutional neural network (CNN;[90]) in which a specific number of nodes in each layer are connected to the next layer through different kernels (see Fig. 2.4). The input of each neuron is computed by a mathematical operation called "convolution". CNN leverages three prominent ideas that help the network be more efficient. These are:

(i) sparse interactions/connectivity: helps to detect local meaningful features

(ii) parameter sharing: improves memory requirements and computing operations

(iii) equivariant representation: the distinctive feature of convolution operation, extract shift-invariant and more robust features compared to FcNNs.

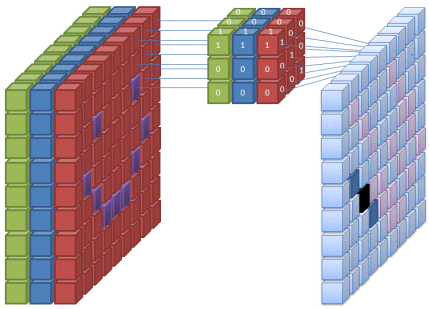


Figure 2.4: The schematic illustration of convolution layer. The middle layer represents a convolutional filter with kernel size (3×3×3). Image credit: Wikimedia[28]

CNNs are composed of a series of layers, each of which performs a specific operation on the data. A typical CNN includes:

- Convolution layer: The convolution layer applies a filter to the input image, which extracts features from the image. The output of the convolution layer is a feature map, which contains the extracted features. Convolution operation [90] is a mathematical operation that takes two functions as input and produces a third function as output. In the context of CNNs, the two functions are the input image and the filter. The filter is a small matrix of weights that is used to extract features from the input image. The convolution operation is applied to the input image at every possible location, and the output is a new image that contains the extracted features. For a 3D image, the convolution layer computes the features as below:

$$(f * g)(x, y, z) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} f(i, j, k) \cdot g(x - i, y - j, z - k), \tag{2.4}$$

where f represents the input image volume, g denotes the filter or kernel, and (x, y, z) represents the spatial coordinates within the output feature map.

- Pooling layers: Pooling is a down-sampling operation that reduces the size of the output image from the convolution operation. Pooling is often used to reduce the computational complexity of CNNs and to improve the generalisation performance of the models. There are two main types of pooling: max pooling and average pooling. Max pooling takes the maximum value from a small neighbourhood of pixels, while average pooling takes the average value from a small neighbourhood of pixels.

- Activation function: An activation function is a non-linear function applied to the output of the convolution/pooling layers. There are several activation functions used in CNNs, such as the Sigmoid function, the hyperbolic tangent function (tanh) function, and the rectified linear unit (ReLU) function.

- Loss function: A loss function is a measure of the difference between the output of the CNN and the ground truth. The loss function is used to train the CNN by minimising the error between the predicted output and the ground truth. There are various loss functions that can be used in CNNs, such as the mean squared error (MSE) loss function and the cross-entropy loss function.

- Optimisation algorithms: It finds the optimal values for the network's parameters that minimise the loss function and improve the network's performance on the given task. The optimisation algorithm plays a crucial role in training the CNN by iteratively updating the model parameters based on the gradients of the loss function with respect to those parameters. There are several optimisation algorithms that can be used to train CNNs. Some of the most popular optimisation algorithms for CNNs include Stochastic gradient descent (SGD) [84], momentum [125] and Adam [86].

- Batch normalisation: It is a technique used in CNN to normalise the output of each

layer. During training, as the network learns, the distribution of the extracted feature can change. This can make training more challenging as the network has to constantly adapt to these changing distributions. Batch normalisation addresses this by normalising the features maps. It calculates the mean and standard deviation of them the and applies a normalisation transformation. This ensures that the mean of the features is zero and the standard deviation is one. It can speed up training, and increased stability and generalisation of the network. It also reduces the sensitivity of the network to the choice of learning rate and can act as a regulariser (regularisers are mathematical functions or penalty terms that are added to the loss function during the training process), reducing the need for other regularisation techniques.

- Dropout layer: Dropout layer randomly sets a fraction of input units to zero during training, which helps to prevent the network from relying too heavily on specific input features. During each training iteration, dropout randomly "drops out" (i.e., deactivates) a certain percentage of neurons in the layer. This means that the remaining neurons must collectively learn to represent the complete set of features, rather than relying on a few dominant features. Therefore, it encourages the network to learn more robust and generalised representations of the input data.

Apart from dropout layer, there are several regularisation approaches that can be applied to deep model to improve generalisation. These techniques aim to prevent overfitting by introducing constraints that discourage the model from learning overly complex patterns that may not be relevant to the general data distribution. Commonly employed techniques include L1 and L2 regularization to control weight magnitudes, early stopping to prevent training beyond the point of diminishing returns, data augmentation to increase training data diversity, and weight decay to penalize large weights.

Deep neural networks and in particular CNN-based models have been very successful in real-world applications ([59]) and they have been applied to many clinical tasks including radio-

therapy outcome prediction, which will be reviewed further in this chapter.

2.2.3 Transfer learning

Use of deep learning models in medical image processing is rapidly growing. However, on of their major limitation is the lack of generalisation for unseen data. This poses a significant challenge when applying these models to clinical practice. Additionally, the performance of the deep models is related to the amount of training data, and medical datasets are often limited. With a small dataset, the network may not be able to accurately learn the data distribution. The obvious solution to this limitation is to acquire more data. However, it is not always possible to obtain the exact supervised data required. In such cases, transfer learning may be considered a viable alternative.

Transfer learning ([118]) is a technique in deep learning where a model trained for a task (called the "source task") is used as the starting point of training for another task (called the "target task"). The general idea of transfer learning is to use the knowledge learned by one model for another where there is not enough data to train exclusively for the second task. Instead of learning a problem from scratch, the model starts the learning procedures using the already learned patterns as initiation. In image analysis with neural networks, the model tries to detect low-level features such as edges and corners in the earlier layers, shapes and textures in the middle layers, and task-specific features in the last layers. Therefore, the earlier and middle layers can share knowledge across specific tasks as they solve a similar feature extraction problem across different medical image analysis applications. However, depending on how similar the target and source tasks are, the transferring weights of different layers may differ. As an example, the weights of a network trained on object detection task can be assigned to the initial weights of a network for segmentation task (see Fig.2.5).

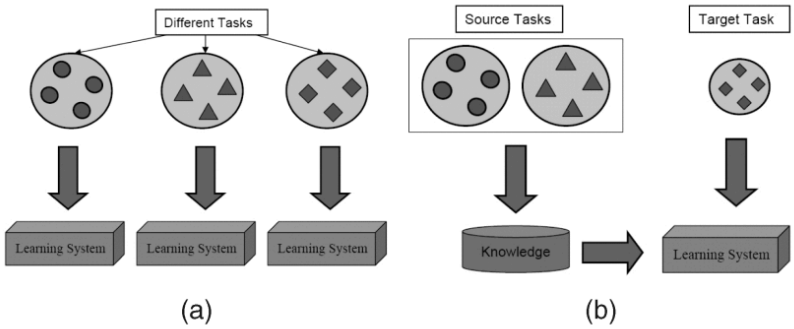


Figure 2.5: Comparison between (a) traditional learning process and (b) transfer learning. Image credit: S.J. Pan et al.[118]

.

Use of pre-trained networks can be performed in two modes; feature extraction mode, where the transferred weights are not involved in the training and their values do not change, and fine-tuning mode, where the transferred weights are in the set of learnable parameters and their values are updated by the training algorithm.

This knowledge sharing approach offers several advantages; The main advantage is that it helps the learning process for applications with comparatively small datasets or unlabelled data. Acquiring and labeling large amounts of data can be time-consuming, expensive, and impractical in many domains. however, with transfer learning, the pre-trained model has already learned to extract meaningful features from the big dataset, which can be used for the new task with a small dataset (by fine-tuning the model's parameters)[118]. Donahue et al. [40] introduce DeCAF, a deep convolutional activation feature, showcasing the effectiveness of pre-trained models on generic visual recognition tasks with limited data. Moreover, transfer elarning can improve the network generalisation [174]; by leveraging knowledge from pre-trained models it encourages the new model to learn robust features that are not specific to the training data. Yosinski et al. [174] demonstrate that features learned from one task can be transferable to another, enhancing the model's ability to generalize to new, unseen data. Finally, it significantly reduces the training computations, resulting in faster convergence and a shorter training time;

by adopting a superior initialization for the weights in a deep network, the minimisation of the loss function can be accelerated, preventing the network from becoming ensnared in local minima. This improved starting point within the loss function landscape offers a more favourable commencement, particularly amplifying the benefits of transfer learning. [180].
In summary, transfer learning is an important technique for improving generalization in deep learning models. It can reduce overfitting, encourage robust feature learning and allow to leverage existing knowledge. This makes transfer learning a valuable tool for researchers and practitioners in a variety of domains.

Demis Hassabis, The CEO of DeepMind argues that transfer learning is the key to artificial general intelligence [64]. In his YouTube video, he illustrates transfer learning using a gamer analogy. If you are an experienced gamer who has mastered multiple video games, and you and someone who has never played any video game before, start playing a new game together for the first time, you would have a better performance. This is because your previous gaming expertise allows you to recognise and apply repetitive patterns. This is exactly the concept of transfer learning.

While transfer learning is a powerful technique, there are some potential pitfalls and challenges to consider. For example, one of the main challenges of transfer learning is domain mismatch. This occurs when the source domain (the domain from which the pre-trained model was trained) is different from the target domain (the domain for which the model is being fine-tuned). For example, if you are using a pre-trained model that was trained on images of cats and dogs to classify images of cars, you may encounter domain mismatch. This is because the pre-trained model may not have learned the features that are important for classifying cars. Transfer learning can also introduce bias to the model. This is because the pre-trained model may have been trained on a dataset that is biased. For example, if the pre-trained model was trained on images of cats and dogs that were mostly male, the model may be biased towards classifying male cats and dogs as cats and female cats and dogs as dogs.

To overcome these pitfalls, careful consideration should be given to the compatibility between the pre-trained model and the target task. Additionally, fine-tuning strategies and data augmentation methods can be applied to mitigate overfitting and improve the performance of transfer learning.

2.2.4 Evaluation metrics for classification

The most important task in developing a machine learning model is evaluating its performance on a test set. One evaluation should be performed during the training to make sure the model is not over-fitted, and one should be done on an unseen dataset after the training to evaluate the performance of the trained model. There are various evaluation metrics specifically defined for classification problems. Prior to deploying an ML-based model to new unseen data, there should be a proper assessment of the performance with different evaluation metrics. Without it, the model may face problems such as poor generalisation and wrong predictions. Because this thesis research question is a binary classification, the classification metrics are explained based on binary prediction.

Accuracy, as the simplest classification metric, measures how often the model correctly predicts. Considering TP as true positive, TN true negative, FP false positive and FN false negative, accuracy is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{2.5}$$

Accuracy can show the performance of the model when the training data is well-balanced. In case of imbalanced dataset (which is very common for medical applications), a biased model can report accuracy > 0.9 where most of the samples in the evaluation set are from the majority class. Therefore, it is not enough to only report the accuracy for the model performance. The confusion matrix is defined to describe the performance by reporting the number of false/true predictions. It reports TP, FP, FN and TN in a matrix. Combining the values in a confusion

matrix to a single metric, F1_score reports:

$$\text{F1\_score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}. \tag{2.6}$$

A higher value for the $F1\_score$ shows higher prediction performance for a model. For many medical tasks where the detection of rare events is important (including toxicity), it is necessary to have an evaluation metric that measures a model's performance in predicting minority class. The sensitivity metric indicates how the model learned the positive class distribution and how well it can predict the true positive labels. It can be written as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2.7}$$

Specificity is defined as the proportion of all the actual negative samples that were correctly predicted. The mathematical notation of specificity is as

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{2.8}$$

A report of both sensitivity and specificity can declare if the model is biased. In the case where the model is biased to generate negative/positive labels, the sensitivity/specificity value will be very low, while specificity/sensitivity metric will be relatively high.

All of the evaluation metrics mentioned above analyse the performance based on the final results of the classification. Consider two binary classifiers, A and B, where model A predicts label one with a probability of 0.6 and model B predicts it with a probability of 0.9. In this case, while the prediction threshold of 0.5 yields the same accuracy for both models, clearly model B is more convinced in its prediction. The receiver operating characteristics (ROC) curve is a chart depicting the true positive rate (TPR) against the false positive rate (FPR) at various threshold values. Plotting the ROC curve can illustrate the superiority of model B over

model A. Changing the threshold value in $[0, 1]$, ROC curve shows the ability of the classifier to distinguish between two classes. In order to quantify the superiority of the two models, the area under the ROC curve, known as the AUC, can be reported. When 0.5 <AUC< 1 (true positive rate is higher than the false positive rate), the classifier is able to distinguish between positive and negative classes better than a random guess.

2.3 Review of current models for radiotherapy outcome prediction

In radiotherapy, tumour control probability (TCP) is a metric to determine the proportion of tumour killing with a given radiation dose, while normal tissue complication probability (NTCP) characterises the predicted damage to normal tissues as a function of dose to those tissues. It can be challenging to attain a treatment outcome that has a high TCP and a low NTCP since the correlation between dosimetric and clinical data, and the desired outcome is not clearly understood. In 1991, one of the earliest and most highly-cited studies carried out by Emami et al.[48], proposed that certain significant complications are influenced by both cumulative dose and the volume of organ-at-risk exposed to radiation. They provided practical guidelines based on basic calculations, suggesting a 5% and 50% NTCP for complete, two-thirds, and one-third uniform irradiation of individual organs-at-risk, assuming the remaining volume receives zero dose. After this work, Kutcher et al.[89] proposed a dose–volume histogram (DVH) reduction algorithm that reduced an arbitrary non-uniform dose distribution into a partial volume that receives the highest dose. This DVH-based analysis inherently presumes that organ function is evenly distributed throughout the organ, while experimental animal studies on the volume effect have provided crucial proof-of-principle that challenges this assumption [14]. However, most NTCP modelling has still relied on this basic assumption, discarding spatial heterogeneity in dose-response.

Since 1991, a lot has changed, and numerous clinical studies have been published on the analysis of outcome prediction based on dose and clinical data. As an example, one category

of NTCP models transfers 3D dose data into a single value, commonly referred to as "effective volume," which is the effective dose for a specified reference volume. This value is then connected to the probability of normal tissue toxicity using a Sigmoid function [152], [33]. The DVH reduction algorithm proposed by Kutcher et al. is one specific example of such an 'effective volume' reduction approach. Dawson et al.[34] proposed using principal component analysis (PCA) to analyse the partial volume effects of normal tissues under radiation. Employing PCA, they identified the variance in cumulative DVH. Features with the largest variance in the DVH were further studied as related to complication risks. Another model proposed by Bonta et al.[18] relied on Critical Volume NTCP models. The authors assumed that organ complications happen when radiation damage to a non-target organ exceeds a specific threshold, which depends on the size of the organ. They used logistic regression to predict the probability of a patient developing toxicity and the maximum likelihood for model parameter estimation. In the same year, Stavreva et al. [142] proposed a new model that is a more complex version of the traditional Critical Volume NTCP models. Other modelling approaches have also been proposed, such as multivariable modelling [43], Bayesian networks[92], machine learning classifiers, and artificial neural networks. Machine learning and neural networks are reviewed in the following sections.

In summary, there are certain difficulties associated with traditional outcome modelling approaches, and care should be taken when employing these models; typically, they are based on data extracted from DVH, which is not an accurate representation of the 3D doses for three reasons: (i) the spatial information of doses is discarded; (ii) it imposes that all the regions have equal functional importance; and (iii) the radiobiological fraction size effects are not taken into account.

2.4 Review on machine learning for radiotherapy outcome prediction

Radiation oncology has leveraged machine learning techniques for both therapeutic and prognostic purposes in recent years ([81] [44], [168]), and many studies have been proposed to improve the radiotherapy treatment workflow. These new techniques have gained rapidly growing interest in this field of study. Deep learning, as an extremely popular branch of machine learning models, has shown promising success in radiotherapy-related tasks [113]. Fig.2.6 illustrates the number of publications applying ML and deep neural network models for radiotherapy since 2015.

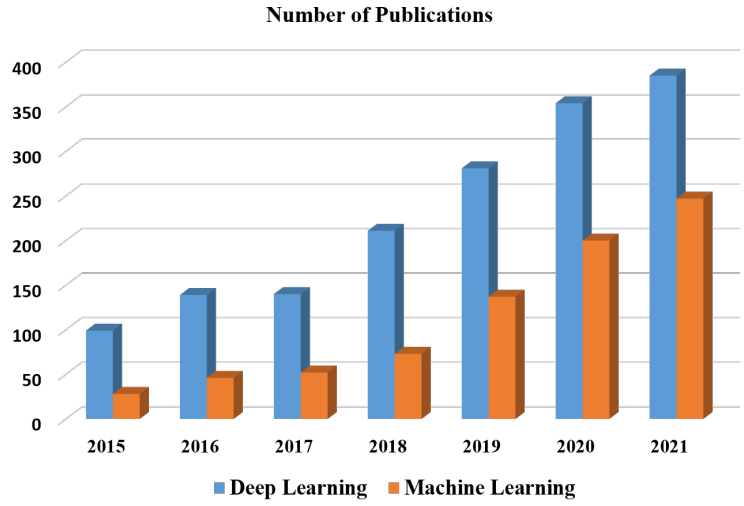


Figure 2.6: The number of publications applied machine learning and deep learning models in the radiotherapy field. The search has been done with at least the phrases "radiotherapy" or "radiation oncology" or "radiation therapy" and for machine learning approaches, at least the terms "support vector machine' or "logistic regression" or "random forest" and for deep learning, at least the terms "deep neural networks" or "convolutional neural network" in their titles. The statistics are obtained from Google Scholar.

Among all machine learning methods, logistic regression, support vector machines and random forest are the most popular classification models utilised in several studies. Logistic regression is ideal when the predictor features are not complex and are linearly connected to the outcome.

Sini et al. [140] adopted a multivariate logistic regression model to predict patient-reported intestinal toxicity in prostate cancer. They had access to selected clinical and dose-volume data for 206 patients in their dataset. In a similar study, Lee and colleagues [93] used a multivariate regression model to show that both dosimetric and non-dosimetric features were associated with xerostomia toxicity. They evaluated the significance of risk factors and noted that age, smoking, financial status, local tumor extent, and alcohol abuse had relationships with xerostomia toxicity. Robertson et al. [131] explored the dose-volume relationship of bowel irradiation and severe diarrhoea using a logistic regression model. With a dataset of 152 patients, their results revealed highly significant correlations ($p < 0.001$) between small bowel receiving at least 15 Gy and the presence of toxicity.

To cope with more complex and non-linearly separable data, support vector machines were widely used for binary classification in radiotherapy outcome prediction. Chen et al. [23] proposed an SVM model to predict grade$\geq$ 2 of radiation-induced pneumonitis for patients with lung cancer. They trained the SVM on a dataset of 219 patients, where a total of 93 features consisting of dose and non-dose factors were collected for each patient. Experimental results were performed to evaluate dosimetric features with and without clinical factors. The results demonstrated that the combination of both significantly improved the prediction performance; the AUC was 0.76 when both types of predictors were used, while it was 0.71 when only dosimetric features were trained. By analyzing the coefficients of SVM, they reported that the mean lung dose and chemotherapy prior to RT were the most powerful predictor features.

In another work by the same group, [134], the authors proposed to combine multiple ML models in order to empower the prediction. They used ensemble decision trees, SVM, and neural networks to predict pneumonitis. The average of the predictions from each model was used as the final prediction. The results showed that combining multiple models was more robust than each model individually. Klement et al. [87] conducted a comprehensive analysis and discussion comparing SVM and traditional outcome modeling. Klement and colleagues investigated

the support vector machine to predict the local TCP after radiation therapy for non-small cell lung cancer. A cohort of 399 patients was included in their study, and different combinations of prognostic features in the dataset were considered for training. The results showed that the strongest predictor was the biologically effective dose at the isocenter. The comparison of AUC showed that SVMs were superior to traditional outcome modeling in stereotactic body radiation therapy. Wang et al. [161] designed three different SVM models to evaluate which variables were important for locally advanced nasopharyngeal carcinoma. Patients' characteristics, including age, gender, cancer stage, therapeutic regimen, and 38 molecular biomarkers, were considered for 49 cases for model training. They reported the significance of each clinical variable by extracting the model coefficients.

For further evaluation of correlated features and radiotherapy-induced toxicity, some studies utilized random forest models. In [98], the authors applied RF to investigate the correlation of multiple variables with pneumonia caused by radiotherapy in patients with esophageal cancer. A dataset of 118 patients, of whom nearly 61% developed radiation pneumonia, was included in this study. Wang et al. [164] employed random forest with the aim of exploring risk factors for reactivation of hepatitis B virus after radiotherapy in patients with liver cancer. They established various RF models to investigate the key factors in the prediction and compared the results with Bayesian classifiers. The Bayesian prediction model was trained using the top five prognostic features as determined by random forest. The experimental results showed that RF could predict the hepatitis B virus with an accuracy 1% higher than the Bayesian model. In another work, [36], RT dose-volume and spatial dose metrics were used to predict acute mucositis after head-and-neck radiotherapy with three ML models: logistic regression, support vector machine, and random forest. The comparison of AUC demonstrated that the RF model had the best prediction power.

2.5 Review on deep learning for radiotherapy outcome prediction

Deep neural networks have transformed many areas of medicine, and they have the potential to handle various challenges faced in radiation oncology. The radiotherapy workflow consists of several complex tasks, including initial decision-making for treatment, organ and tumour segmentation, treatment planning and dose treatment optimisation, tumour motion management, outcome modelling, and quality assurance. Deep learning can potentially play an important role in most of the RT steps in order to assist the clinical team [74], [157], [138]. For the initial treatment steps, deep neural networks can extract clinically crucial features that help with treatment decisions. For example, the analysis of the pathological response of lymph nodes in patients treated with chemotherapy can help the clinician decide on the radiotherapy treatment [82]. In the tumour/organs segmentation step, deep neural networks have shown promising performance, and they are currently used in many real-world medical tools [69], [41]. As an example, automated organ at risk segmentation of CT images using deep networks before treatment planning in head-and-neck cancer radiotherapy [116]. In terms of treatment planning and optimisation of dose prescription, DL models can assist in personalising the treatment. There are various neural networks that generate optimised dose treatment plans [107] or predict the optimal individual patient radiation dose distribution [24], [106]. Predicting tumour motion using 4D CT images [100], and quality assurance of treatment [149] are other examples of deep learning assisting in the radiotherapy workflow.

Outcome modelling can also benefit from neural networks, as they have the potential to take in more detailed information for response prediction. In recent years, there has been an increasing amount of literature on employing deep learning only for radiation therapy outcome prediction.

Much of the current literature has used three-dimensional images as input into the network. However, there are relatively few studies that have focused on two-dimensional networks. Zhen et al. [177] introduced a 2D CNN model to predict rectum toxicity for patients after cervical cancer radiotherapy. They used a pre-trained CNN with 16 convolutional layers to predict

grade≥ 2 rectum toxicity. For the model input, a 2D dose surface map was constructed by unfolding the 3D dose distribution of the rectum. They also compared the prediction performance with logistic regression that was trained with only dose information. The comparison results showed that the features extracted by the CNN are more powerful than 1D dose-volume parameters for outcome prediction, and the proposed deep network outperformed the logistic regression model. Although unfolding 3D dose can be applied to the rectum, it cannot be extrapolated to most of the other organs; the rectum is a hollow structure and approximating it with an unfolded 2D surface does not lead to information loss, unlike many other organs.

2.5.1 Analysing only dose data

Liang et al.[99] used a three-dimensional convolutional network to determine the relationship between the 3D dose image and the RT outcome. They employed C3D [151], a well-known convolutional network for the task of video action recognition pretrained on the UCF101 dataset [141], to predict radiation pneumonitis in patients with non-small cell lung cancer. Moreover, they analysed three multivariate logistic regression models with numerical data, including dosimetric factors, NTCP, and dosiomics features. The comparison results showed that the CNN outperformed the LR models, with an AUC of 0.84 compared to 0.78 for the latter. Similarly, another group [76] reported that deep learning methods produced almost two times fewer false positives for toxicity prediction compared to DVH-based models. They proposed a 3D CNN to investigate patterns in the dose treatment plan and their association with hepatobiliary (HB) toxicity after liver stereotactic body radiotherapy (SBRT). The authors also proposed a fully-connected neural network to explore the correlation of clinical features with HB toxicity. The weighted sum of the two predictors was considered the best prediction model.

2.5.2 Analysing dose and CT data

To increase the prediction power, several studies have considered the effects of other data in addition to the dose distribution for outcome prediction. In [75] the authors used a concatena-

tion of dose images and CT scans as the inputs in their network. They proposed a 3D CNN with residual blocks to investigate the relationship between CT scans and dose with late HB toxicity. In the same vein, Yan et al. [172] proposed utilising both dose distribution and CT scans to train a CNN with the aim of predicting post-treatment gastro-urinary function after prostate radiotherapy. The input of their network was a two-channel (dose and CT images) path, extracting significant features within 3D convolutional blocks. The results showed there is a substantial association between the dose irradiated to the bladder and side effects after treatment. This view was supported by Men and colleagues [112], who propose a three-path residual 3D CNN for predicting xerostomia after radiotherapy for head and neck squamous cell carcinoma. The network consisted of three separate convolutional paths, each processing dose distribution images, CT scans, and region of interest contours. The output of the three paths was summed and passed through four linear layers to predict toxicity. They reported the prediction performance with various experiments; first, they investigated the importance of each input. They trained the network for four input modes when they have (i) all three inputs, (ii) all but the contour data, (iii) all but the CT data, and (iv) all but the dose distribution data. With an AUC of 0.84, the best performance was obtained when all three data were used as inputs. The contour images were the least associated input data, with an AUC of 0.82 (when trained with dose and CT data), while the dose image was the most associated with toxicity (AUC of 0.70 when trained without dose data). They also compared the results of their network with LR models. The inputs of LRs were clinical and dosimetric variables, and the best AUC was reported at 0.74.

In a study of predicting local failure, Aneja et al., [4] used a convolutional neural network to analyse only CT scans. They employed a 3D CNN with the input of CT images to predict local failure following SBRT. Additionally, they analysed the prediction power of clinical risk factors (age, histology, biology, etc.) with three ML techniques: random forest, SVM, and LR. The comparison results demonstrated that the best performance is achieved when the deep neural network analysed CT imaging, with an AUC of 0.81. The second-best performance was

for random forest with an AUC of 0.69, and this was followed by SVM and LR with 0.65 and 0.59, respectively.

2.5.3 Analysing other imaging modalities

There are relatively few works that have considered other imaging modalities than standard 3D CT for RT outcome prediction. Bin and colleagues [15] focused their study on dose distribution, ventilation imaging (VI; derived from four-dimensional CT), and functional dose (obtained by weighting the dose image with VI). They used C3D network [151] as a feature extractor and applied ML methods including SVM, LR, K-nearest neighbour, and random forest to predict pneumonitis after thoracic radiotherapy. In [163], slice concatenation of pre-RT CT, pre-RT positron emission tomography (PET) imaging, and dose distribution were used for training an 8-layer CNN in order to predict the outcome of radiotherapy for oropharyngeal cancer. In another study, Wang et al. [162] developed a recurrent neural network with input of MRI and cone-beam CT scans to predict the outcome of lung cancer treatment. Hongming and colleagues [96], proposed to analyse only CT scans and PET images to predict the survival rate after radiotherapy for rectal cancer. They compared the results with traditional survival prediction (based on radiomic features) and reported that their deep network had higher prediction performance.

2.5.4 Analysing imaging combination with clinical factor

Although much of the current literature attempted to evaluate the association of 3D (imaging and dose) data with RT-induced toxicity, there are a number of studies that have explored the combination of image and clinical features for outcome prediction. Predicting cancer recurrence and survival rate in SBRT [77], the authors proposed a multi-path convolutional network; one path for the input of the 3D dose distribution and the other path for the treatment features (including patients' demographics, OAR properties, tumour size, laboratory measurements of the liver function, etc.). Each path extracted the important features, and their concatenation

was passed through a fully-connected layer to predict the outcome. Similar to [77], Welch et al. [166] developed a pipeline to combine clinical data with dose distribution, CT scans, and contour data to predict locoregional failure at 3 years after head and neck radiotherapy. They compared the results of the neural network pipeline with LR and RF methods (trained on numerical features). According to their findings, the logistic regression model had the highest AUC. The authors argued that these results could be due to the breadth of information included in their clinical data, which requires a less complicated modelling technique.

2.5.5 Analysing only clinical data

There are a limited number of studies that employed deep neural networks for DVH data. In a study conducted by Qi and colleagues [124], the authors investigated the potential correlation between dosimetric features and patient-reported quality of life (QOL). They applied numerical predictors extracted from the dose-volume histogram to a fully-connected deep network to predict the score of QOL for urinary functions and the rectal domain. An accuracy of 0.90 showed that radiotherapy for prostate cancer can affect rectal-related QOL, and there is an association between urinary functions and the dose irradiated for treatment. In a similar work [30], the authors used multiple neural networks (with different architectures) to predict local control after radiotherapy for patients with non-small cell lung cancer. They trained the networks with biological features (levels of cytokines, micro-RNAs, and single nucleotide polymorphisms), features extracted from PET imaging data, and dose-volume histograms.

Overall, there is a large and growing body of literature investigating RT outcome prediction with the help of deep learning models. The studies are reviewed and summarised focusing on specific aspects. Table 2.1 illustrates a summary of the reviewed works.

Table 2.1: Summary of studies utilising neural networks for radiotherapy outcome prediction

Ref	Cancer	Dataset	Summary of Network	Predicted Outcome	Visual Explanation
[177]	Cervical	42 patients. 3D Dose treatment plans	16-Layer 2D-CNN (Adopted from VGG-16 [139])	Grade≥ 2 rectum toxicity	Grad-CAM
[99]	Lung	70 patients. 3D dose treatment plans	5-layer 3D-CNN pretrained on video classification	Grade ≥ 2 Radiation pneumonitis	Grad-CAM
[76]	Liver	125 patients. 3D dose treatment plans, non-dosimetric features	3-layer 3D-CNN pretarined on anatomical images. Non-dosiemtric features are used to train FcNN	Grade ≥ 3 hepatobiliary toxicities	Saliency maps created by systematically varying dose
[75]	Liver	122 patients. 3D dose treatment plans, 3D CT scans	10-layer 3D-CNN pretrained on anatomical images. The input is a concatenation of dose and CT	Grade ≥ 3 hepatobiliary toxicities	same as their previous work [76]
[172]	Prostate	52 patients. 3D dose treatment plans, 3D CT scans	3-layer CNN with two input channels each processing one input.	Urinary and bowel symptoms	N/A
[112]	Head and neck	784 patients. 3D dose treatment plans, 3D CT scans, contours	6-layer 3D-rCNN with three input channels	Grade ≥ 2 xerostomia	N/A

Table 2.1: Summary of studies utilising neural networks for radiotherapy outcome prediction

Ref	Cancer	Dataset	Summary of Network	Predicted Outcome	Visual Explanation
[15]	Thoracic	217 patients. 3D dose treatmen plans, VI, functional dose	8-layer 3D-CNN pretrained on video classification (C3D [151])	Grade $\geq$ 2 Radiation pneumonitis	N/A
[163]	Oropharyngeal	66 patients. 2D dose treatment plans, 2D CT scans, 2D FDG-PET images (all on axial slices)	8-layer 3D-CNN. A concatanation of 2D PET/CT and dose is taken as the input	2D axial PET images at mid-treatment	N/A
[162]	Lung	11 patients and a public dataset [73] with 13 patients.	2D-RNN with 6 residual blocks	Acute esophagitis	N/A
[77]	Liver	120 patients. 3D dose treatment plans, clinical features	two-path network from [76] for dose data and numerical data.	Post-SBRT survival and local cancer progression	Saliency maps created by systematically varying dose
[166]	Oropharyngeal	160 patients. 3D dose treatment plans, 3D CT, contours, clinical features	4-layer 3D-CNN with three channels for dose, CT and structure set	Locoregional failure at 3 years	N/A

Table 2.1: Summary of studies utilising neural networks for radiotherapy outcome prediction

Ref	Cancer	Dataset	Summary of Network	Predicted Outcome	Visual Explanation
[4]	NSC Lung	344 patients. CT scans, clinical factors	3D-CNN trained with CT scans and FcNN trained with clinical factors. The CNN is pre-trained on 1,009 CT scans from an outer dataset LIDC [7]	Local failure	N/A
[96]	Rectal	84 patients.CT scans, PET images	Two 3D-CNN with the same architecture for analysing CT and PET separately.	Time of local tumour recurrence	N/A
[124]	Prostate	86 patients. Dosimetric parameters	Deep fully-connected network with 200 numbers of neurons in hidden layers	Quality of life score for urinary and rectal functions	N/A
[30]	NSC lung	98 patients. Biological variables and features extracted from dose/PET images	Three different architeccutres: 1D-CNN, locally-connected network and FcNN	Local control	N/A
[45]	Lung	219 patients, 1D Clinical features	three-layer feed-forward neural network	Grade $\geq$ 2 Radiation pneumonitis	N/A

Table 2.1: Summary of studies utilising neural networks for radiotherapy outcome prediction

Ref	Cancer	Dataset	Summary of Network	Predicted Outcome	Visual Explanation
[143]	NSC Lung	142 patients, 1D features extracted from lung dose-volume	FcNN with three layers	pneumonoitis	N/A
[21]	Prostate	664 patients, 1D Clinical features	1 layer feed-forward network	Faecal incontinence	N/A
[148]	Prostate	718 patients, Clinical + features from DVH	3 layer fcNN	Late rectal bleeding	N/A
[63]	Prostate	119 patients, dosimetric features extracted from DVH	1 layer feed-forward network	bladder and rectum complications, Biochemical control	N/A
[122]	Prostate	321 patients, clinical and dosimetric features	Different ANN architectures are tested. The best result is for FcNN with 2 layers.	astro-intestinal and genito-urinary toxicities	N/A
[171]	Prostate	754 participants, clinical and dosimetric features	1 layer feed-forward network	Urinary toxicities	N/A

Abbreviations: CNN, convolutional neural network; FcNN, fully-connected neural network; rCNN, residual CNN; RNN, recurrent neural network; SBRT, stereotactic body radiation therapy; NSC, non-small cell; VI, ventilation image; DVH, dose-volume histogram; ANN, artificial neural network.

2.5.6 Current challenges and opportunities

The use of artificial neural networks in radiotherapy outcome prediction is a rapidly evolving field of study that holds great promise in the improvement of cancer treatment. However, it is still in its early stages, and many challenges yet need to be addressed.

Classification modelling

The majority of papers summarised in this chapter used a binary variable to predict toxicity outcomes, usually predicting the occurrence of toxicity as "yes" or "no," despite the fact that toxicity presentation is more complex than a binary number. Clinically, toxicity often scored with ordinal grades, reflecting the severity of the side effects, or, in relatively few cases, continuous scales capturing patient-reported outcomes. Modelling the toxicity with a binary value causes a loss of information. This issue is not specific to neural networks, and traditional NTCP modelling is also affected by it. It is also important how the toxicity labels are generated; the severity of side effects reported by patients is a more accurate representation of toxicity compared to grades clinicians assign. However, progress has been made in both fields in terms of ordinal regression/classification and data labelling.

Several machine learning models have been altered to deal with the ordinal classification problems [127], [67], [137]. In particular, for neural networks, a "ranking learning" strategy has been proposed. Ranking learning models are typical neural networks with minor changes in their formulation [29], [53]. These changes are generally based on two approaches. First, they can convert the ordinal problem into a pairwise binary classification problem. In this case, the on-hot label encoding changes to group encoding in order to be compatible with the ordinal problem. As an example in [25] a target formulated as [0,0,1,0] (showing the toxicity grade 3) can be reformulated to [1,1,1,0] for ordinal classification and the standard exponential function for classification can be changed to Sigmoid function. Li et al. [95] used the idea of target changing for their neural network to predict the early diagnosis of Alzheimer's disease for four

outputs in the order of: Alzheimer's disease, mild cognitive impairment (MCI), progressive MCI, and normal control. In another work, [10], authors used the same ordinal coding to predict ulcer severity in patients with Crohn's disease. The other approach to tackling ordinal classification is to change the network architecture in order to learn multi-threshold thresholds for ordinal classification. Cao and colleagues ([19]) changed the connection of biases and weights in the last layer to restrict the neural network to make rank-consistent predictions.

Although neural networks and machine learning methods can easily deal with ordinal data, due to the small size of medical datasets, the performance of the prediction is lower in comparison with binary classification problems; because there are not enough examples for some classes (for instance, the number of patients with grade > 3 toxicity in a 4-graded problem is very low), the network is not able to learn the data distribution for the rare category, which results in biased training and low performance.

Lack of data

The size and quality of the dataset available has a pivotal role in learning algorithms, and it directly affects the performance of the model. Although different aspects come into play during network development, the data are the backbone of the entire model, and without an adequate dataset, various crucial tasks cannot be accomplished. In the medical area, it is generally difficult to create a dataset that is large enough to train a neural network. Collecting medical images needs professional expertise for labelling and contouring. It is thus common for datasets used in medical image analysis to be small. Consequently, if a small dataset is used for training, the network may not be able to effectively learn the data distribution. Additionally, in various clinical problems, including radiotherapy outcome prediction, the population of the different classes may be unbalanced. Training a deep network with a highly imbalanced dataset results in a model biased towards the larger class, as it considers rare samples as noise.

There are three main solutions that can help alleviate this problem. The first approach is to

use transfer learning, as previously discussed. This process applies information from a learned task to improve performance on a goal task, typically reducing the amount of required training data. All the reviewed studies, except the work by Men et al.,[112], which includes a dataset of 784 patients (Radiation Therapy Oncology Group; RTOG, 0522 trial, head and neck cancer), employ transfer learning to overcome the problem of their small dataset. Although transfer learning can be a solution to domain-specific data scarcity, it relies on the assumption that both "source-task" and "goal-task" are sufficiently similar in terms of input and output data. For non-medical image applications, neural networks are often transferrable (due to the similarity of low-level features), but for radiotherapy imaging data, it is more complicated; dose data are quite different from other images, lacking sharp variations and exhibiting smoother shapes over larger receptive fields. Therefore, an optimal architecture for radiotherapy might require wider kernel sizes, decreasing the performance of transfer learning. In general, there is a risk of task dissimilarity, and it requires careful investigation before applying transfer learning. However, some studies used networks that have been pre-trained on natural images or videos, and their results are convincing; for example, Ibragimov et al. [77] and Liang et al. [99]transferred the learned weights from the C3D network, which had been trained for video action classification.

There is also a specialised form of transfer learning known as domain adaptation which eliminates the necessity for retraining a model on a new dataset. With domain adaptation, a pre-trained model can be fine-tuned to achieve optimal performance on new data, saving considerable computational resources [51]. This can be achieved through various methods; instance-based adaptation, feature-based adaptation and deep domain adaptation are the three main techniques used in the literature [176],[123].

Another technique that can be employed is data augmentation, which involves making slight modifications to the existing data in order to generate new data with the same data distribution. Rotation, scaling, cropping, colour range changes, adding random noise and horizontal/vertical flipping are the most common data augmentation techniques for medical images. In [76] the

authors applied Gaussian noise with zero mean and 0.1 standard deviation to their original dataset, Liang et al. [99] augmented their dataset by flipping along three directions and Welch et al. [166] enriched their dataset by performing affine transformations prior to the training.

Unlike in other domains, certain transformations may not be appropriate for medical images. For example, horizontal or vertical flipping may not be appropriate for maintaining the anatomical integrity and spatial consistency of the human anatomy. Therefore, careful consideration should be given when applying augmentation techniques to medical datasets.

The third approach is generating synthesised data. There are generative models that are able to produce unreal or fake data with the same statistical properties and schema as the "real" samples. Generative adversarial neural networks (GANs) ([60]) and variational autoencoders(VAE), are examples of deep neural networks to create new examples of their training data. Generative data augmentations have been applied to medical and non-medical applications to generate larger datasets ([133], [5]). Oversampling techniques such as SMOTE ([22]) and registration methods such as Fischer-Modersitzki ([55]) are also utilised as data synthesising methods in radiotherapy outcome prediction. Due to the high complexity of generating 3D medical images, generative techniques are not common for medical data augmentation.

Visual explanation and interpretability

Machine learning and artificial neural networks perform challenging tasks with a high accuracy; ResNeXt-101 ([105]) with 829 million parameters has achieved 97.6% accuracy in a classification challenge on the Imagenet dataset ([39]). This model complexity makes it difficult to interpret network behaviour. In contrast, classical rule-based models are interpretable but may lack high accuracy and robustness when applied to high-dimensional data. There is usually a trade-off between interpretability and accuracy. However, the difficulty in understanding the networks' learning and responding procedures hinders their usage in real-world medical systems. In radiation oncology, there exists a need for models with explainability/interpretabil-

ity; oncologists personalise treatment for individual patients, and to empower this, they need to understand which aspects of dose distributions/radiation treatments impact the outcome. For this reason, some groups have explored visual explanations for the predicted outcome in their network.

Gradient-weighted class activation mapping (Grad-CAM)[136] is a technique for convolutional neural networks which highlights those input regions that are "important" for the prediction. In a CNN, the last convolutional feature maps (features extracted by the last convolutional layer) are the most informative features, and they carry spatial information as well as high- and low-level features. Grad-CAM uses the last feature maps and explores their impact on the network's final decision. It computes the gradient of the feature maps with regard to the final output. The gradient operation shows how any changes in each feature map affect the final decision. Higher gradients for a feature map indicate that the features extracted from it are more strongly correlated with the output of the network. More recent attention has focused on the Grad-CAM method, and a few studies have used it to explain their convolutional networks. [177] and [99] used Grad-CAM to highlight the anatomical regions with the highest activation when predicting toxicity.

Ibragimov et al. ([76], [77], [75]) also adopted the idea of gradients of features. After training their CNN, for each pixel x in the dose data, they created two artificial dose distribution, one with the value of x increased and one with the value of x decreased. Then, the two new dose plans were separately passed through the CNN, and subtraction of the predicted outputs showed how changes in x affect the prediction. They specifically examine whether "change" in input in a region affects the network's prediction. In another work by Men et al., [112], the different feature maps of the network were visualised to detect critical regions (not only the last feature maps). Although the results showed that the CNN can extract high-risk regions of the input image, it is still not clear how these regions are related to the output.

In summary, the transparency and interpretability of artificial neural networks, and clarifying

their processes to their outputs can help users trust the network's outcome, which makes them feasible to employ in real-world problems. This trust is pivotal for the practical application of neural networks in real-world problem-solving.

Visual explanation in RT outcome prediction is the main challenge, and there are relatively few methods that address this issue (see Table 2.1).

2.6 Summary and conclusions

The primary objective of this chapter was to investigate the current state of research regarding radiotherapy outcome prediction. Initially, traditional models used for prediction were examined, along with an analysis of their advantages and disadvantages. This analysis led to the exploration of how machine learning techniques can enhance prediction models. A comprehensive review of ML, particularly deep learning models, was provided. Since the focus of this research is on applying deep learning models for toxicity prediction, the chapter also elucidated the challenges faced by deep learning models. This raises the question of whether traditional ML models outperform deep learning models, given that the latter were introduced to address the limitations of the former. To address this question, experiments comparing both models are necessary.

In the later Chapters, a comparison between machine learning and deep learning models and the answer to the questions such as "whether analysing 3D data can be helpful for prediction", "which data can predict toxicity with the highest accuracy", and "whether combining clinical data with dose and imaging improves the results" are provided.

However, prior to conducting any experiments, it is essential to provide a comprehensive introduction to the dataset and the evaluation metrics employed for conducting the experiments. The next chapter will focus on thoroughly describing the dataset, including its preprocessing steps and the challenges encountered.

Chapter 3

Dataset

In this chapter, a comprehensive overview of the dataset used in this research is provided. The dataset comprises of two main sections: the clinical dataset and the image dataset, both of which are described in detail. The approaches to data preprocessing for both clinical data and 3D CT and dose data are described in details. This information is crucial for readers to understand the context and limitations of the analysis and to assess the validity and generalizability of the findings of this research.

3.1 Introduction

Data for 315 patients with four types of pelvic cancer - including anal, rectal, endometrial, and cervical cancer - were included. All patients were treated between 2004 and 2014 at the Leeds Cancer Centre, United Kingdom, with curative 3D conformal radiotherapy or IMRT. A summary of the standard treatments received for each cancer type are shown in Table 3.1.

The National Research Ethics Service Leeds East Committee approved the data collection study following ethical review with reference number 13-YH-0156. Further use of data for the current project was provided by the LeedsCAT research database with reference 19-YH-

0300.

Median duration of follow-up from radiotherapy was 2 years (IQR: 1.4-3.5 years). Patient-reported toxicities were assessed using the validated European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life questionnaires – including the core questionnaire (QLQ-C30) and disease-specific modules (QLQ-CX24, EN24 and CR29) and additional items from the EORTC Item Library to cover those missing for patients with anal cancer (for whom a disease-specific module was not available at the time of data collection) [1], [62], [167], [61], [155].

Table 3.1: Summary of usual treatment received for each cancer type

Cancer type	Treatment
Anal	Curative chemoradiotherapy: 50 Gy in 25 fractions EBRT combined with concurrent chemotherapy (mitomycin and 5-flurouracil)
Rectal	Two neoadjuvant (pre-surgery) radiotherapy schedules depending on patient fitness and stage: 25 Gy in 5 fractions EBRT or 45 Gy in 25 fractions of EBRT combined with concurrent chemotherapy (5-flurouracil) followed by curative surgery.
Endometrial	Post surgery patients received EBRT 45 Gy in 25 fractions and some patients received vaginal vault brachytherapy 12 Gy in 3 fractions.
Cervical	Curative chemoradiotherapy: 48 Gy in 24 fractions EBRT combined with concurrent chemotherapy (cisplatin) followed by brachytherapy 21 Gy in 3 fractions.

Abbreviations: Gy, gray (absorbed energy per unit mass of tissue); EBRT, external beam radiotherapy;

The dataset consists of 75 male and 240 female patients, of which 265 reported at least one of the toxicities under study. 258 clinical features, including the patients' demographics, medication and pre-RT treatment status, radiotherapy treatment information, and anatomical features, are included in the numerical dataset. 3D dose distributions, 3D CT scans, and contour structure sets for organs in the pelvis are also collected in the image dataset.

Based on the Radiation Therapy Oncology Group (RTOG) guidelines [56], the intestinal cavity structure was contoured as a "Bowel Bag" for each patient. The bowel bag is the organ at risk

for all three toxicities considered here.

Table 3.2 and Fig. 3.1 summarise the patients' statistics in the dataset. Information about cancer stages for patients in the dataset is shown in Table 3.3 and Table 3.4 (please note that the two patients who lacked clinical data were excluded, so both tables represent the data for a total of 313 patients).

Table 3.2: Summary of the treatment statistics included in the dataset.

Cancer type	Sex	Num	Age (years)	Total Dose (Gy)	Min dose (Gy)	Max dose (Gy)
Anal	F	68	62.12[10.30]	45.37[7.93]	26.06[18.65]	2.18[4.08]
	M	27	63.95[11.03]	45.478.46]	21.90[19.61]	1.46[3.09]
Rectal	F	26	61.90[17.80]	36.10[11.33]	35.98[11.92]	0.0[0.0]
	M	48	65.15[9.90]	37.0[10.84]	19.23[9.23]	0.37[1.57]
Endometrial	F	49	67.36[11.70]	46.60[5.72]	61.77[17.71]	1.19[3.20]
Cervical	F	97	49.82[12.95]	51.50[8.43]	71.97[12.31]	6.23[7.75]

Abbreviations: Num, number; F, female; M, male. For age, total dose, min dose and max dose, the mean [standard deviation] is shown in the table.

Table 3.3: Details of FIGO stage of primary diagnosis for patients in the dataset.

Cervical cancer			Endometrial cancer		
FIGO stage	Num	Percent	FIGO stage	Num	Percent
1a2	1	1.0%	1a	2	4.2%
1b	2	2.1%	1b	5	10.4%
1b1	9	9.3%	2	6	12.5%
1b2	14	14.4%	2a	2	4.2%
2a	1	1.0%	2b	1	2.1%
2b	59	60.8%	3	1	2.1%
3a	2	2.1 %	3a	13	27.1%
3b	3	3.1%	3b	1	2.1%
4a	4	4.1%	3c1	4	8.3%
4b	1	1.0%	3c	6	12.5%
			3c2	3	6.3%
			4b	2	4.2%
Unknown	1	1.0%	Unknown	2	4.2%
Total	97	100.0%	Total	48	100.0%

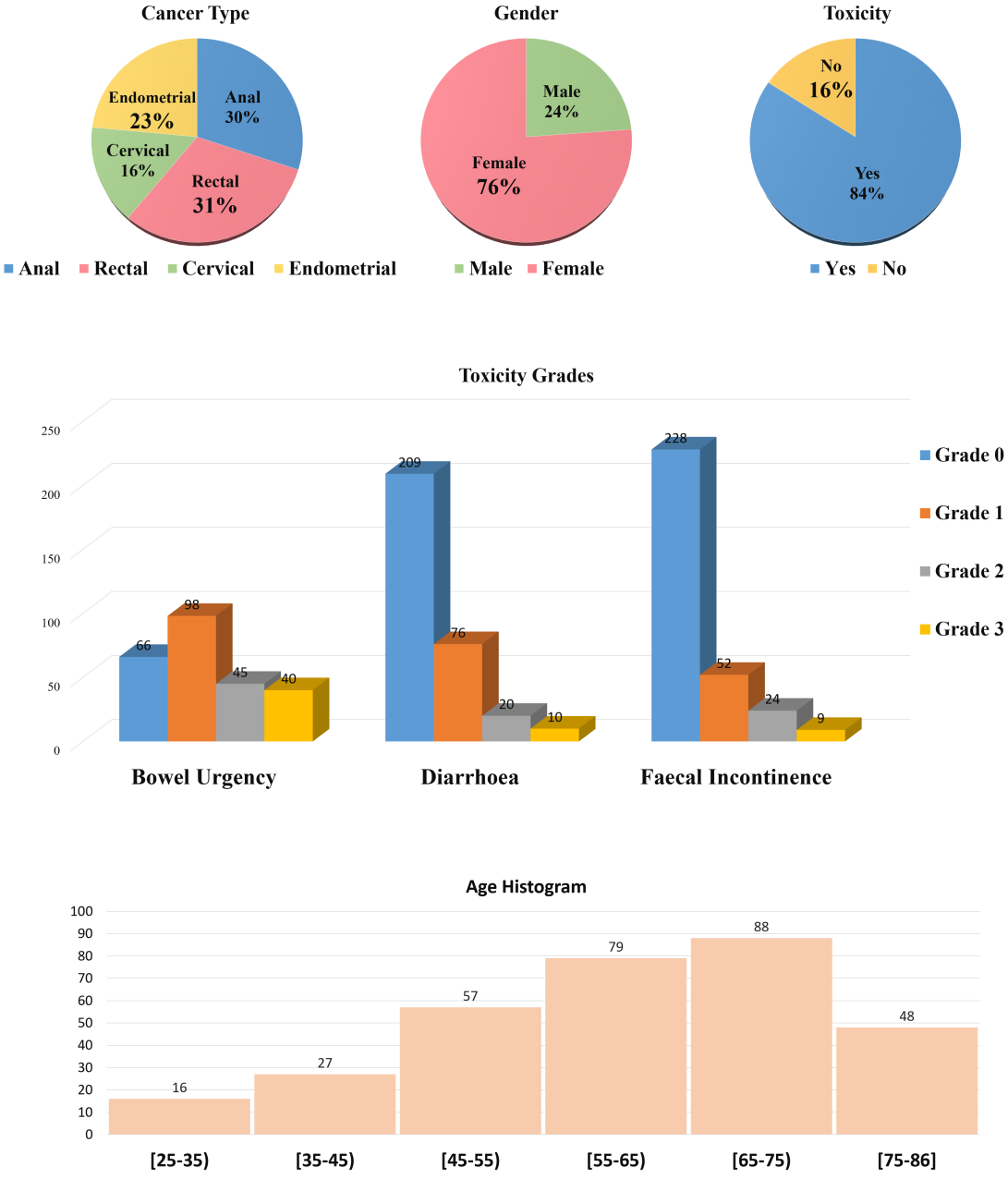
Abbreviations: Num:number of patients. **Note:** FIGO (International Federation of Gynecology and Obstetrics) provides a standardized system for describing the extent and spread of the cancer [153].

Table 3.4: Details of TNM stage for patients in the dataset.

Anal cancer			Rectal cancer		
T stage	Num	Percent	T stage	Num	Percent
1	15	16.0%	1	2	2.7%
2	39	41.5%	2	18	24.3%
3	19	20.2%	3	39	52.7%
4	15	16.0%	4	9	12.2%
X	4	4.3%	X	2	2.7%
Unknown	2	2.2%	Unknown	4	5.4%
Total	94	100%	Total	74	100%
N stage	Num	Percent	N stage	Num	Percent
0	51	54.3%	0	22	29.7%
1	13	13.8%	1	31	41.9%
2	13	13.8%	2	12	16.2%
3	4	4.3%	X	5	6.8%
X	11	11.7%			
Unknown	2	2.1%	Unknown	4	5.4%
Total	94	100.0%	Total	74	100.0%
M stage	Num	Percent	M stage	Num	Percent
0	84	89.4%	0	47	63.5%
1	3	3.2%	1	11	14.9%
X	7	7.4%	X	12	16.2%
			Unknown	4	5.4%
Total	94	100.0%	Total	74	100.0%

Abbreviations: T, local tumour extent; N, nodal involvement; M, metastasis, Num, number of patients.

Figure 3.1: Patients summary statistics included in the dataset. The toxicity chart displays the percentage of patients who have reported experiencing at least one of the toxicities.



3.2 Data analysis plan

The primary question of the data analysis is whether ML can predict those patients that experience RT-induced toxicity by considering factors such as tumour site, treatment features, individual characteristics, CT scan and dose distribution. The strategy for addressing this question involves utilising ML and in particular ANN models to explore possible relationships between various factors and the toxicity.

For the purpose of this research, three most common bowel-related side effects of pelvic radiotherapy were explored: bowel urgency, diarrhoea, and faecal incontinence. These toxicities were selected and graded based on the patient's EORTC questionnaire, as below:

- for bowel urgency: "When you felt the urge to move your bowels, did you have to hurry to get to the toilet?"
- for diarrhoea: "Have you had diarrhoea?"
- for faecal incontinence: "Have you had any leakage of stools from your back passage?" or "Have you had leakage of stools from your stoma bag?" (results combined as per EORTC scoring Manual)

Responses used an ordinal scale with the following categories allocated to the data for this dataset: 0 indicating "not at all," 1 indicating "a little," 2 indicating "quite a bit," and 3 indicating "very much." To classify patients for the purposes of this study, they were divided in two different ways as bellow:

- for chapter 5: patients with grade ≥ 2 for experiencing moderate/severe toxicity and patients with grade < 2 for experiencing no/mild toxicity. This approach is taken in Chapter 5 to identify moderate or severe bowel urgency symptoms. Due to significant imbalances, it was not feasible to use this grading style for diarrhoea and faecal incontinence. However, when categorising bowel urgency, dividing it into these two categories resulted in a balanced dataset that could be effectively learned by the network.

- for chapter 6: patients with grade ≥ 1 for experiencing toxicity and patients with grade < 1 for not experiencing toxicity were classified. The reason for employing this categorisation approach is that Chapter 6 aimed to investigate symptoms of toxicity across all three types of toxicity. This approach of categorisation enabled the network to: (i) be trained effectively, as there were sufficient data available for all three types of toxicity, and (ii) predict toxicity even in cases where symptoms were mild.

For bowel urgency, a total of 66 patients were excluded from the study because their data were irrelevant to the research question (they had a stoma). Additionally, for faecal incontinence, the toxicity grades for two patients were missed, therefore, they were also removed for analysis.

Clinical data for 315 patients were available in a STATA file and were linked to the patients' imaging and dose data with an Excel spreadsheet file. Two patients were excluded from the dataset as there were no study ID available for them (the Excel file consisted of 313 patients).

Imaging and dose data were analysed with an artificial neural network model and numerical data were assessed with three basic machine learning models, including SVM, random forest, and logistic regression, and in combination with imaging and dose data with an artificial neural network. All the models were implemented with the Python Software Foundation; Python Language Reference, version 3.7. Available at Python.org. Deep learning models were developed using PyTorch framework [120], Torch library [119] version 1.5.1 and basic ML models (SVM, RF, LR) were from Scikit-learn Python library version 1.0.2 [121]. All methods were evaluated with the "metrics" package of the Scikit-learn library. The numerical dataset was pre-processed with the Pyreadstat package [49]. Image and dose data were stored in Dicom format, and image preprocessing was performed with the Pydicom Python library version 2.2.2 [109] and SimpleITK toolkit Python version 2.1.1 [173]. 3D volumes were plotted using the 3D Slicer platform [52].

A few risks must be considered regarding the current dataset. Firstly, the dataset is limited in size (patient numbers), which can potentially impact the model's ability to generalize. Sec-

ondly, the majority of cervical cancer patients in the dataset received brachytherapy, but the dose delivered with brachytherapy is not recorded in the external beam dose distribution, and it was consequently not possible to factored in the effect of brachytherapy dose this study. This is due to limitations in available methods to allow for the integration of brachytherapy data into external beam radiotherapy toxicity models. Thirdly the dataset may be unfair for the population of patents as an example the number of female patients are nearly three times more than male patients.

Additionally, the dataset has missing values and it has limited follow-up time for some patients (some patients may develop toxicity after the follow-up time), which are common challenges in medical datasets. To address issues related to generalisation and missing values, ~~I have implemented~~ data augmentation and imputation techniques have been employed. However, it is important to note that certain challenges such as time to event and unfairness (as the data selection process is not completely fair, exemplified by an overrepresentation of female patients with cancer in the dataset) are inherent in the nature of medical datasets and cannot be entirely eliminated. The small size of the dataset also impose a challenge for the model's generalization.Despite these challenges, ~~I will continue to strive for~~the best possible outcomes have been pursued with the available data. Table 3.5 shows the summary of data analysis plan.

3.3 Clinical dataset

Twenty-two clinical features relevant to bowel urgency, diarrhoea and faecal incontinence were selected based on previous research potentially demonstrating a link to radiotherapy and bowel toxicity [26], [31], [37], [12].

Features associated with toxicity were subdivided into five categories; demographic, dosimetric, comorbidity, treatment and medication features. Numerical values for demographic fea-

Table 3.5: Summary of data analysis plan

Patient cohort	Outcome to predict	Input data	OAR	Competing risks
Variable: Bowel urgency				
All patients except Stoma	When you felt the urge to move your bowels, did you have to hurry to get to the toilet?	3D dose distribution, CT scans, 22 clinical features	Bowel bag	small dataset, missing data, unfairness, generalisation
Variable: Diarrhoea				
All patients	Have you had diarrhoea?	3D dose distribution, CT scans, 22 clinical features	Bowel bag	small dataset, missing data, unfairness, generalisation
Variable: Faecal incontinence				
All patients	Have you had any leakage of stools from your back passage or stoma bag?	3D dose distribution, CT scans, 22 clinical features	Bowel bag	small dataset, missing data, unfairness, generalisation

Abbreviations: OAR, organ at risk.

tures corresponded to the patients' age, gender, body mass index (BMI), cancer type and smoking status. Dosimetric factors were estimated from rela differential dose-volume histograms, using relative (rather than absolute) volumes. Seven dosimetric feature corresponded to the relative volume of bowel bag receiving 10, 20, 30, 40, 50, and 60 Gy dose and the total mean dose received. Comorbidity features were defined by three binary values including diabetes, cardiac diseases and previous abdominal surgery. Four treatment features were considered as concurrent chemotherapy, surgery, conformal or Volumetric Modulated Arc Therapy (VMAT) and tumour recurrence [83], [170]. Statins and ACE inhibitors as medication are features that can affect bowel toxicities [165] are also considered in the candidate clinical features. Time since RT was also included in the clinical features to account for varying follow-up times. Table 3.6 summarises the candidate numerical data included in this study.

Table 3.6: Candidate clinical features included in this study for toxicity prediction.

Feature	Mean(std)/Num	Missing data	Description
Diagnosis			
Cancer type	anal:95, rectal:74, endometrial:49, cervical:97	0	value in {1,...,4}
Demographic			
Age	60.48 (13.54)	0	year
Gender	female:240, male:75	0	1 for male 2 for female
BMI	27.62 (5.83)	59/315	-
Current smoker	yes:48, no:205	62/315	binary value in {0,1}
Comorbidities			
Diabetes	yes:27, no:287	1	binary value in {0,1}
Cardiac	yes:103, no:211	1	binary value in {0,1}
Previous Ab surgery	yes:136, no:178	1	binary value in {0,1}
Medication intake			
ACE Inhibitors	yes:38, no:275	2/315	binary value in {0,1}
Statins	yes:54, no:259	2/315	binary value in {0,1}
Treatment			
Total dose	44.91 (10.47)	0	total irradiated in Gy
Concurrent chemo	yes:208, no:107	0	binary value in {0,1}
Received surgery	yes:119, no:196	0	binary value in {0,1}
Received VMAT	yes:19, no:296	0	binary value in {0,1}
Time since RT	2.45 (1.21)	0	years after treatment
Recurrence	yes:47, no:268	0	binary value in {0,1}
Dosimetric			
VBowelBag10Gy	7.24 (6.34)	10/315	% of bowel bag received 10Gy dose
VBowelBag20Gy	11.28(10.34)	10/315	% of bowel bag received 20Gy dose
VBowelBag30Gy	7.28 (6.80)	10/315	% of bowel bag received 30Gy dose
VBowelBag40Gy	16.88 (14.80)	10/315	% of bowel bag received 40Gy dose
VBowelBag50Gy	2.28 (4.16)	10/315	% of bowel bag received 50Gy dose
VBowelBag60Gy	0.54 (3.47)	10/315	% of bowel bag received 60Gy dose

Abbreviations: std, standard deviation; Num, number of patients; BMI, body mass index; Ab, abdominal; ACE, angiotensin-converting enzyme; VMAT, Volumetric modulated arc therapy; VBowelBagXGy, relative bowel bag volume receiving X Gy effective doses.**Note:** all VBowelBagXGy are converted to 2 Gy equivalent dose fractions.

There are two main challenges with the current candidate features. First, for some entries in the dataset data are not complete (see missing data in Tab.3.6). The missing value issue causes practical problems for machine learning models and they require to be identified and replaced. Second, the training dataset for each class label is imbalanced, with one class (in this case, class toxicity) having fewer examples. As a result, it is more difficult for the model to learn the features of the minority class and distinguish them from the majority class. This can lead to biased models, where the model performs well on the majority class but poorly on the minority class. Most classification models are built assuming an equal distribution of classes, which can cause them to overlook the minority class and prioritize learning from the abundant observations instead [65]. This is problematic as the predictions for the minority class are often more important and valuable.

3.3.1 Statistical imputation for missing values

There are two broad categories of imputation methods: univariate imputation and multivariate imputation [102].

Univariate imputation methods fill in missing values using only information from the individual variable that contains the missing values. These methods calculate a statistical value (mean, median, mode, etc.) for the missing data. While univariate imputation methods are simple and easy to implement, they do not take into account any correlations that may exist between variables in the dataset. Additionally, they may not be suitable for all types of datasets. For instance, mean imputation assumes that the data is normally distributed, which may not always be the case. Moreover, using these methods may result in biased estimates and can lead to incorrect conclusions. Therefore, it is essential to carefully consider the nature of the data and the specific requirements of the analysis before deciding on an univariate imputation method.

Multivariate imputation approaches, on the other hand, use information from multiple variables to impute missing values. These methods take into account the correlations between variables

and can produce more accurate imputations. Examples of multivariate methods include regression imputation and k-nearest neighbour imputation [102].

To analyse each imputation approach, both univariate and multivariate methods were tested for the model's performance. Ten missing entries for dosimetric features (VBowelBagXGy) were not imputed as they can not be imputed based on other data in the dataset, and those patients were excluded from the analyses.

For categorical values, including smoking, diabetes, medications, etc., univariate imputation with replacing the most frequent data (see mode in table 3.6) was performed. For the only continuous predictor, BMI, the average of the BMI column was estimated and replaced. Multivariate imputation was also applied and experiments were performed to assess the two imputation methods (all the experiments are described in chapter 4, where toxicity classification with different ML models are reported).

For the multivariate feature imputation, the approach proposed by Van et al.[156] was implemented; each missing feature was replaced in a repetitive procedure; at first, the feature column including the missing values was considered an output y and the rest of the dataset (other feature columns) were treated as input x. A regression model fitted inputs and outputs (only for known y) and it predicted the missing value for each feature. This was repeated until all the missing values were replaced. The prediction performance of the model when using the K-nearest neighbour (KNN) model for imputation was also tested. KNN imputer fills the missing cells using the Euclidean distance. Considering the K-closest features (in terms of Euclidean distance), it uniformly averages their values and replaces the missing sample. Scikit-learn [121] Python library version 1.1.1, was used for all types of statistical imputation.

3.3.2 Data augmentation for imbalanced dataset

To address the problem of imbalanced datasets, data augmentation was utilised; a technique that increases the size of the minority class by creating new synthetic data from the existing data.

Data augmentation is often the primary solution to help balance the dataset and improve the performance of the model [88]. Some common data augmentation techniques for imbalanced datasets include:

- Oversampling: This involves creating new samples for the minority class by replicating existing samples or generating new samples based on the existing samples.
- Undersampling: This involves removing samples from the majority class to balance the dataset. This can be done either randomly or by selecting samples that are similar to the minority class.
- Synthesizing new samples: This technique involves creating new samples for the minority class by applying a function to the current samples. For example, interpolating between existing samples or adding random noise.

In this study, due to the small number of dataset, oversampling and undersampling was not possible. Therefore the synthesising approach was implemented and tested on the model's performance. Further details are provided in Chapter 4.

3.3.3 Data normalisation and feature scaling

The numerical dataset has variable scales, and the range of values is different. For instance, the range for the dosimetric features is in $[0, 75.8]$ and the range for the 'time since radiotherapy' is $[-5.05, -1.01]$. For gradient-based machine learning algorithms, including logistic regression and neural networks, having features on a same scale helps the model converge more efficiently and quickly to the minima; the feature value of x affects the step size of the gradient, and a similar range for x results in smooth movements towards the minima. For distance-based algorithms, including SVM, the feature range drastically impacts the performance of the model. These models use the distance between the feature points to learn the data distribution; for example, in SVM the support vector distances determine the optimum hyperplane for classification. Training the model with a dataset consisting of uneven ranges of features creates a

biased model in which features with higher values gain higher weights [129]. Tree-based algorithms, including random forest, are fairly insensitive to the range of features. For example, in a RF model, the features are split based on their entropy, which is related to the order of the features rather than their value.

Normalisation for continuous and categorical values were preformed separately. For continuous features, the range of datapoint was scaled in [0, 1] with Min/Max normalisation as follow:

$$x = \frac{\mathrm{x} - \min(\mathrm{x})}{\max(\mathrm{x}) - \min(\mathrm{x})}. \tag{3.1}$$

This normalisation keeps the distance and orders of the datapoints and changes their scale. For continuous variables this distance is meaningful while for categorical features distance and order are randomly selected by an encoder and they do not have significant information. For example, the value assigned for gender is 1 for male patients and 2 for female patients and this could be the opposite. Using Min/Max normalisation assigns more value for the category with higher number. To decrease the significance of the distance the coding of categorical numbers was changed. Target encoding and one-hot encoding [114] are two well-known methods that change the coding for categorical values. One-hot encoding creates an sparse matrix and inflates the categorical feature; in case of gender, instead of one column, two columns are considered as gender where they belong to male and female feature separately. For example for a female patient, this new encoding assigns zero/one for male/female column. Although it vanishes the importance of the feature, for numerical value with too many categories it increases the dimensionality - particularly when most of the categories are rare and useless for prediction.

Target encoding encodes the categories with respect to the impact they might have on the target (label for classification). For a binary classifier, it computes the posterior probability of target = 1, given the input x is the category c_i, or $p(t = 1|x = c_i)$. Since the output falls within the range of zero to one, it does not require any further normalisation. Table 3.7 illustrates the

range of numerical data before and after normalisation.

Table 3.7: Target category encoding results for categorical features.

Feature	Value	New Value BU	New Value DI	New Value FI
Gender	{1,2}	{0.33,0.38}	{0.02,0.11}	{0.08,0.11}
Cancer type	{1,..,4}	{0.49,0.37,0.15,0.29}	{0.23,0.12,0.08,0.11}	{0.26,0.12,0.16,0.22}
Current smoker	{0,1}	{0.33,0.35}	{0.18,0.07}	{0.12,0.10}
Diabetes	{0,1}	{0.32,0.50}	{0.09,0.11}	{0.10,0.07}
Cardiac	{0,1}	{0.32,0.37}	{0.10,0.07}	{0.13,0.03}
Previous Ab surgery	{0,1}	{0.31,0.37}	{0.07,0.12}	{0.09,0.11}
ACE inhibitors	{0,1}	{0.33,0.32}	{0.10,0.03}	{0.12,0.01}
Statins	{0,1}	{0.32,0.33}	{0.09,0.07}	{0.11,0.02}
Concurrent chemo	{0,1}	{0.22,0.39}	{0.06,0.11}	{0.03,0.14}
Received surgery	{0,1}	{0.40,0.19}	{0.06,0.11}	{0.13,0.05}
Received VMAT	{0,1}	{0.36,0.05}	{0.10,0.05}	{0.10,0.05}
Recurrence	{0,1}	{0.35,0.25}	{0.10,0.04}	{0.10,0.08}

Abbreviations: Ab, abdominal; ACE, angiotensin-converting enzyme; VMAT, Volumetric modulated arc therapy; BU, bowel urgency; DI, diarrhoea, FI, faecal incontinence.

3.4 Imaging and dose data

Imaging and dose data for 313 patients were available in the dataset. For each patient, three types of data were collected: CT scans, dose distributions (RD files), and structure sets (RS files) were all collected and stored in DICOM format. CT files provide 3D anatomical information, including the bones, blood vessels and soft tissues inside the patient's body. RD files include dose data (dose images) and dose metadata information. Structure set contours (organ segmentation) were included in RS files. The contours with the bowel bag structure were extracted as it is the organ at risk available for all patients and associated with bowel toxicities. Although the anorectum is an organ at risk, as this is target organ for treatment for patients with anal and rectal cancer, this was not used as a training contour for the dataset. Fig. 3.2 illustrates a CT scan, dose, and overlapped mask image for one example patient in the dataset.

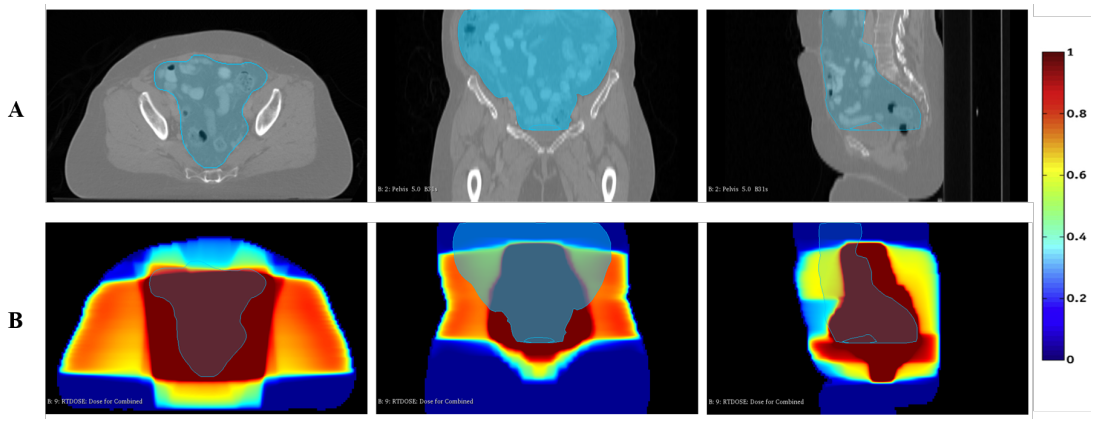


Figure 3.2: An example of patient data in the dataset. A shows the CT scan and B shows the dose distribution plan. The bowel bag contour extracted from RS file (mask image) is overlapped with CT and dose images. From left to right: axial, coronal and sagittal views. Higher values for dose image demonstrates higher dose irradiated. The scale for dose data is not in units of Gy (Gray). To provide a more visually informative representation, the dose data is plotted using a heat colour map.

Original image data are usually not ready for training machine learning models and they require data preprocessing to be ready for inference and training. Appropriately transforming and scaling the entire dataset, as the preprocessing step, facilitates the training/testing procedure of the model and improves its learning performance. In order to make the dataset ready for neural network input, various various data pre-processing steps were performed.

3.4.1 Dose distribution correction

There were two different sets of dose treatment data in the dataset; for the first group, including data of 168 patients, multiple dose distributions per patient were available, each representing a single radiation beam. For the second cohort (145 patients), only one dose distribution per patient was stored. For the first cohort, the beams are delivered from 2–7 different directions as part of each treatment (see Fig. 3.3).

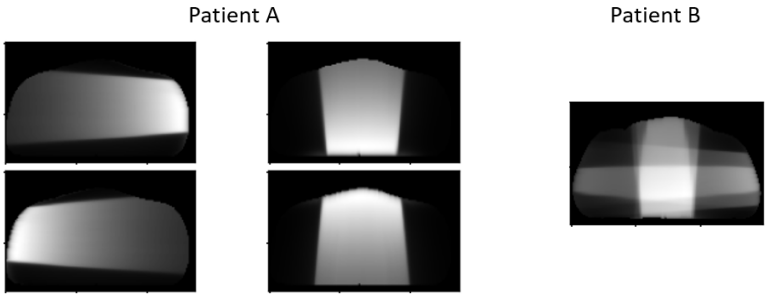


Figure 3.3: Examples of different dose data available in the dataset. Patient A, who received dose treatment from four different angles, and patient B has only one treatment file. The dose files for patient A need to be combined as one dose file.

To have a representation of the full treatment, combining of the beam dose distributions from different directions into a single dose distribution was needed. Because all of the DICOM files used the same coordinate system, a sum of all of the different beams can present the final dose distribution. For radiobiological effect correction (i.e. taking into account fraction-size effects), for each voxel in the dose distribution D, the dose was recalculated to the equivalent dose in 2 Gy fractions (EQD2; [13]) by:

$$\mathrm{EQD2} \;=\; \mathrm{nd}\;\frac{(d+\alpha/\beta)}{(2+\alpha/\beta)}, \tag{3.2}$$

where n is the number of treatment fractions for patient's RT treatment, $d = D/n$ is the dose per fraction, and α/β is a constant controlling the fraction-size sensitivity that is set to 3 Gy [12]. Fig. 3.4 shows a dose re-scaling operation. For patients treated with multiphase treatments (e.g. larger fields followed by cone down), each phase was rescaled to EQD2 prior to calculating the total dose across phases.

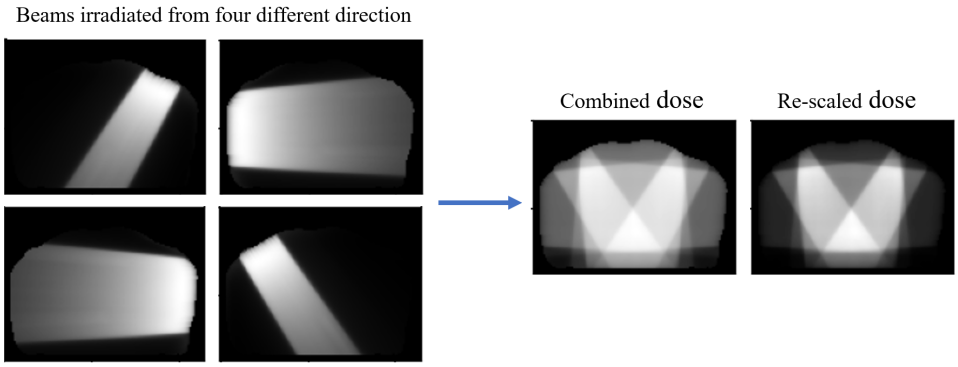


Figure 3.4: Example of dose correction for a patient with four dose treatment files. The combined dose is the result of summation of all the four files and the rescaled dose is computed with the combined dose and Equation 3.2.

3.4.2 Image registration and re-sampling

Patient position in radiotherapy is determined based on the position that yields the minimum irradiated dose to the organs at risk. Patients with prone and supine positions were available in the datasets. Due to that, CT scans and their corresponding dose distributions had different dimensions and orientations for patients with different positions (See Fig.3.5, original data). To integrate all patients' data, dose and contours data were re-sampled to their corresponding CTs, and to improve registration time and accuracy, all the CT scans resampled to be the same orientation as the prone position. Then, CT, dose and contour images were spatially re-sampled with linear interpolation to voxel sizes of 0.97mm× 0.97mm and thickness of 5mm with the SimpleITK toolkit [173] in Python. Fig. 3.5 illustrates the result of re-sampling and CT corrections.

Moreover, the number of CT slices for each patient in the dataset varied from 49 to 224 with an average of 98 (See Fig.3.6). In order to ensure that all the images used for training had a consistent orientation, scale, and position, CT scans of all the patients were transformed into a reference patient. This could help to eliminate variability between images and made it easier for the neural network to learn the features that were important for the toxicity prediction.

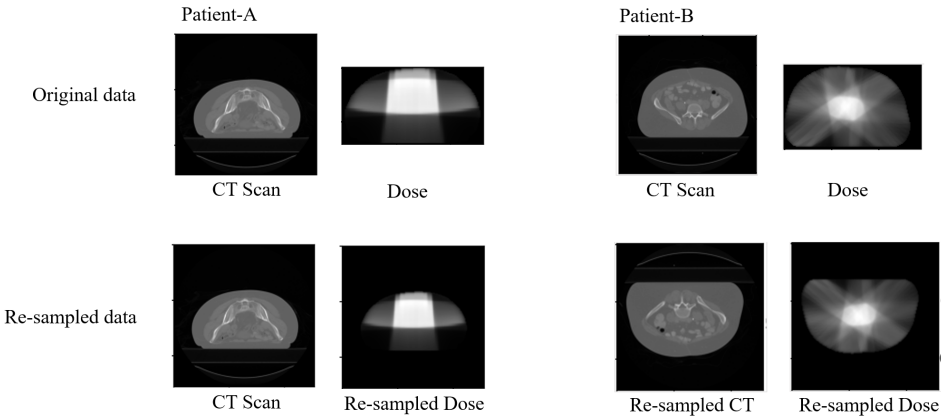


Figure 3.5: Example of image matching and re-sampling for two patients in different positions. The top row shows CT and dose distributions for patient A in the prone position and patient B in the supine position, with different dose orientations for the two patients. The bottom row presents the outcomes of the re-sampling and CT orientation correction. Specifically, the CT scan for patient B was re-oriented to match the orientation of patient A.

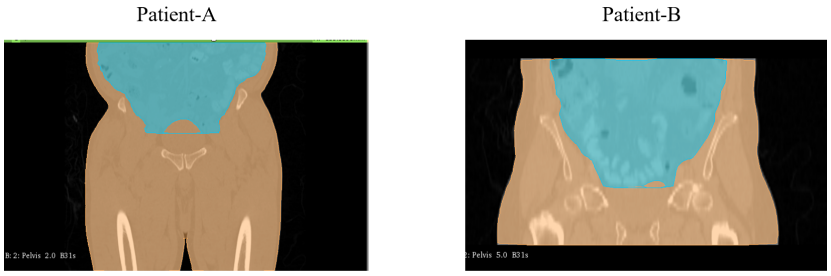


Figure 3.6: The coronal view of CT scans for two patients, A and B, with the maximum and minimum number of slices, respectively. The bowel bag structure (region of interest for bowel-related toxicity prediction) is shown in blue.

Image registration or image alignment involves spatially transforming the source images to align with a target image. Prior to registration, all CT slices containing bowel bag structure were extracted for each patients, and then these volumes were rigidly registered to a reference image. The patient with the least number of slides (35) for bowel bag was selected as the reference patient. The height and width of CT scans were the same for all the patients ([512,512]). The registration was performed using the SimpleITK [173] toolkit in Python. After registra-

tion, the CT data for all the patients had a dimension of [35,512,512] voxels. Dose and mask images were also registered using the computed transformation for their corresponding CT. Two examples of CT registration are shown in Fig.3.7.

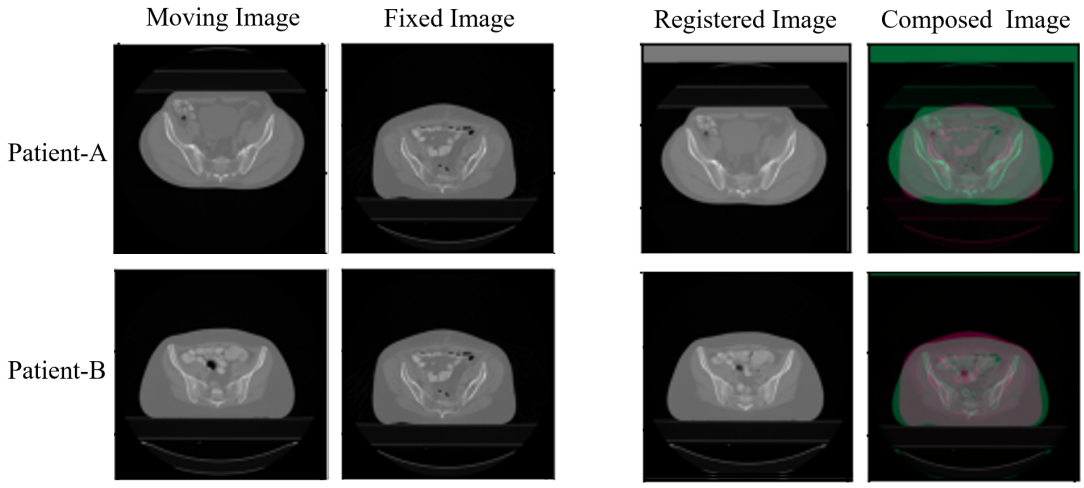


Figure 3.7: Example of CT registration. Composed image illustrate how moving image (green shade) is registered to the fixed image (red shade).

3.4.3 Region of interest masking

To train the deep learning model for toxicity prediction in the bowel bag region, irrelevant information outside of this area should be removed from the input data. This is because the model only requires information related to the area of interest. Therefore, the bowel bag contours were extracted as mask images; a mask is a tensor containing binary values of either one or zero based on the volume's shape. Specifically, the region of interest, which was the bowel bag, was assigned a value of one in the mask, while the rest of the area in the mask was set to zero. All re-sampling and registration were also applied on the mask image. Then this mask was superimposed to its corresponding CT and dose images (in order to remove non-essential information); the element-wise multiplication of the CT/dose volumes with the corresponding mask was performed as the third step of image pre-processing. Fig. 3.8 illustrates the results

of masking.

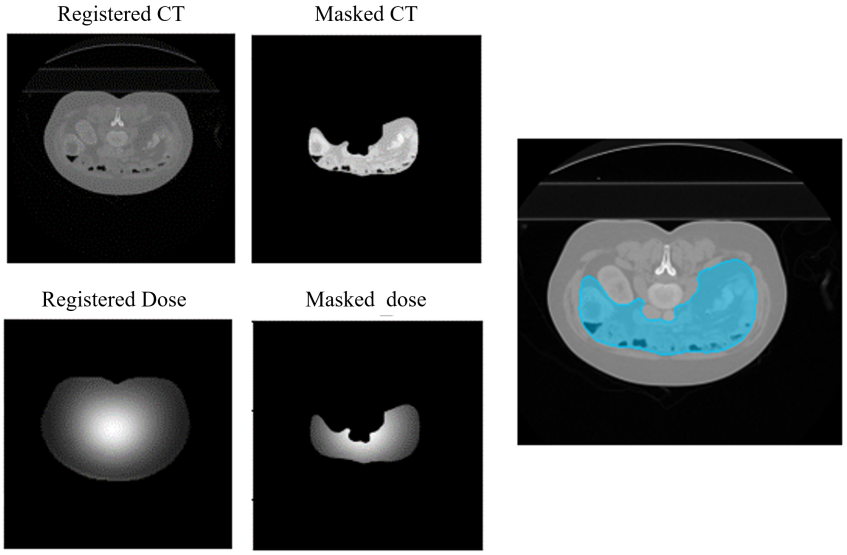


Figure 3.8: Example of region of interest masking. The blue contour in the right CT image is the bowl bag extracted form structure file. The result of the CT and dose masking are shown as “Masked CT” and “Masked dose”, respectively.

3.4.4 Data normalisation

The final step of image data pre-processing was normalisation . Data normalisation can improve neural network convergence time and accuracy [91]. In classification networks, the output is a number in the range [0, 1], therefore all image data was normalised to be in the range [0, 1]. The final CT and dose images were re-scaled based on the whole population of the dataset. Considering X as the normalised value for x, Min-Max normalisation was applied with:

$$X = \frac{x - \text{Min}(D)}{\text{Max}(D) - \text{Min}(D)}, \tag{3.3}$$

where Min(), Max() are functions that return the minimum and maximum of the whole dataset D, respectively.

~~I did not crop~~ The non-zero area of the input image were not removed because ~~my~~ the proposed model extracts the small patches in the image if they contain non-zero values (explained in the chapter 5), therefore the zero patches (patches without information) will be automatically removed.

3.5 Other approaches and limitations

It is important to acknowledge that there exist alternative approaches to data preprocessing. For instance, in the case of normalisation, techniques such as Z-score normalisation can be utilised. Similarly, for registration, non-rigid transformations can provide alternative options. Additionally, various methods of imputation and augmentation are available to address missing values and enhance the dataset, respectively. However, it is crucial to carefully consider the nature of the dataset, including its size, data distribution, and other relevant factors, when selecting the most appropriate preprocessing approaches. Adapting the preprocessing techniques to align with the specific characteristics of the dataset is key to obtaining optimal results. In this chapter, the aim was to meticulously choose the preprocessing techniques that yield the highest prediction performance, taking into account the nature of the dataset and the models employed.

The approaches that have been employed for data preprocessing were not without limitations. Rigid registration, for instance, can be affected by the quality of the images, and not all images in the dataset possessed optimal quality. As for data normalisation, one limitation is the potential loss of information. By basing the normalisation on the entire population, data with smaller values (specifically in dose distribution data) relative to the overall population may be overshadowed or lost entirely. Furthermore, normalisation can impact rare data points, as their normalised values may not accurately represent their significance or differentiate them from more common events.

Concerning data augmentation, one major limitation in this research was the small size of

the dataset, which impacted the effectiveness of augmentation. While augmenting a small dataset can increase its size, it may not necessarily introduce new information. However, given the circumstances, the available options were utlised effectively to generate an appropriate amount of new data without encountering the issue of overfitting.]. Furthermore, the method of augmentation must be carefully considered based on the dataset characteristics. for the aim of this research the shape of the bowel bag should be reserved which, restricted options like rotation or scaling for data augmentation.

Imputation also presents its own set of challenges, particularly regarding the potential impact of new imputed data on existing relationships and correlations within the dataset. Despite these challenges, the most suitable imputation approach was employed to address the available gaps. It is important to note that these limitations and challenges within the data preprocessing methods are present. Nevertheless, the best approaches possible was utlised, given the circumstances, to mitigate these limitations and achieve optimal results.

3.6 Summary and conclusions

In this chapter, the dataset utilised in the study was introduced, consisting of two types of data: 3D CT imaging, dose distributions, and contours, as well as 1D clinical data. Each data type required different pre-processing steps. For the 3D CT, dose, and contour volumes, registration, re-sampling, and normalisation were applied. Additionally, fraction-size correction was performed for the dose data. Regarding the clinical dataset, several pre-processing approaches, including imputation for missing values, augmentation, and normalisation were needed. The specific details of these pre-processing procedures were elaborated upon in this chapter.

The next step is exploring the ability of machine learning to predict toxicity. Here an important question arises: Can machine learning approaches effectively predict toxicity? In the forthcoming chapter, the findings on how different ML models can predict toxicity based on the available dataset are presented. The aim is to assess the predictive capabilities of these mod-

els and determine if they can contribute to detecting clinical risk factors. By conducting this investigation, the thesis aims to shed light on the potential of machine learning techniques in toxicity prediction and their potential impact on enhancing accuracy and overall outcomes.

Chapter 4

Conventional Machine Learning Models for Toxicity Prediction

In this chapter, three conventional machine learning models (LR, SVM, and RF; as the most common ML models for classification in medical tasks 2.2.1) are utilised to analyse clinical metadata and predict three bowel-related toxicities. Additionally, the importance of each clinical feature is assessed, and various methods of imputation for missing values in the clinical dataset are tested. The main goal of this chapter is to demonstrate the potential of clinical metadata alone in predicting RT-induced toxicity, given their frequent usage in current toxicity prediction models. The outcomes of this chapter are needed in the following chapters to compare the impact of image data (CT and dose) and their combination on toxicity prediction.

4.1 Introduction

In order to investigate the correlation between clinical data and different bowel toxicities, 22 pre-selected clinical features (see Chapter 3) with three ML techniques: LR, SVM and RF for the three different types of patient-reported bowel toxicity (bowel urgency, diarrhoea and

faecal incontinence). The main objective is to predict grade≥ 1 toxicity with pre-selected clinical features.

4.2 Implementation details

The choice of the best kernel for svm depends on the nature of the data and the problem. The linear kernel is suitable for linearly separable data, and due to being computationally efficient, it is worth to explore whether this method is effective for the dataset. The model was created using Scikit-learn Python library and the C value was set to 1 (C parameter is a regulator term that controls the trade-off between the training error and the margin. A larger value of C implies a smaller margin and a higher penalty for misclassification, while a smaller value of C implies a larger margin and a lower penalty for misclassification). Exploring a curved decision boundary, the SVM with polynomial kernel was trained with C=1 and a degree of 5 (this implies that the decision boundary created by the SVM is a polynomial function of degree 5 in the input space). Finally SVM with the RBF kernel was trained with the C=1 and gamma=0.05 (the gamma determines the influence of each training example on the decision boundary. A higher value of gamma indicates that nearby training examples have a significant impact on the decision boundary, while a lower value of gamma means that training examples farther away can also have an influence).

The RF model consisted of 100 randomly created trees. The "max_features" hyperparameter was set to the length of the training columns (max_features shows the number of features randomly considered at each split when building the individual trees within the forest).

The LR model was trained with the default parameter of the Scikit-learn library except that the C values was selected 0.01. Due to the imbalanced dataset, data augmentation was applied to the training set and ML models were tested for both modes with and without data augmentation.

The class weight property for all methods was set to "imbalanced" in model training. All the hyperparameters were selected based on a exhaustive search over specified parameter using the GridSearchCV function of "model_selection" class of Sikit-learn library.

4.3 Predicting bowel urgency with machine learning techniques

Seventy-five patients were excluded from the dataset due to missing data for dosimetric features or a label for bowel urgency. From 240 patients, 20 patients with and 20 without bowel urgency toxicity were randomly selected for the test set. The reminder (200 patients) were used for training, in which 155 patients had toxicity and 45 did not. Data augmentation was only applied to the training set.

To determine the most effective imputation method, an LR model using different imputation techniques was trained on the training set. Among continuous features only BMI was imputed and the missing entries for dosimetric features (VBowelBagXGy) were not imputed (as dose data are patients specific, imputation can potentially lead to inaccurate or false information). Consequently, patients with missing dosimetric features were excluded from the study. For categorical features, the mean, mode (most frequent data) and median of the missing variable distribution were replaced with univariate imputation. Multivariate imputation and KNN were also tested for both continuous and categorical variables.

The experiment results are presented in Table 4.1. KNN yielded the highest accuracy and AUC values among all the imputation methods. Univariate and multivariate imputation produced similar results and notably, the performance was poor without any imputation. Having a better performance, for all the experiments in the study, the clinical dataset was imputed using a KNN approach.

Table 4.1: Comparison of the different imputation methods based on the LR classifier. Best performance in each metric is shown in bold.

Imputation	Accuracy	AUC	Sensitivity	Specificity	Comment
-	0.50	0.52	0.45	0.55	Patients with missing entries were removed from the training dataset.
Univariate-1	0.57	0.60	0.45	0.70	BMI is imputated with "mean" and categorical are imputed with "mode" strategies.
Univariate-2	**0.58**	0.57	0.40	**0.75**	BMI is imputated with "mean" and categorical are imputed with "median" strategies.
Multivariate	0.55	0.60	0.40	0.70	BMI and categorical are computed with multivariate imputation.
KNN	**0.58**	**0.64**	**0.55**	0.60	BMI and categorical are computed with KNN imputation.

Abbreviations: AUC, area under the receiver operating characteristic curve; BMI, body mass index; KNN, k nearest neighbour; LR, logistic regression.

An LR classifier with various data augmentation techniques was trained to investigate how data augmentation affects model performance. Synthetic Minority Oversampling Technique (SMOTE) [22] is a synthesising technique which generates synthetic data for each sample of the minority class based on its k-nearest neighbours; for all the samples, the first k neighbours are identified, and then between each pair of points and neighbours, a new synthetic data is interpolated.

Another approach is Adaptive Synthetic (ADASYN) sampling, introduced by Haibo He et al. [66] for imbalanced learning. The technique is an improved version of SMOTE that finds the first k neighbours of the "harder-to-learn" examples in the minority class. Then after creating those samples, it adds small random values to them, thus making them more realistic. In other words, instead of all the synthesised samples being linearly correlated to the real samples, they have more variance, i.e., they are more scattered. It is important to note that all augmentation and imputation techniques were applied after target encoding and normalisation processes. As a result, adding small values to categorical variables is permissible, as after normalisation, all variable values (both categorical and continuous) are scaled between 0 and 1.

Table 4.2: Comparison of the different augmentation methods based on the LR classifier. The best performance in each metric is shown in bold.

Augmentation	Accuracy	AUC	Precision	F-1	Sensitivity	Specificity
No augmentation	0.50	0.60	0.50	0.49	0.45	0.55
SMOTE	0.67	0.65	0.70	0.64	0.60	0.75
ADASYN	**0.75**	**0.77**	**0.81**	**0.72**	**0.65**	**0.85**

Abbreviations: AUC, area under the receiver operating characteristics curve.

The two augmentation approaches were separately applied to the dataset and the model performance was evaluated for each method. Table 4.2 shows that the performance significantly improved when the model was trained on the augmented dataset. The model's accuracy was calculated at 0.50 while this increased to 0.75 when it was trained on the dataset with augmentation. The other performance metrics also improved. Comparing SMOTE and ADASYN, the latter performs superiorly; since the dataset is highly imbalanced, it is essential to focus on generating samples for those minority class examples that are harder to learn. Synthesizing new samples by interpolating between all the examples, can result in overfitting and the production of redundant samples. Therefore, ADASYN performs better than SMOTE by not only increasing the number of minority class examples but also enhancing the diversity of the synthetic samples, which is crucial for improving the classification performance for imbalanced datasets.The ADASYN technique for oversampling was used for the rest of the experiments.

The three classifiers were trained on the augmented dataset and tested on the unseen test set.

Table 4.3 illustrates the results of the experiments.

Table 4.3: Comparison of the different machine learning methods for bowel urgency prediction. The best performance in each metric is shown in bold.

Method	Parameter	Accuracy	AUC	Precision	F-1	Sensitivity	Specificity
LR	-	**0.75**	**0.77**	**0.81**	**0.72**	**0.65**	**0.85**
SVM	Polynomial kernel	0.70	0.70	0.75	0.66	0.60	0.80
SVM	Linear kernel	0.65	0.72	0.72	0.68	**0.65**	0.65
SVM	RBF kernel	0.55	0.65	0.58	0.61	0.55	0.55
RF	100 trees	0.65	0.60	0.75	0.56	0.45	**0.85**

Abbreviations: AUC, area under the receiver operating characteristics curve; LR, logistic regression; SVM, support vector machine; RF, random forest; RBF, radial basis function.

LR had superior performance than other models in all the evaluation metrics. The worst performance was achieved by random forest; having a high specificity (0.85) and a low sensitivity (0.45) shows that random forest is biased towards predicting label zero (the class without toxicity).

To find the importance of each predictor feature, (i) the magnitude of the coefficients in the logistic regression model, (ii) SVM weights and (iii) the Gini importance computed by the random forest model were extracted. Fig 4.1 shows the results of the experiment. Considering all models, "BMI" and "cancer type" are among the top five important features; this implies that both features can affect bowel urgency toxicity. Additionally, the dosimetric features (VBowelBagXGy) consistently scored highly across all models, suggesting a strong association between the dose irradiated to the bowel bag and the risk of toxicity.

However, there was some overlap among the top fifteen features, including "time since RT", "current smoker", "cardiac", and "diabetes". In contrast, at the end of the spectrum, "ACE inhibitors" and "statins" gained the lowest weights.

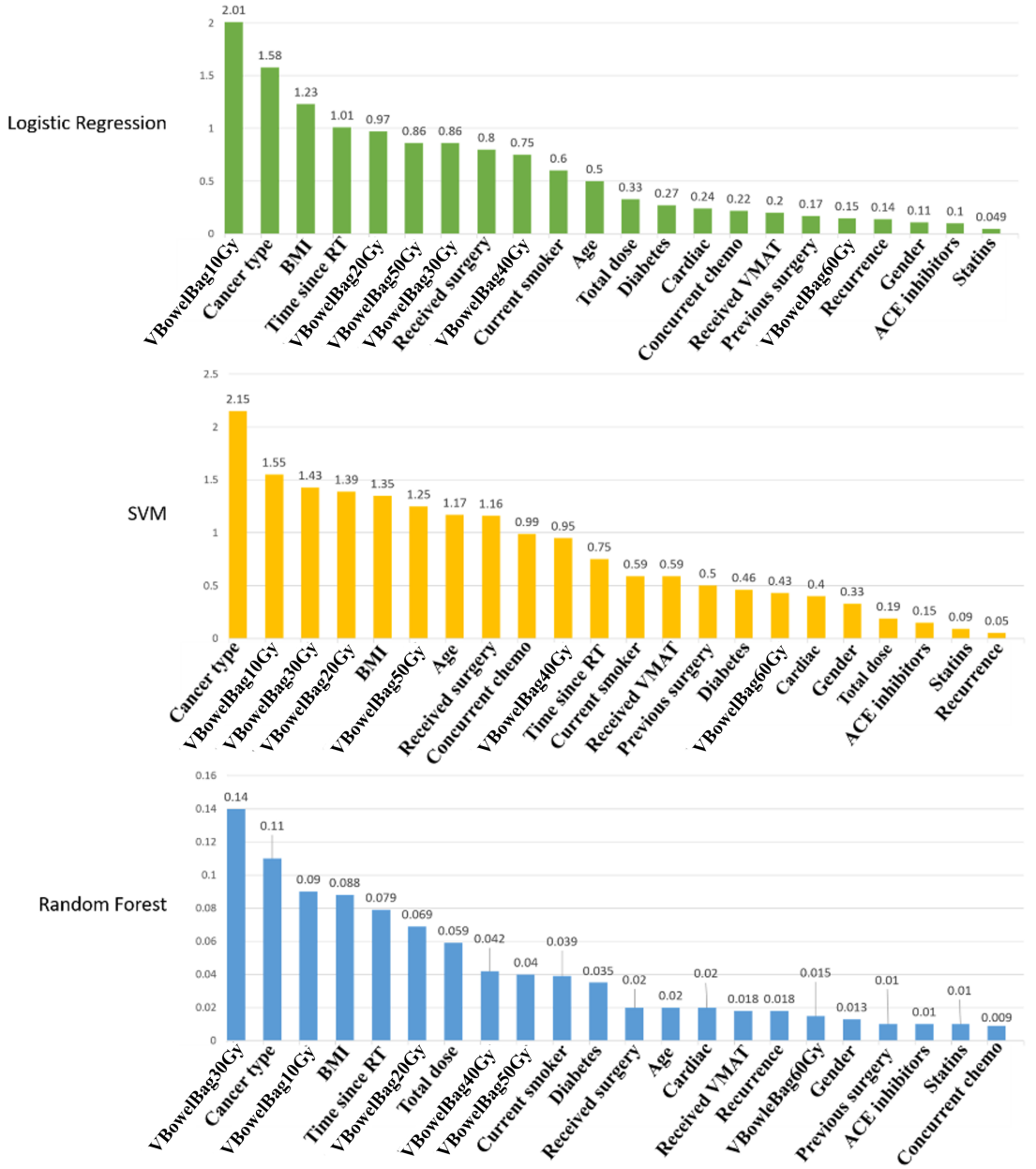


Figure 4.1: Analysis of risk factors for bowel urgency. The importance of the features for logistic regression, SVM, and random forest models are extracted. The x axis presents the significance of features; for logistic regression the coefficients of the model, for SVM the models weights and for random forest the Gini score of each feature are shown. **Abbreviations:** BMI, body mass index; VMAT, Volumetric modulated arc therapy; RT, radiotherapy; ACE, angiotensin-converting enzyme. **Note:** Total dose denotes the total prescribed dose.

4.4 Predicting diarrhoea with machine learning techniques

For diarrhoea prediction, 10 patients were excluded due to missing values for VBowelBagXGy. Twenty patients with diarrhoea and twenty without were randomly selected for the test set. The remainder (265 patients) was left aside for the training set including 184 patients without and 81 patients with symptomatic diarrhoea. ADASYN data augmentation was applied to the minority class, and the new training set included 346 patients. Table 4.4 shows the results of the experiment. Like the results for bowel urgency, LR performed slightly better than other models, and the RF model had the worst performance. The low sensitivity score of 0.25 for RF shows that the model did not learn the minority class distribution and is highly biased.

Table 4.4: Comparison of the different machine learning methods for diarrhoea prediction. The best performance in each metric is shown in bold.

Method	Comment	Accuracy	AUC	Precision	F-1	Sensitivity	Specificity
LR	-	**0.70**	**0.65**	**0.68**	0.61	0.65	**0.75**
SVM	Polynomial kernel	0.60	0.61	0.63	**0.64**	**0.70**	0.60
SVM	Linear kernel	0.65	0.63	0.54	0.54	0.60	0.50
SVM	RBF kernel	**0.70**	0.63	0.58	0.59	**0.70**	0.50
RF	100 trees	0.50	0.60	0.50	0.33	0.25	**0.75**

Abbreviations: AUC, area under the receiver operating characteristics curve; LR, logistic regression; SVM, support vector machine; RF, random forest; RBF, radial basis function.

To understand the influence of each predictor feature, the importance of each feature was extracted from all models similar to the previous section. Fig.4.2 illustrates the significance of the clinical variables.

Fig.4.2 shows that the dosimetric features representing the portion of the bowel bag receiving higher dose are associated with toxicity. The total dose is also located in the top ten features for all models. Non-dosimetric features, including “cancer type”, “BMI” and “received VMAT” are also correlated with risk of toxicity, as they are selected in the top 15 features by all the models.

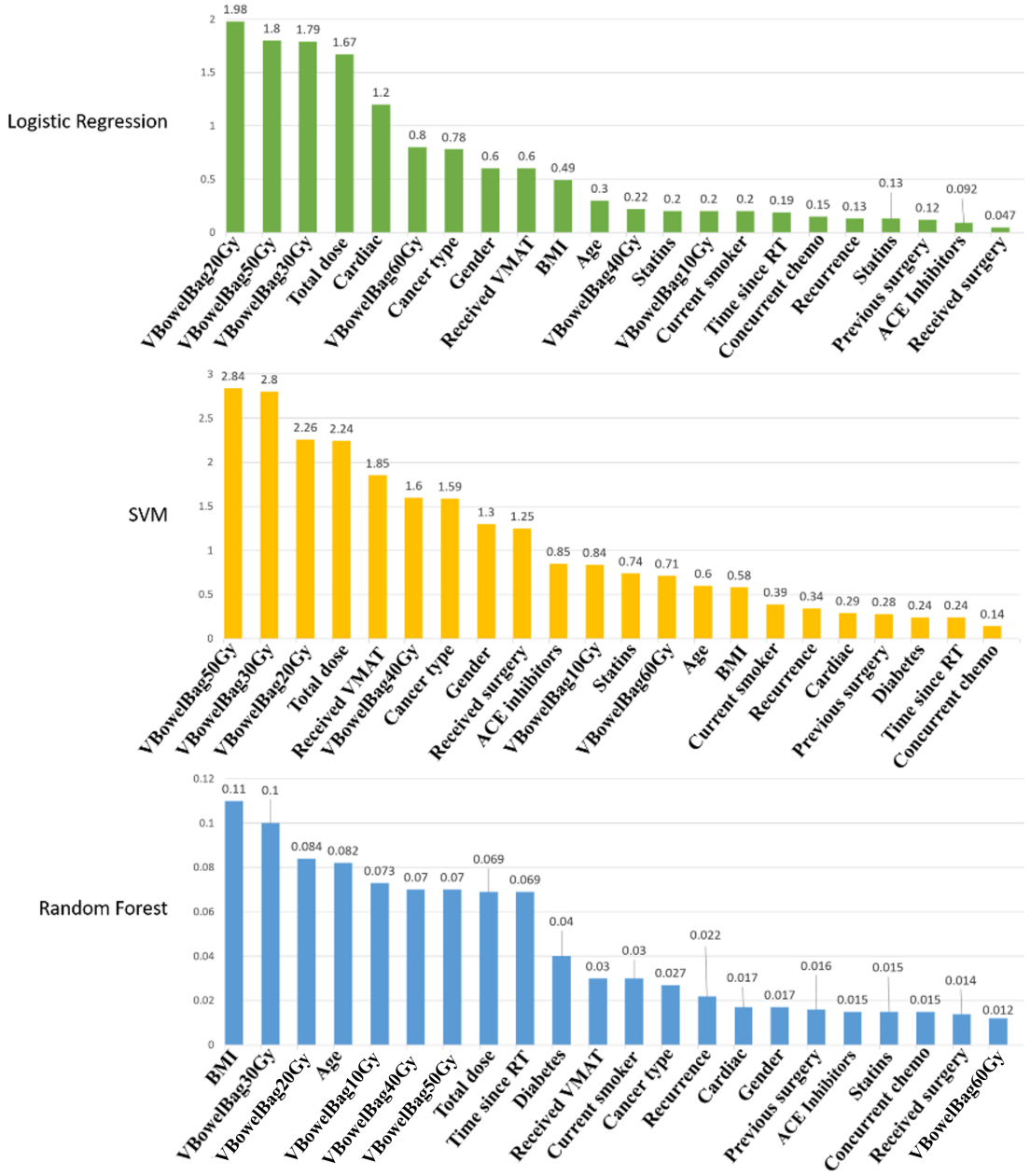


Figure 4.2: Analysis of risk factors for diarrhoea. The importance of the features for logistic regression, SVM, and random forest model are extracted. The x axis presents the importance of features; for logistic regression the coefficients of the model, for SVM the models weights and for random forest the Gini score of each feature are shown. **Abbreviations:** BMI, body mass index; VMAT, Volumetric modulated arc therapy; RT, radiotherapy; ACE, angiotensin-converting enzyme. **Note:** Total dose denotes the total prescribed dose.

4.5 Predicting faecal incontinence with machine learning techniques

Twelve patients were removed from the dataset due to missing data, for dosimetric features or the faecal incontinence label. 20 patients with and 20 without symptomatic faecal incontinence were randomly chosen for the test set. Among the 263 patients remaining for the training set, 199 of them reported no symptoms of faecal incontinence toxicity. Due to the imbalanced training set, ADASYN data augmentation was applied to the minority class (64 patients). Therefore all ML models were trained on the augmented dataset with 327 data. The results of the experiment are shown in Table 4.5. Similar to the prediction of diarrhoea and bowel urgency, logistic regression and RF models received the highest and lowest scores for faecal incontinence prediction; LR had an accuracy of 0.70 while RF had an accuracy of 0.55.

Table 4.5: Comparison of the different machine learning methods for faecal incontinence prediction. Best performance in each metric is shown in bold.

Method	Comment	Accuracy	AUC	Precision	F-1	Sensitivity	Specificity
LR	-	**0.70**	**0.71**	**0.78**	**0.64**	**0.55**	**0.85**
SVM	Polynomial kernel	0.65	0.63	0.68	0.61	**0.55**	0.75
SVM	Linear Kernel	0.60	0.65	0.66	0.50	0.40	0.80
SVM	RBF kernel	0.60	0.64	0.62	0.55	0.50	0.70
RF	100 trees	0.55	0.60	0.62	0.35	0.25	**0.85**

Abbreviations: AUC, area under the receiver operating characteristics curve; LR, logistic regression; SVM, support vector machine; RF, random forest; RBF, radial basis function.

The coefficients of the LR and SVM models and the Gini score for RF were extracted to explore the importance of each clinical factor (see Fig4.3). For all models, "BMI" is among the top three features. Furthermore, being among the top ten features, dosimetric variables (VBowelBagXGy) are also highly correlated with the faecal incontinence toxicity. Other non-dosimetric features, including "cancer type", "time since RT", and "ACE Inhibitors" are also highlighted by all models within the top 15 features.

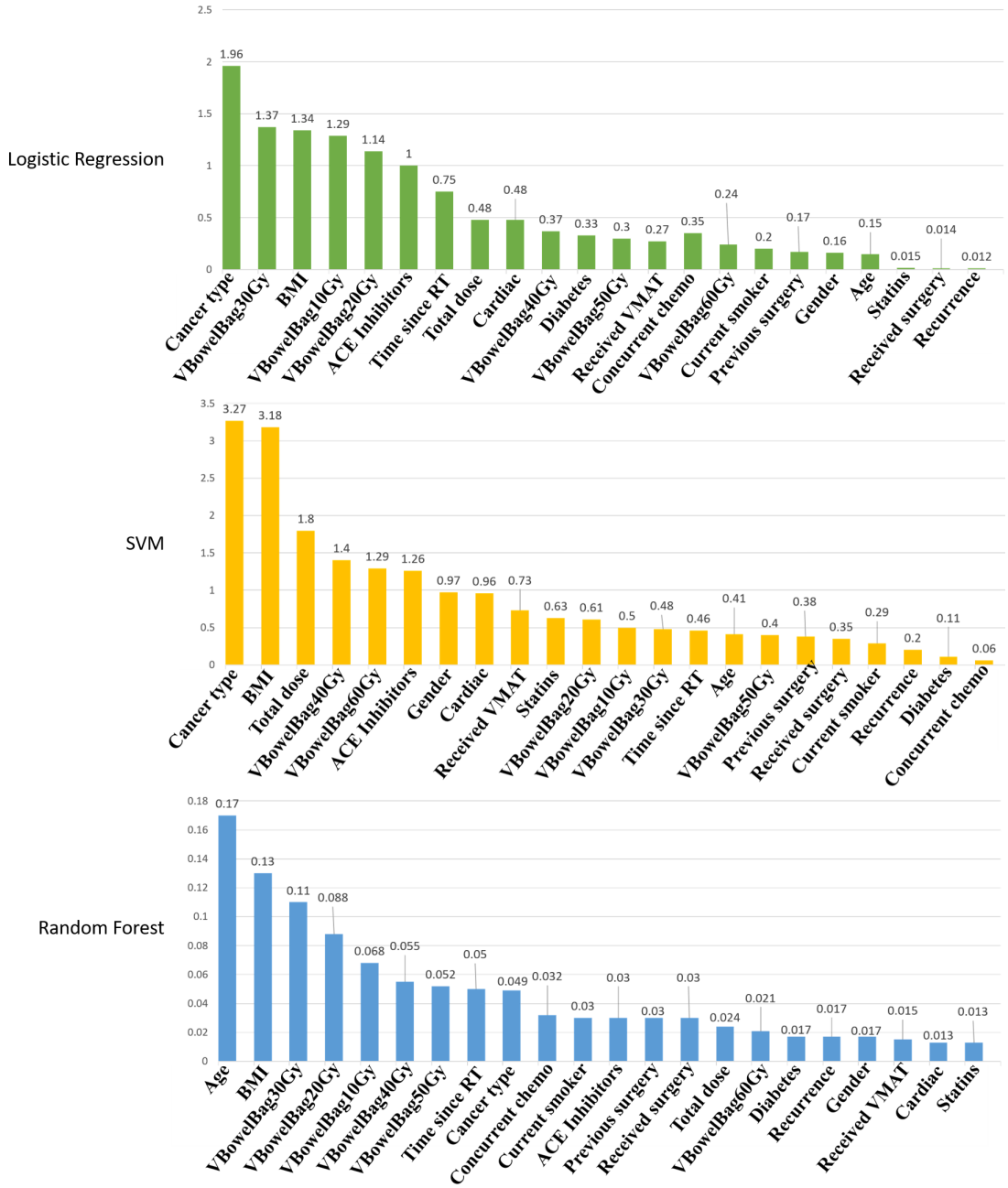


Figure 4.3: Analysis of risk factors for faecal incontinence. The importance of the features for logistic regression, SVM, and random forest model are extracted. The x axis presents the importance of features; for logistic regression the coefficients of the model, for SVM the models weights and for random forest the Gini score of each feature are shown. **Abbreviations:** BMI, body mass index; VMAT, Volumetric modulated arc therapy; RT, radiotherapy; ACE, angiotensin-converting enzyme. **Note:** Total dose denotes the total prescribed dose.

4.6 Discussion

In a general view of all the models, logistic regression performed slightly better than SVM and RF. This may be because the dataset is relatively small, and although ADASYN data augmentation was used to balance the dataset and generate the hard-to-learn features, it is insufficient. This indicates that data points in the decision boundary may not accurately reflect the true decision boundary. As a result, it can create a false maximum margin boundary in SVM and a false decision in the random forest model. Furtheremore, considering the class imbalance in the dataset and the fact that LR is a probabilistic classifier that does not explicitly take the margin into account, it can be justifed why the LR outperforms the SVM and RF models. On the other hand, the RF model had the lowest AUC and accuracy among all the models. It may be due to that RF tends to distinguish the class distribution by evaluating the value of each feature, while in LR and SVM, the correlation is determined by the product of coefficients and values rather than just the values themselves. Considering the complex relationship between different factors and toxicity, SVM and LR can perform superior to RF.

Comparing performance among three toxicities showed that prediction of bowel urgency had the highest AUC and accuracy. This may be due to that the class distributions for diarrhoea and faecal incontinence are imbalanced towards the negative class (no toxicity), meaning that there are more samples for patients without toxicities. Additionally, the grades/labels are ordinal, which means that the class distributions are close together. For example, the difference between “No” and “Mild” toxicity might be small, as might the difference between “Mild” and “Moderate” toxicity. Although data augmentation was used to balance the dataset for all three types of toxicity, the difference between the grades showed that there were not enough good samples for augmentation for diarrhoea and faecal incontinence. For example, the class with toxicity for faecal incontinence includes 64 patients, where the majority of them (52 patients) belong to the category of mild toxicity, which can be overlapped with no toxicity (228 patients). Therefore, data augmentation was not as effective for diarrhoea and faecal incontinence as it

was for bowel urgency, where there were enough samples for each grade.

The analysis of feature importance reveals that dosimetric features ranked as top predictors across all models, indicating a strong association between the irradiated dose and toxicity. These findings suggest that incorporating 3D dose distribution, which provides spatial information as well as dosimetric information, could enhance the prediction performance.

4.7 Summary and conclusions

In this chapter, three ML models were employed to predict toxicity using clinical data. This chapter was important because it provided results for comparison and addressing the question of whether deep learning and analysing 3D spatial information can enhance performance. As mentioned earlier, one of the drawbacks of the traditional models is their inability to incorporate spatial information. Hence, the emergence of deep learning models aimed to overcome this challenge. However, certain challenges persist, particularly regarding the interpretability of the network's process. In the next chapter, a novel deep learning model will be introduced, which leverages spatial information to predict toxicity while addressing the challenges associated with explainability.

Chapter 5

Toxicity Prediction in Pelvic Radiotherapy Using Multiple Instance Learning and Cascaded Attention Layers

This chapter uses a deep learning model to explore 3D dose distributions and CT scans to predict RT-induced toxicity. The main objective of this chapter is to address three key challenges faced by current outcome modelling techniques, as discussed in Chapter 1: exploring the spatial information of input data, detecting the bowel regions that are involved in the toxicity, and explaining the network's behaviour. The results of this chapter can help clinicians have a better understanding of how the irradiated dose and different anatomical regions are correlated with toxicity. This information can be beneficial in guiding the development of optimal dose planning.

A modified version of this chapter originates from my paper "Toxicity Prediction in Pelvic

Radiotherapy Using Multiple Instance Learning and Cascaded Attention Layers" published in *IEEE Journal of Biomedical and Health Informatics*, in February 2023 [47].

5.1 Introduction

Using CT-based models of patient anatomy, modern radiotherapy provides treatment plans that are optimised on an individual patient basis. This optimisation is mostly based on straightforward hypotheses on the relationship between normal tissue and the radiation dose applied to the tumour (an increase in dose will increase cancer control as well as the rate of side effects). Particularly in the case of radiation-induced toxicity, the relationship between the delivered dose and the toxicity is not clearly understood. In this work, a deep learning model is proposed to explore this relationship; the correlations of image data, including CT scans, dose distribution plans, and bowel urgency toxicity are investigated with a 3D convolutional neural network. The model is based on multiple instance learning (MIL) and attention mechanism that generates three outputs:

(i) a binary value predicting the toxicity,

(ii) a toxicity risk map illustrating the anatomical regions association with the toxicity,

(iii) an input importance map presenting the relative importance of CT and dose for the toxicity.

As discussed in earlier chapters, conventional methods for predicting RT outcomes often discard the spatial information of dose distribution. To address this issue, several deep learning models have been proposed to incorporate spatial information. However, a significant challenge remained: these models were unable to detect the correlation of anatomical regions with toxicity. The novelty of the proposed model lies in its ability to detect the anatomical areas that are related to the toxicity.

In Chapter 3 it is explained that RT-induced toxicity was classified with two different ways.

For the purpose of this work, the focus is on analysing mild/severe bowel urgency toxicity, which is defined as reported bowel urgency with grade $\geq$ 2. The reason for this is that the dataset does not have a major imbalance for bowel urgency (161 patients without toxicity vs 79 patients with toxicity). It is not possible to analyse other toxicities in the same way because they are highly imbalanced and it drastically affects the training of the network (even with data augmentation dataset is still imbalanced).

5.2 Methodology

The main structure of the network consists of: (i) two encoders extracting the most significant features from each input (i.e., CT scans and dose distribution plans) separately, (ii) two attention modules describing the network's behaviour, and (iii) one classification module predicting the toxicity. The final output of the network is a binary variable that determines the occurrence of the grade$\geq$ 2 bowel urgency toxicity.

The architecture of the network is illustrated in Fig. 5.1.

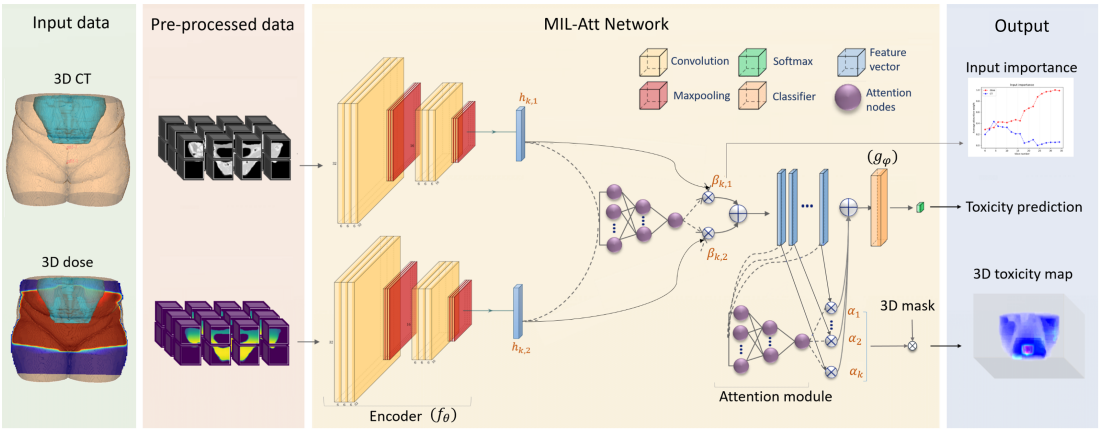


Figure 5.1: The schematic illustration of the proposed model. 3D input images are pre-processed and fed into the MIL-Att network. The output of the network is a binary variable defining toxicity prediction. Attention weights $\alpha_1, .., \alpha_k$ are utilised to generate toxicity maps; and weights of the first attention module, $\beta_1, ..., \beta_k$, are extracted to analyse the impact of each input on the network's decision.

5.2.1 Multiple instance learning with convolutional blocks

Multiple Instance Learning (MIL) is a form of supervised machine learning algorithms designed for datasets where individual instances lack direct labels but they are instead grouped into "bags" that has a labels. MIL are especially beneficial for tasks where accurately labeling individual instances is challenging or unfeasible [20]. For example, in histopathology image processing, a pathologist may assess an entire slide image of a tissue sample to ascertain the presence of cancer cells. However, labeling each individual cell in the slide may be impractical or even impossible [97]. A traditional MIL, the algorithms includes main components as below:

- Instances: the basic unit of data, often represented as a feature vector. In the histopathology example, each instance is the patch extracted from the slide. These instances may or may not have positive/negative labels.
- Bag: a set of instances. Each bag is labelled with a binary class label. The bag label is positive if there is at least one instance with a positive label, otherwise it is negative.
- Learning task: the classifier is trained to identify whether a bag is positive or negative. The training process includes: (i) extracting features from each instanc within the bag. These features capture relevant information for the classification task. (ii) fitting a classification model; there are various models than can be fitted on the extracted features to perform classifcation [68]. An appropriate MIL algorithm should be selected based on the characteristics of the data and the desired outcome. the most common MIL algorithms include SVM, nearest neighbour, ensemble models and neural network.

For this study, a neural network has been chosen and adapted to fulfill the objectives, specifically targeting the automated extraction of features and classification for toxicity prediction. The difference between the proposed MIL and the conventional MIL lies in the labeling strategy. In the proposed model, instances within a bag do not have known labels, while the bag

label is known. In simpler terms, the network is unaware of which particular instances within a positive bag have positive labels, but it has the information that the entire bag is positive.

The proposed MIL model is desgined with a CNN classifier with modification. In a typical binary classification using a CNN, the goal is to detect the best model that assigns a label $y \in \{0, 1\}$ for a given input data x. However, in multiple instance learning one aims to find the label $y \in \{0, 1\}$ for multiple instances belonging to the same category.

Let $X^{(K)} = \{x_1, x_2, ..., x_K\}$ denote a bag of K instances, the MIL model predicts the label $y^{(K)}$ for the entire bag. Fig5.2 shows the difference between a typical deep learning model and multiple instance learning model.

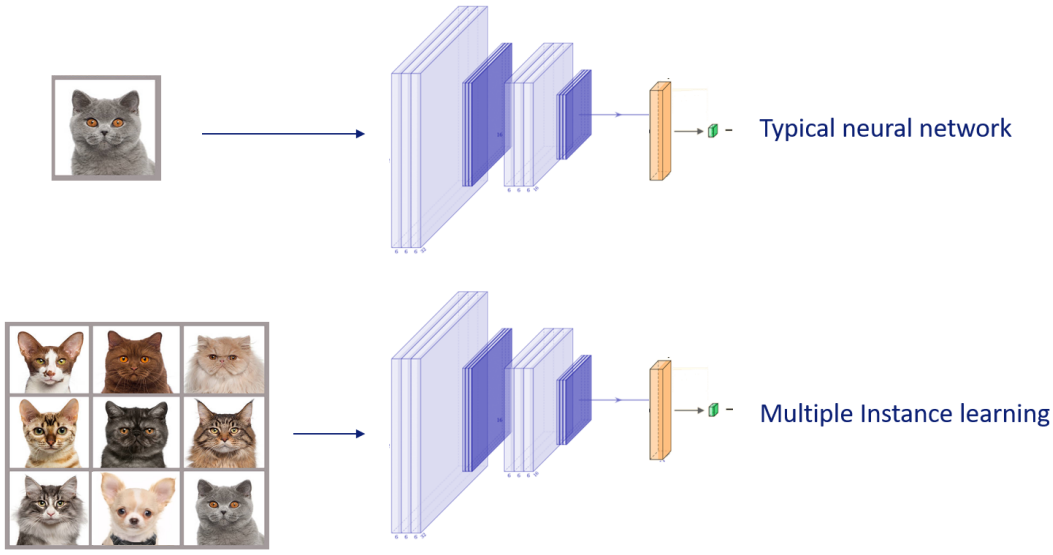


Figure 5.2: The difference between deep learning and multiple instance learning. The deep learning model aims to predict if the input image is a dog, while for MIL the aim of the network is to predict if there is any dog in the input bag.

Multiple instance learning is ideal for a variety of medical imaging jobs due to the issue of poorly annotated data [126], [150], [27], [169], where the input image data can be divided into multiple instances/patches. In this study, MIL is employed for two main reasons; first, medical data are big 3D images, and it is impractical to train a deep neural network with two different 3D inputs; CT scans and dose plans were divided into smaller 3D cubes, and this drastically

reduced time and resource (i.e., memory and GPU) complexity while offered a model that investigated both modalities. Second, a key challenge in an outcome prediction problem is identifying the anatomical regions associated with the toxicity. By attaching attention modules to the MIL model, those instances within the bag (areas in the image) that trigger the final prediction can be discovered.

5.2.2 Attention mechanism

Machine learning-based attention mechanism mirrors the human cognitive process: deliberately focusing on specific relevant features, while ignoring other features.For instance, consider a photograph of a family where the grandfather and grandchildren are seated in a row. When searching for the grandfather, an individual might focus on the color of their hair while disregarding the color of their clothes. Essentially, this person has acquired the knowledge that the color of hair is a distinctive feature of elderly individuals. The same applies to the attention mechanism in neural networks, where the network learns to prioritize certain features as more crucial for determining the final output.

The attention mechanism has been introduced in various fields, but its predecessors in neural networks were primarily used in recurrent models [110], [135]. These early attention mechanisms calculated weights for words within a context sequentially, but they paved the way for more sophisticated attention mechanisms that have found applications in translation models [9], vision models [128] and perceiver machines [79] and transformer [158].

In the context of CNNs, various attention mechanisms, such as self-attention [178], channel attention [72], spatial attention [179], and more, have been implemented. However, the main approach for most of them involves calculating weights for specific components of the model, such as features, channels, or spatial dimensions. A straightforward formulation could be as follows: Let C represent a specific component of a CNN; the attention module calculates

weights for the ith element of C can be calculated in the following manner:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{k=1}^{N} \exp(e_k)}, \tag{5.1}$$

where e_i is the score generated by an alignment model showing how well the component C_i is impact on the output. In the proposed framework, the aim is to detect the most impactful (impact on the toxicity) regions within the input image, therefore the attention computes the weight for features (the component C) that are extracted from convolutional blocks. the alignemt model e is parameterised as a feed forward network which is simultaniusly trained with the main CNN.

Since the aim of this framework was to provide an explanation for the network's output in addition to toxicity prediction, to meet this goal, two attention modules were attached along with the convolutional encoders (see Fig.5.1). In a standard MIL classification task, the features collected from all of the instances are sent to a classification module with equal weight. However, in this work, using the attention mechanism, the weighted average of all the features were computed (the extracted features do not have equal weight). These weights were determined by a feed-forward neural network - the attention module - that was jointly trained with other sections of the network.

Both attention modules are desinged with a two fully-connected layers with ReLU acitivation function and a softmax function for the final output of the module that computes the weights for all features extracted from different location in the input image (or instances in the input bag). The number of layers in the attention module can vary depending on the complexity of the input data. While a larger number of layers can potentially capture more complex relationships and enhance the efficiency of the learned weights, it also increases the risk of overfitting. In this research, due to the relatively small input size (small patches extracted from original image), employing more than two layers led to overfitting and was therefore inappropriate for

the model.

The novelty of this chapter is employing two attention modules that separate the attention weights over feature and space. This allows: (i) easily distinguish the contributions of CT and dose inputs to the predicted toxicity and (ii) locating the anatomical regions associated to bowel urgency toxicity. Thus, for each individual patient, the network creates two risk maps; one highlights the critical regions, where the patient's OAR and dose delivered together drive the toxicity risk, and the other describes how CT and dose plans trigger the network's outcome.

5.2.3 Model formulation

Consider a bag with K instances as $X^{(K)}$ where: $X^{(K)} = \{\mathsf{X}_{1,i}, \mathsf{X}_{2,i}, ..., \mathsf{X}_{K,i}\}$, and $\mathsf{X}_{k,i}$ represents the k^{th} instance from i^{th} input of the bag. To simplify matters, X and y are denoted instead of $X^{(K)}$and $y^{(K)}$ for each bag and label, respectively. More details of the notation are shown in Table 5.1.

Table 5.1: Summary of the notations.

Notation	Description	Value
k	Instance number	$1 \leq k \leq K$ (K can be different for each bag)
i	Input number	$i \in \{1, 2\}$, CT: $i = 1$, dose: $i = 2$
$\boldsymbol{h}_{k,i}$	Feature vector for instance k and input i	$\boldsymbol{h}_{k,i} \in \mathbb{R}^{1\times l}$
$\boldsymbol{w}, \mathbf{V}$	Weights for input attention	$\boldsymbol{w} \in \mathbb{R}^{d\times 1}, \mathbf{V} \in \mathbb{R}^{d\times l}$
$\beta_{k,i}$	Attention weights for instance k, input i	$\beta_{k,i} \in \mathbb{R} \mid\ 0 \leq \beta_{k,i} \leq 1$
$\boldsymbol{z}_k$	Weighted feature for instance k	$\boldsymbol{z}_k \in \mathbb{R}^{1\times l}$
$\boldsymbol{q}, \mathbf{R}$	Weights for region attention	$\boldsymbol{q} \in \mathbb{R}^{p\times 1}, \mathbf{R} \in \mathbb{R}^{p\times l}$
α_k	Attention weights for instance k	$\alpha_k \in \mathbb{R} \mid\ 0 \leq \alpha_k \leq 1$
$\boldsymbol{s}$	Weighted feature for input bag	$\boldsymbol{s} \in \mathbb{R}^{1\times l}$

The outcome prediction problem can be formulated as a posterior probability as below:

$$y = \Phi_{\Omega}(X), \quad y \in [0, 1], \tag{5.2}$$

where Ω is the set of the model parameters. Considering $\mathsf{X}_{k,i}$ as an input cube (an instance in the bag) representing cube k from input i ($i = 1$ for CT and $i = 2$ for dose), the encoder module can be written as $f_{\theta_i}(\mathsf{X}_{k,i})$, which is a convolutional neural network with the parameters θ_i. The module's output is a vector, $\boldsymbol{h}_{k,i}$, containing the most important features of the input cube ($f_{\theta_i}(\mathsf{X}_{k,i}) = \boldsymbol{h}_{k,i}$).

Once the features have been extracted from each input cube, the attention module β calculates which input (CT or dose) has more relevant features for toxicity prediction. The attention module is a two-layer fully-connected network that takes feature vectors as input and outputs the importance weights of the inputs. It can be formulated as:

$$\beta_{k,i} = \frac{\exp\{\boldsymbol{w}^T \tanh(\mathbf{V}\boldsymbol{h}_{k,i}^T)\}}{\sum_{j=1}^{2} \exp\{\boldsymbol{w}^T \tanh(\mathbf{V}\boldsymbol{h}_{k,j}^T)\}}, \tag{5.3}$$

where $\mathbf{V}$ and $\boldsymbol{w}$ are weights matrix and vector, respectively. The value of $\beta_{k,i}$ is the importance weight for the cube k in the input i. Therefore, the ultimate feature vector representing the input cube is computed as:

$$\boldsymbol{z}_k = \sum_{i=1}^{2} \beta_{k,i} \boldsymbol{h}_{k,i}. \tag{5.4}$$

Next, the aim is to find those anatomical regions which correlates the most with the toxicity outcome. Attention module α computes the importance weight for each cube based on its location. It can be written as:

$$\alpha_k = \frac{\exp\{\boldsymbol{q}^T \tanh(\mathbf{R}\boldsymbol{z}_k^T)\}}{\sum_{j=1}^{K} \exp\{\boldsymbol{q}^T \tanh(\mathbf{R}\boldsymbol{z}_j^T)\}}, \tag{5.5}$$

where $\mathbf{R}$ and $\boldsymbol{q}$ are the weight matrix and vectors parameters, respectively.

Finally, the feature vector that is ultimately fed to the classification module can be computed

using:

$$\boldsymbol{s} = \sum_{k=1}^{K} \alpha_k \boldsymbol{z}_k, \tag{5.6}$$

Based on 5.2, the toxicity classification can be written as:

$$\begin{aligned} &y = \Phi_\Omega(X) = g_\varphi(\boldsymbol{s}), \\ &g_\varphi : \boldsymbol{s} \mapsto [0,1], \quad \Omega = \{\varphi, \theta, \boldsymbol{w}, \mathbf{V}, \boldsymbol{q}, \mathbf{R}\}. \end{aligned} \tag{5.7}$$

where g_φ refers to a four-layer fully-connected network that generates the probability of the toxicity occurrence (classification module). Assuming $t \in \{0, 1\}$ represents the target class for X, the model can be trained through minimizing the binary Cross-Entropy loss function given by:

$$L(t, \Omega) = -t \log(\Phi_\Omega) - (1 - t) \log(1 - \Phi_\Omega) \tag{5.8}$$

The loss function is summed over all input bags in the training set and minimisation is performed w.r.t. Ω parameters.

5.3 Experimental results

The patients were divided into two classes: those with moderate or severe (responded 2 or 3) toxicity (79 patients) and those with no or mild (responded 0 or 1) toxicity (161 patients). For the sake of simplicity, the former group are referred as patients with bowel urgency toxicity and later without bowel urgency toxicity. 75 participants were not included in the study because of incomplete data or having stomas (item not relevant).

5.3.1 Implementation details

Before the training, a 3D rigid transformation (SimpleITK [104], version 2.0.1) was utilised to register CT scans and dose distributions to a reference patient with CT images of dimen-

sions $[35, 512, 512]$ voxels. The patient with the minimum number of slices for the bowel bag structure was selected as the reference. The size of each pixel was $5mm \times 0.97mm \times 0.97mm$ and each input data was divided into smaller cubes with voxel dimensions of $[6, 32, 32]$. The encoder had two 3D convolutional modules, in which each convolution layer is followed by maxpooling and batch normalisation layers. 30 and 50 convolution filters with a kernel size of $(2, 3, 3)$ were utilised in the two convolutional layers. Table 5.2 shows the parameters and details of the network architecture.

Table 5.2: Parameters of the MIL-Att network

Module	Layer	Kernel size	Filters/ Units	Stride	Padding
Encoder-dose	Conv + ReLU+BN	(2,3,3)	30	(1,1,1)	0
	Maxpool	(1,2,2)	-	(1,2,2)	0
	Conv + ReLU+BN	(2,3,3)	50	(1,1,1)	0
	Maxpool	(1,2,2)	-	(1,2,2)	0
Encoder-CT	Conv + ReLU+BN	(2,3,3)	30	(1,1,1)	0
	Maxpool	(1,2,2)	-	(1,2,2)	0
	Conv + ReLU+BN	(2,3,3)	50	(1,1,1)	0
	Maxpool	(1,2,2)	-	(1,2,2)	0
Attention-α	Linear + Tanh	-	512	-	-
	Linear + Softmax	-	1	-	-
Attention-β	Linear + Tanh	-	512	-	-
	Linear + Softmax	-	1	-	-
Classifier	Linear + ReLU	-	1000	-	-
	Linear + ReLU + DP	p=0.5	500	-	-
	Linear + ReLU + DP	p=0.5	50	-	-
	Linear + Softmax	-	1	-	-

Abbreviations: Conv, convolutional layer; ReLU, rectified linear unit activation function; BN, batch normalisation; Tanh; hyperbolic tangent activation function. DP, drop out.

Adam optimisation [86] with the learning rate of $1e^{-4}$ was employed. The number of neurons in both attention modules was set up to 512 ($p, d = 512$, see Table 5.1). ReLU activation function was used for all inner layers, and the last fully-connected layers in the classification and attention modules were activated using Sigmoid and Softmax functions respectively.

5.3.2 Training strategy

The dataset was divided into training and test sets. The test set consisted of 40 patients: 20 patients with bowel urgency and 20 patients without bowel urgency were randomly selected. The remaining was used for training set (209 patients).

Because the training set was imbalanced, with only 59 patients with toxicity compared to 141 patients without toxicity, data augmentation was used to address this. Specifically, Gaussian noise (with a zero mean and 0.1 standard deviation) was added as well as smoothing recursive Gaussian noise (with a 5 mm sigma across each axis) to the data of 59 patients in the minority. This was performed using SimpleITK Python toolbox [104] version 2.1.1, which provides both filters for 3D image analysis. The augmentations were applied to the dose distributions and CT scans in their original resolution, before dividing the 3D images into smaller cubes.

After augmentation, the training dataset consisted of 321 patients. During each epoch, 40 patients were randomly selected for the validation set, and the network performance was evaluated by observing the results on this validation set.

Cross-validation was not used for two reasons. Firstly, when the dataset is imbalanced, dividing it into smaller subsets may leave certain folds without a positive label, which can affect the accuracy and AUC measures used to assess the classifier's performance. Secondly, when using data augmentation, only non-augmented data can be validated. Therefore, the augmentation process must be repeated for each fold, which can be computationally expensive.

Convolutional autoencoders for transfer learning

To avoid overfitting while increasing the network generalisation, a transfer learning strategy was employed. Two autoencoder (AE) networks, sharing the same architecture as MIL-Att encoders, were separately trained using the pre-processed CT and dose images in the training set, resulting in AE-CT and AE-dose, respectively.

Only 240 of the 315 patients in the dataset were candidates for bowel urgency toxicity (n=64 patients with a stoma did not complete this questionnaire item). Except for the 40 patients in the test set that were left out, AEs were trained with all the patients (both candidates and non-candidates for bowel urgency). The goal of the autoencoders was to identify the key features from the input data that are important for re-creating them. In total, 275 image datasets were divided into cubes with dimensions of [6,32,32] voxels, and the AE networks were trained on 32303 cubes.Note that only cubes with values greater than zero were taken into account for training, resulting in a different number of cubes for each patient. The loss function of the AEs was mean squared error and the reconstruction error was computed as 0.0091. Table 5.3 summarises the details of the autoencoder architecture.

Table 5.3: parameters of the Autoencoder

Module	Layer	Kernel size	Filters/ Units	Stride	Padding
Encoder	Conv1 + ReLU+BN	(2,3,3)	30	(1,1,1)	0
	Maxpool1	(1,2,2)	-	(1,2,2)	0
	Conv2 + ReLU+BN	(2,3,3)	50	(1,1,1)	0
	Maxpool2	(1,2,2)	-	(1,2,2)	0
Decoder	ConvTranspose1+ReLU+BN	(2,3,3)	30	(1,1,1)	0
	MaxUnpool1	(1,2,2)	-	(1,2,2)	0
	ConvTranspose2+ReLU+BN	(2,3,3)	1	(1,1,1)	0
	MaxUnpool2	(1,2,2)	-	(1,2,2)	0

Abbreviations: Conv, convolutional layer; ReLU, rectified linear unit activation function; BN, batch normalisation; Tanh; hyperbolic tangent activation function. DP, drop out.

The proposed framework was tested with different training strategies. First, the network was trained from scratch (MIL-Att-scratch), which means that all the network's weights were initialised at random values. Second, to explore how transfer learning can improve the performance, the network was trained with the following settings:

(i) both encoders in MIL-Att network were initialised with the weights of the autoencoder trained with CT images, called MIL-Att-CT,

(ii) both MIL-Att encoders were initialised with the weights learnt by AE-dose, called MIL-Att-dose,

(iii) CT and dose encoders were initialised separately using the weights of AE-CT and AE-dose, respectively, called MIL-Att-both.

Fig. 5.3 shows the comparison of training and validation losses for different training modes.

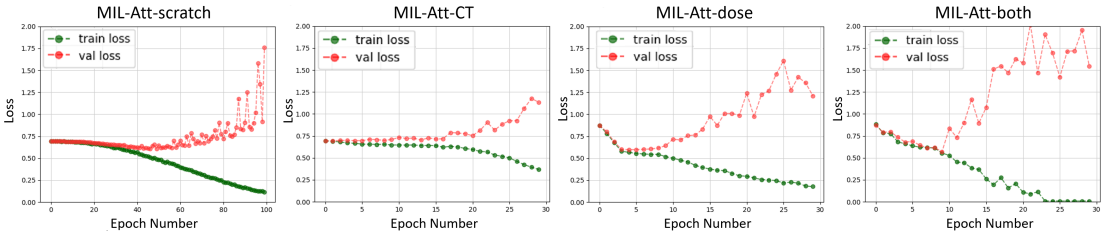


Figure 5.3: Training and validation loss for four modes of MIL-Att network. Training from scratch converged at higher epochs while for pretrained modes it converged at lower epochs.

Each network was trained over a number of epochs, computing training and validation losses for each epoch. The training loss and validation loss both decline and stabilise for all training modes at a particular point that denotes an optimal fit. To avoid overfitting, we performed early stopping; during training in each epoch, 40 data were randomly selected to validate the training procedure and monitor the performance of the model on the validation set. The training was stopped once the performance on the validation set started to decrease. Fig. 5.3 shows that without early stopping there would be severe overfitting. Other approaches including batch normalisation, drop out and ransfer learning are also employed to help with the overfitting problem. On the basis of this, the best model was chosen (model with the lowest validation loss).

For training from scratch (without transfer learning; MIL-Att-scratch), the network was trained for 100 epochs. Up to epoch number 47, both training and validation loss decreased (see MIL-Att-scratch in Fig.5.3). The network was overfitted at the epoch 47 because the training loss drastically reduced while the validation loss highly increased. All networks converged

in lower epochs (epochs < 16) for transfer learning modes (MIL-Att-CT, MIL-Att-dose, and MIL-Att-both), therefore training was stopped after 30 epochs (for this reason, the learning rate for transfer learning modes was set to $1e^{-3}$). After epochs > 16, the validation loss was significantly higher than the training loss, that also shows that the network was overfitted. The best validation results (the lowest validation loss) gained by MIL-Att-both indicating that pretraining both encoders can improve the training process.

Classification performance

The network performance was measured by assessing five evaluation metrics: accuracy, sensitivity, specificity, F1-score, and AUC. The findings were compared with three previously proposed models by Yang et al.[172], Ibragimov et al. [76] and Liang et al.[99] (reviewed in 2.5) each for a specific reason.

Yang et al.[172] proposed a convolutional network (CT-dose-CNN) that, like this proposed network, included two different channels for analysing CT scans and dose images for toxicity prediction after prostate RT. Liang et al. [99] transferred the weights learnt by C3D (proposed in [151]) to predict the toxicity outcome after lung radiotherapy. The network was trained with two different settings; training all layers (C3D-FT) and only training the last layer (C3D-FE). The authors used Grad-CAM to generate a toxicity risk map. The prediction performance of their model was compared to the current network and the associated attention risk map was correlated with their Grad-CAM toxicity map. The last work, Ibragimov et al. [76] used a 3D convolutional network (Dose-CNN) with three layers to predict liver toxicity. Only one input channel is included in their proposed network, and they only analyse dose's impact on toxicity. The current model was compared to theirs to identify how exploring CT data in addition to dose can affect the performance.

All models were implemented with Python 3.7 and followed the given procedures in the respective papers to train. Table 5.4 and Fig. 5.4 summarise the prediction performance experiments.

Table 5.4: Comparison of prediction performance across different methods. Best performance in each metric is shown in bold. All the reported results pertain to the performance achieved on the test set.

Method	Parameter	Accuracy	Specificity	Sensitivity	F-1
MIL-Att	scratch	0.65	0.70	0.60	0.64
	CT	0.65	0.60	0.70	0.64
	dose	0.75	0.75	0.75	0.80
	both	**0.80**	0.80	**0.80**	**0.82**
C3D[99]	FE	0.65	**1.0**	0.30	0.65
	FT	0.72	0.75	0.70	0.75
CT-dose-CNN[172]	-	0.60	0.55	0.65	0.60
Dose-CNN[76]	-	0.60	0.85	0.35	0.46

Abbreviation:: FE, feature extraction mode; FT, fine tuning mode.

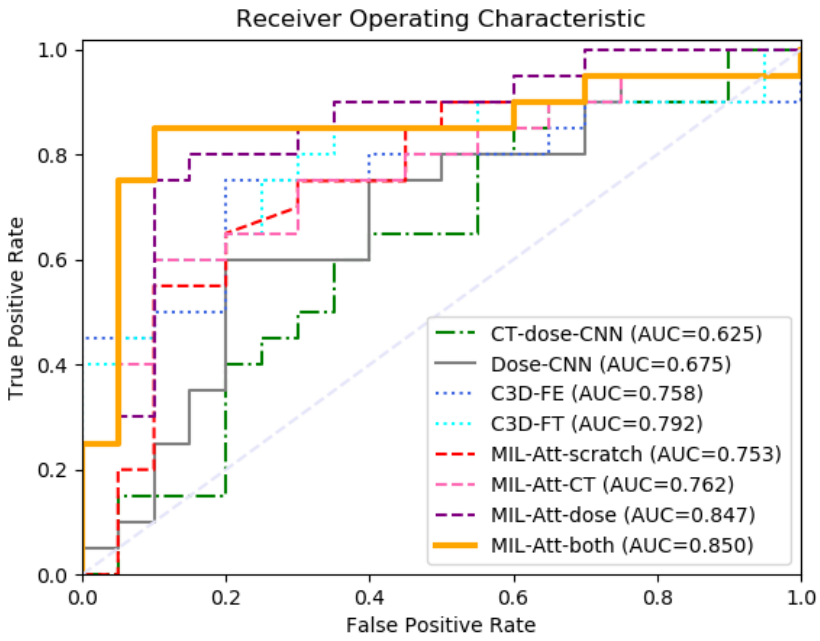


Figure 5.4: Receiver operating curve analysis for toxicity prediction using the test set.

Table 5.4 shows that the prediction performance of MIL-Att-both gained the highest values for accuracy, sensitivity and F1-score, while for specificity, the C3D-FE model achieved the best value (1.0). The specificity summarises how effectively the negative class is classified, and sen-

sitivity metric is the complement to it. The low value of 0.3 for sensitivity in C3D-FE indicates that the C3D-FE network was biased towards predicting negative class (without toxicity) and it did not adequately learn the data distribution for the positive class (with toxicity). With the same inference, Dose-CNN was also biased toward predicting negative labels. The next best value for specificity was achieved by MIL-Att with pre-trained encoders. The ROC analysis (Fig.5.4) also supports the superiority of the MIL-Att-both model compared to the others.

AUC evaluation with DeLong's test

To investigate significant differences between models, AUC metrics were compared using DeLong's test [38]. DeLong et al. proposed a technique for determining whether the AUC of one machine learning model differs significantly from an alternative model. They proposed to compute empirical AUC rather than traditional (binomial) AUC for this reason: because the normal AUC relies on strong normality assumptions, while the empirical AUC does not. In other words, the typical AUC is invariant with respect to the rate of positive samples and in the case of a small and imbalanced test set, comparing AUCs may not reflect the difference between the performances of the two models. The empirical AUC can be calculated as follows:

$$\hat{\Theta} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\Psi(x_i, y_i), \qquad where \quad \Psi(x_i, y_i) = \begin{cases} 1 & \text{y<x} \\ \frac{1}{2} & \text{y=x} \\ 0 & \text{y>x} \end{cases} \tag{5.9}$$

In order to investigate whether one model is better than the other in terms of AUC, the z score can be calculated between two models as below:

$$z \triangleq \frac{\hat{\Theta}_{(A)} - \hat{\Theta}_{(B)}}{\sqrt{\mathrm{Var}[\hat{\Theta}_{(A)} - \hat{\Theta}_{(B)}]}} \tag{5.10}$$

After calculating the z score between two models, the corresponding p-value can be extracted

from a lookup table presented by [154]. If p-value< 0.05 it can be concluded that Model A has a statistically significant different AUC from Model B. DeLong's test was performed using Python library [144] on the test set to compare different models. Fig.5.5 illustrates the experiment results.

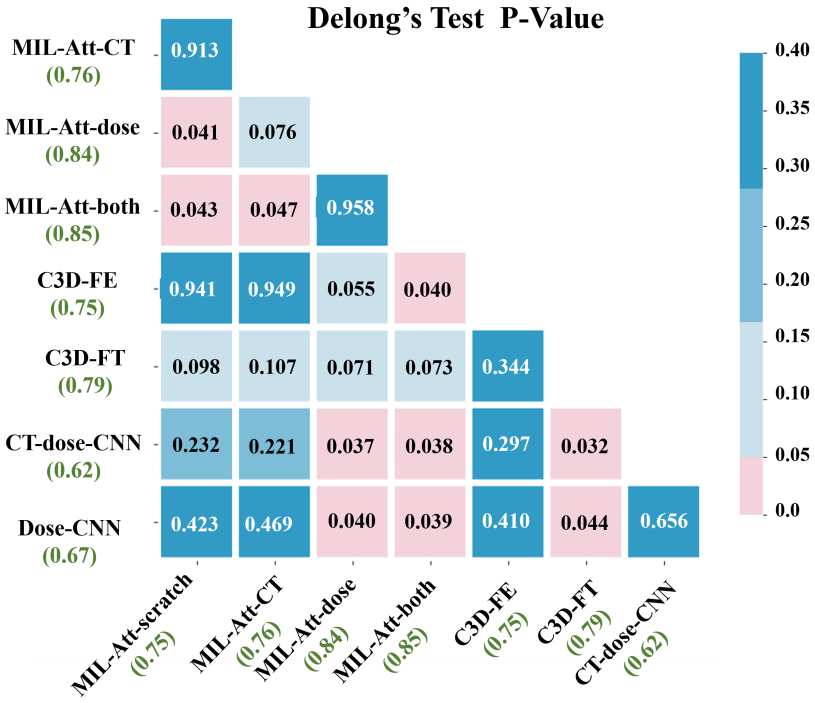


Figure 5.5: AUC comparison using DeLong's test. Smaller p-values demonstrate significant differences. AUC values are in parentheses.

From Fig. 5.5 it can be seen that CT and dose pre-trained networks improved the AUC compared to training from scratch (p-value< 0.05). Considering p-value= 0.913, it can be concluded that the pre-training with CT did not statistically improve the model's performance compared to training from scratch. Additionally, the network that was pre-trained with both CT and dose did not perform noticeably better than the network pre-trained with only dose data (p-value=0.958).

These findings imply that the model's performance can be enhanced by pre-training the network using dose data. This may be due to the features of dose distribution are different than those

of CT data; dose data is a grayscale image with subtle intensity variations, whereas CT data contains several edges, corners, ridges, and other distinctive features.

5.3.3 Prediction performance discussion

When comparing the proposed model (MIL-Att) and C3D with other methods, it can be observed that they outperform CT-dose-CNN and Dose-CNN. This may be related to the depth of the networks or how the features are extracted; three convolutional layers in both networks (CT-dose-CNN and Dose-CNN) collect features from the input data. In contrast, C3D used eight convolutional layers to transform the input into latent features.

On the other hand, MIL-Att network extracts the features locally using two encoders that each has two convolutional layers. The proposed network explores the 3D CT scans and dose distributions locally for every cube. The average number of cubes for input data was 209, which means the network represents the input on average with 209 cubes and extracts feature from each of them separately. One might conclude that the architecture of the layers for CT-dose-CNN and Dose-CNN may not be sufficient to represent the dataset's natural pattern of the data distribution.

The performance of C3D is very close to that of MIL-Att. This is because both methods learn deep features either from the entire input or by exploring local areas in the data. However, MIL-Att outperformed C3D overall on the assessment metrics on the test set. This is because, opposite to C3D, the MIL-Att network analyses CT scans and dose data. This finding emphasises the fact that combining CT scans with dose distributions can provide more useful information and, as a result, improve prediction performance.

Comparing network architectures, MIL used fewer convolutional layers than C3D. This noticeably reduces the number of parameters to learn; for example, MIL-Att learns about 11 million parameters, compared to C3D's about 78 million. It should also be considered that C3D only investigates one input, while MIL-Att analyses two 3D inputs.

The ultimate software developed from these models will be utilised in hospitals and it must be stored in conventional computers (it is not common for hospitals to have GPU machines with large memories). Therefore, it could be said that when analysing large medical data, MIL can be more efficient regarding time and resource usage. It should be noted that while the concept of MIL may share similarities with other efficient models that explore adjacent patches or slices (such as 2.5D networks), the main difference is that in MIL, the entire volume is still considered as the input. Therefore, the output is determined based on the complete volume, not just each stack or cube. In problems likes classification, where the final result is determined based on the whole input (despite 3D segmentation), multiple instance learning is particularly memory and GPU usage efficient.

Comparing CT-dose-CNN and dose-CNN reveals that despite analysing both CT and dose images, CT-dose-CNN had a lower AUC. This might be because CT-dose-CNN applies convolutional filters to the original size of the CT and dose, whereas Dose-CNN scales down and crops the input to the size of [19,19,19] images. This suggests that the latent space dimensions in the CT-dose-CNN are not enough to project (present) the inherent structure of the input data (i.e., CT and dose) in the dataset.

In summary there are three broad points:

1) The best performance is gained when both CT and dose data are sufficiently explored.

2) When memory is limited and input data are huge 3D volumes, a multiple instance learning model can encode the input more efficiently than conventional deep learning networks.

3) Extracting the most discriminating features is a key point in the toxicity prediction problem.

5.3.4 Patient specific risk map

Generating toxicity risk map

The main novelty of the MIL-Att network is that it can present visual explanations for the network's behaviour by using an attention mechanism. In MIL-Att, the features are locally extracted from different regions of the bowel bag and the attention layers investigate how they affect the final decision (output). Finding these critical areas is necessary for clinicians, as they need to know how the network makes its decisions. Consequently, it can help them with optimal radiotherapy treatment planning.

After training, the second attention weights (α_k) were retrieved for every patient in the test set to create the risk map. 3D visualisations of the toxicity maps for two patients without bowel urgency and two patients with bowel urgency are shown in rows A and B of Fig. 5.6, respectively.

Results were plotted in three distinct views for easier visualisation. Greater values in the risk map show more critical regions for the bowel urgency toxicity. For both patients that reported bowel urgency, the attention weights are gathered in the anterior and iliac fossa anatomical areas of bowel bag, indicating that these locations are linked to a higher risk of bowel urgency. Contrarily, in patients without bowel urgency, the attention weights are dispersed across the entire intestinal bag and are not concentrated in any one area.

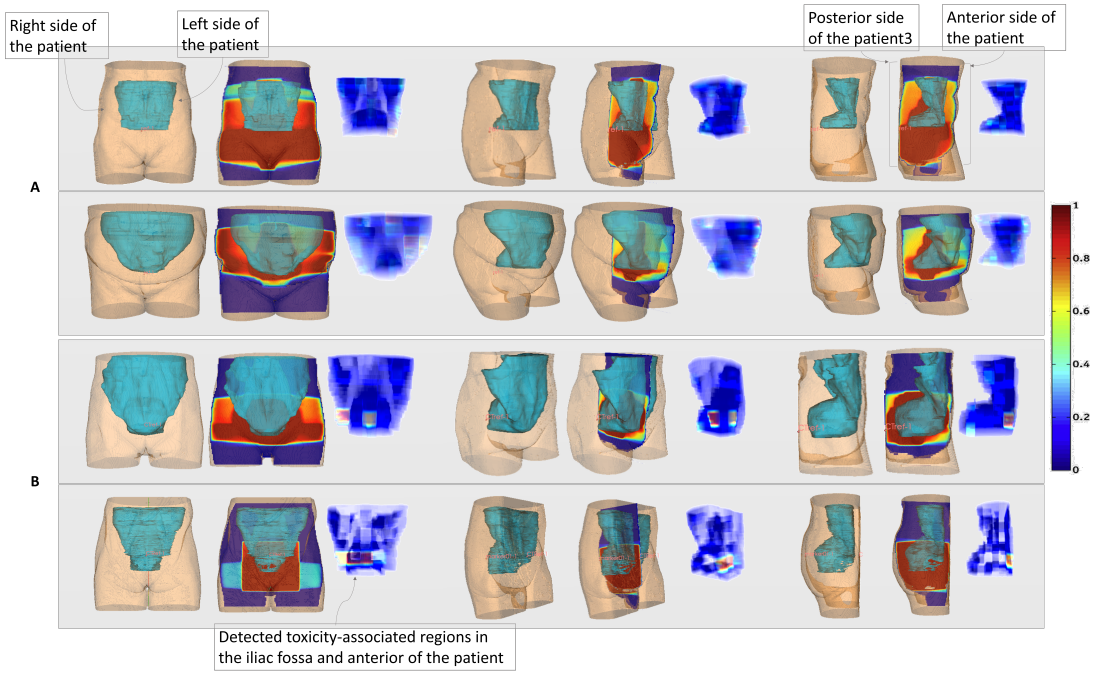


Figure 5.6: 3D comparison of generated toxicity map by MIL-Att for four patients in two groups of A and B without and with bowel urgency, respectively. From left to right: the first image shows patient CT scan where the bowel bag is detected with blue contour. The second image is the dose distribution overlapped with the CT scan, and the third image is the attention map generated by the proposed model. The attention map is shown with heatmap, where the higher numbers show the more association with toxicity. For each patient, three different views are plotted to clarify the toxicity-associated regions.

Comparison with Grad-CAM

As mentioned in Section 2.5.6, many studies have employed the Grad-CAM technique to explore network attention, in an attempt to understand the spatial features affecting toxicity risk. The current model was compared with Grad-CAM based on the C3D-FT network to further assess the derived attention risk maps. The PyTorch Grad-CAM library[57] was adopted and adapted for 3D data. Fig.5.7 illustrates the 3D comparison of risk maps for two sample patients, A and B, without and with bowel urgency, respectively.

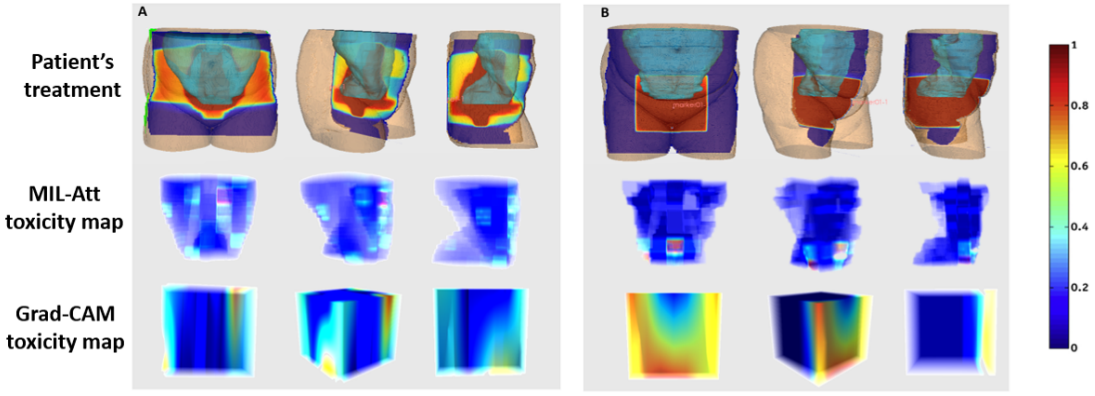


Figure 5.7: 3D comparison of generated toxicity map by MIL-Att and Grad-CAM for two patients A and B without and with bowel urgency, respectively.

Attention weights are distributed throughout the bowel bag for the patient without bowel urgency (patient A). In contrast, for the patient with toxicity (patient B), all the attention is on the front side (anterior) of the patient's bowel bag. Grad-CAM also produced similar results; for patient without bowel urgency, gradients of the features are distributed across the cube, while for patient with bowel urgency, the highest gradients are concentrated anteriorly.

The 2D presentation of the toxicity risk maps were also plotted to have a thorough comparison. Fig. 5.8 depicts three slices of the patient's abdomen, covering various parts from the bottom (slice number = 0) to the top (slice number = 35).

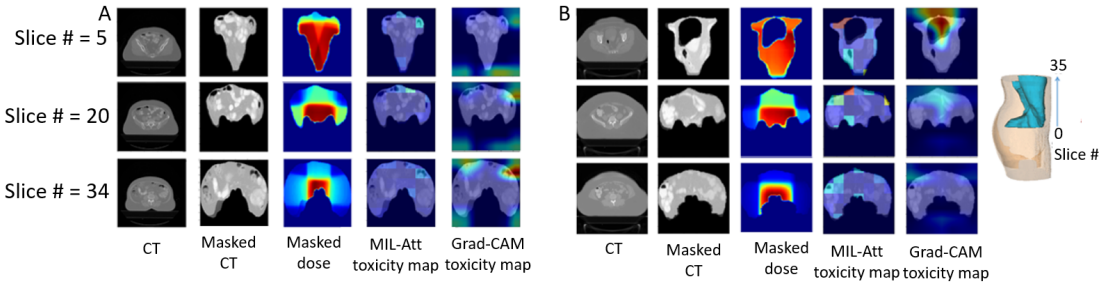


Figure 5.8: 2D comparison of generated toxicity map by MIL-Att and Grad-CAM for two patients A and B without and with bowel urgency, respectively.

From Fig.5.8, it can be seen that the Grad-CAM activation map varies slightly as the number

of slices increases. This is because the feature maps in the last convolution layer (those that are used to generate the risk map) are noticeably smaller than the input image's size, and to generate the activation map, they should be up-sampled. This generates an estimated representation for the high-risk regions. The input size in the C3D network is [18,112,112] voxels, and the size of the features in the final convolution layer is [2,7,7] voxels. Due to the up-sampling on a grid, the Grad- CAM map does not match the morphology of the bowel bag anatomy. As opposed to Grad-CAM, which highlights the entire anterior area of the activation maps as crucial locations, the MIL-Att generates attention weights individually for each cube in the bowel bag, making toxicity localisation more accurate and reliable.

5.3.5 Association of input data with high-risk toxicity

In addition to the dose distribution, recently developed deep learning algorithms incorporated additional input parameters (such as CT, PET) for their model (previously addressed in Section 2.5). Men et al. [112] showed that prediction performance improved when the network was trained with both CT and dose distributions. At the time of writing, no published research has examined the relative significance of the inputs on toxicity outcome. Using an attention module, each input's importance for the network decision was assessed. The attention module β computes the significance of each cube based on its location in dose and CT. The module is a fully-connected network that outputs a value in $(0, 1)$, which shows the importance of the input cube. The attention weights β were extracted for cubes in CT and dose data and computed the average of the weights for each slice inside the pelvis (see Fig. 5.9)

For the bottom part of the pelvis (inferior slices of the bowel bag; slice number$<$ 10), CT and dose slices obtained equally high attention. This suggests that the anatomical information (feature extracted from the CT scans) and received dose are both related to the toxicity risk. For the more cranial portion of the pelvis (slice number$>$ 10), features collected from the dose distribution gained higher attention (weights) to predict the bowel urgency toxicity. From a clinical standpoint, this makes sense: the structure of bowel present in the lower part of the

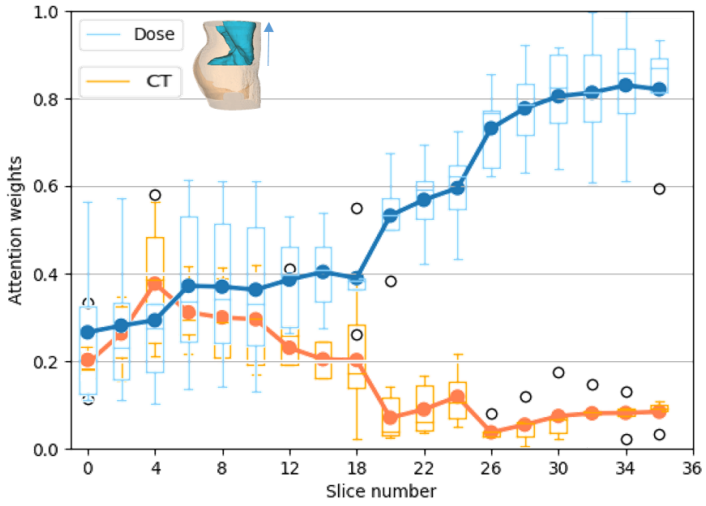


Figure 5.9: Quantitative evaluation of input association with toxicity. Higher value of attention weights shows higher impact on toxicity prediction.

pelvis can be different for each patient (for example, due to bladder size/location); and the risk of toxicity not only depend on the dose received, but also on the patient's bowel shape (for example, if the patient has low-lying bowel loops). On the other hand, the upper section of the pelvis exhibits a more consistent bowel structure, as all patients will have both large and small bowels in this area. The sole factor in this region distinguishing patients with and without bowel urgency is likely the dose received. As an illustration, in Fig. 5.8 shows that patients A and B have nearly the same bowel bag structure in slices 20 and 34, but distinct morphologies in slice 5, representing the inferior aspect of the bowel bag.

After averaging the attention weights across all slices, the CT and dose inputs had average weights of 0.17 and 0.53, respectively. When the overall weights were normalised, it was observed that the dose distribution had a strong association of 76% with bowel urgency toxicity. In contrast, the CT image had a comparatively weaker association of only 24%.

5.3.6 Atlas for toxicity modelling

Because patients' organs vary in size and shape, the toxicity risk maps differ for each person. Instead of using separate risk maps to study correlations between anatomical regions and bowel urgency toxicity, a toxicity atlas was created from all the maps produced by the proposed model. Using the 3D Diffeomorphic Demons registration technique [159], the CT images for all patients with toxicity were co-registered to the reference patient. Dose slices and toxicity maps was also registered with the same transformation (computed from their corresponding CT). The average of all the transformed risk maps was computed (called atlas). Fig. 5.10 depicts the result of the generated atlas. Considering the attention atlas, one can conclude that bowel urgency toxicity is related to the irradiation dose delivered to the bowel bag's anterior and right iliac fossa regions.

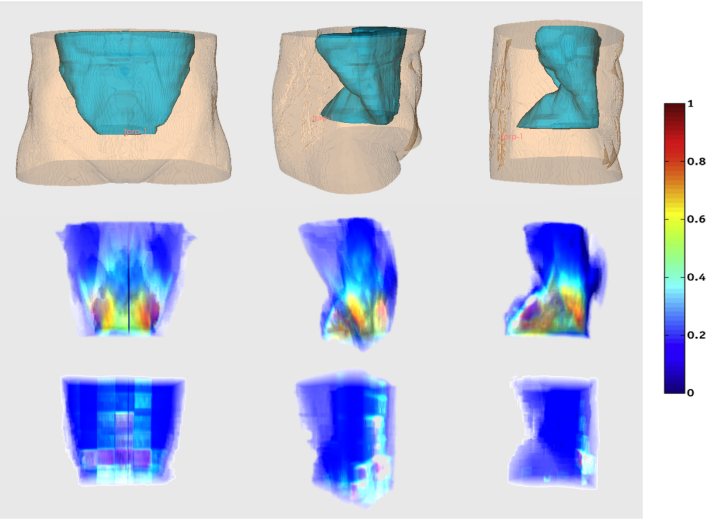


Figure 5.10: Toxicity model. From top to bottom: reference patient, average of irradiated dose and atlas for toxicity. Generated atlas localises high-risk regions for toxicity with higher values.

From a clinical perspective, these findings could potentially be applied to radiation therapy treatment planning to minimize radiation exposure to these sensitive anatomical regions. It is most probable that the anterior dose region is related to small bowel loops (within the radiotherapy field). As the use of IMRT becomes more prevalent in clinical practice, where radiation

therapy dose can be shaped to more easily avoid organs at risks, the dose to this region is likely to be lower than what was observed in this series (where most patients received 3D-conformal radiation therapy). On the other hand, the right iliac fossa region is likely linked to the radiation dose to the terminal ileum. This region is commonly impacted in Crohn's disease, and it is known to be associated with bowel symptoms, such as diarrhoea and urgency. Damage to this region may be related to nutritional deficiencies such as bile acid malabsorption, which is identified as a known side effect after pelvic radiotherapy [3]. This anatomical region is not currently an avoidance structure in standard radiotherapy practice and should be an area of future research.

5.4 Summary and conclusions

In this chapter, a novel deep learning model was proposed for predicting bowel urgency toxicity following pelvic radiotherapy.

The novelty of this work is that it goes beyond mere prediction of toxicity and instead provides explanation of how the network detect that toxicity. To do that, the model integrated two attention modules to clarify: (i) which anatomical regions are correlated to a higher risk of toxicity and (ii) how CT and dose images impact the network's prediction.

Additionally, a toxicity atlas was constructed which integrates information from all toxicity risk maps, summarising and visualising the toxicity based on the bowel bag structure. This proposed visual interpretation helps enhance understanding of the network's function and can assist clinicians having a better view of how each input data are involved with the toxicity. The comparative experiment demonstrated that this framework provides clinically convincing tools for radiotherapy outcome prediction.

As mentioned previously, the informative factors for toxicity prediction extend beyond imaging and dose data. Clinical data has also proven to be valuable in predicting toxicity and are

currently utilised in NTCP models. The upcoming chapter further explores this analysis by incorporating clinical factors into the model, aiming to evaluate the impact of adding these variables to enhance the predictive capabilities.

Chapter 6

Deep Learning Combining Imaging, Dose and Clinical Data for Predicting Bowel Toxicity After Pelvic Radiotherapy

In this chapter, a new architecture for MIL-Att model is proposed which simultaneously analyses CT scan, dose distributions and patients' clinical features. The primary objective of this chapter is to predict toxicity by analysing different input data and provide an explanation of how each input can impact toxicity. Specifically, the question as to whether clinical features provide additional useful information to the toxicity prediction model is explored. The novelty of this chapter is that, in contrast to a few previous works that integrated clinical data with imaging and dose, this neural network can identify which clinical features are associated with a higher risk of toxicity. The findings of this chapter can assist clinicians in gaining a better understanding of how different data can influence treatment outcomes and which features may

pose a risk for each patient.

A modified version of this chapter originates from my paper "Deep learning combining imaging, dose and clinical data for predicting bowel toxicity after pelvic radiotherapy", submitted to the Medical Physics journal and currently under review. Additionally, a section of the analysis from this chapter was published as a scientific abstract entitled "Deep learning with visual explanation for radiotherapy-induced toxicity prediction", presented at the SPIE Medical Imaging, Computer-Aided Diagnosis conference held in San Diego, California, United States in February 2023 [46].

6.1 Introduction

As discussed in earlier chapters, traditional methods for analysing radiotherapy dose data often overlook the complete 3D treatment information. In contrast, deep learning models have shown promise in their ability to automatically process 3D information and have achieved notable success. However, side effects caused by radiotherapy not only depend on dose but also various other factors, including cellular properties, organs' physiology and anatomy, treatment features and patient's characteristics[78], [146], [54]. Therefore, for a thorough understanding of toxicity, both 3D and clinical data must be taken into account. It was also discussed in earlier chapters that a limited number of studies have explored the combination of CT, dose data and clinical features, but despite the efforts to integrate data with neural networks, the complexity of these models presents challenges in explaining the relationship between input data and toxicity.

In this chapter, the model outlined in Chapter 5 is expanded to evaluate three different aspects of bowel toxicity; as well as bowel urgency, faecal incontinence and diarrhoea are also evaluated. In addition, the model is further expanded by incorporating clinical data alongside the dose and imaging dataset to achieve two goals: (i) predicting three types of induced bowel toxicity and (ii) identifying potential clinical risk factors.

6.2 Dataset

Based on the EORTC standard Quality of Life questionnaires [1], [62], [167], [61], [155] the patient's responses to three questions related to bowel urgency ("when you felt the desire to move your bowels, did you have to rush to get to the toilet?"), diarrhoea ("have you had diarrhoea?") and faecal incontinence ("have you had faecal incontinence?") were extracted (see Table 3.5). Responses to the questions were based on the severity of the symptom as follows: grade zero for "not at all", grade one for "a little", grade two for "quite a bit" and grade three for "very much". For the purposes of this chapter, patients were stratified into two classes: without toxicity for grade= 0, and with toxicity for grade≥ 1.

3D CT scans and 3D dose distributions were collated for the study cohort. 'Bowel Bag' was masked for each patient (see Table 3.5). Furthermore, 22 clinical features relevant to bowel toxicity were chosen (see Table 3.6 and Section 3.3). For each toxicity, 40 randomly selected patients (20 with and 20 without toxicity) were left aside for a test set and the reminder were used for training and validation sets. Table 6.1 shows the number of patients for each experiment. Data augmentation was only applied to the training set.

Table 6.1: Number of patients included in the experiments

Toxicity	Grade=0	Grade≥ 1	Excluded	Train	Test	Augmented
Bowel Urgency	65	175	75	200	40	110
Diarrhoea	204	101	10	265	40	100
Faecal Incontinence	219	84	12	263	40	135

Note: total number of patients available in the original dataset was 315. Train + Augmented sets were used for model training.

6.3 Neural network for combining image, dose and clinical data

The MIL-Att model was expanded with a new path for the analysis of clinical data (MIL-Att-H). Fig.6.1 shows the new architecture of the proposed model. 22 clinical features were passed

through the attention module γ, which is a fully-connected neural network. The goal of this module is to explore whether each clinical feature is important regarding toxicity. Having $\boldsymbol{h}_3$ as the feature vector presenting clinical data ($\boldsymbol{h}_3 \in \mathbb{R}^{1\times 22}$), the output of attention module γ (weights for clinical feature) n are as follows:

$$\gamma_n = \frac{\exp\{(\mathbf{M}\boldsymbol{h}_3)_n\}}{\sum_{j=1}^{22} \exp\{(\mathbf{M}\boldsymbol{h}_3)_j\}}. \tag{6.1}$$

Considering $\circ$ as element-wise product, the weighted feature vector $\boldsymbol{s}^C$ can be computed as:

$$\boldsymbol{s}^C = \gamma \circ \boldsymbol{h}_3. \tag{6.2}$$

Unlike MIL-Att, where there was only one feature vector (extracted from CT and dose), in the new model (MIL-Att-C), there are two feature vectors for the final classification. The concatenation of these two vectors can be considered as the final feature vector representing the input data. The weighted average of extracted features was used to compute the final feature vector for image data (CT and dose) as follows: $\boldsymbol{s}^{\boldsymbol{I}} = \sum_{k=1}^{K} \alpha_k * \boldsymbol{z}_k$ (see Section 5.2.3). With the network configuration discussed in the previous chapter, $\boldsymbol{s}^{\boldsymbol{I}}$ had a dimension of 1×5400, while $\boldsymbol{s}^C$ had a dimension of 1×22. Training the network with the concatenation of these two vectors resulted in the model refusing to consider clinical data. The γ weights were small numbers (close to zero) that did not show any significance for clinical data. To avoid this issue, a fully-connected module l_μ was used to reduce the dimension of $\boldsymbol{s}^{\boldsymbol{I}}$ before concatenation. Consequently, the setting for the classification module (g_φ) was changed, and the number of layers decreased in it to two. The feature vector $\boldsymbol{s}^{\boldsymbol{I}}$ had the form as below:

$$\boldsymbol{s}^I = l_\mu(\sum_{k=1}^{K} \alpha_k \boldsymbol{z}_k). \tag{6.3}$$

Considering $\boldsymbol{s}$ as the concatenation of $\boldsymbol{s}^{\boldsymbol{C}}$ and $\boldsymbol{s}^{\boldsymbol{I}}$, the final output of the model can be written

as: $y = g_\varphi(\boldsymbol{s})$, where y is the predicted label and (g_φ) is the classification block. The MIL-Att-C network (Φ_Ω) is formulated with:

$$\begin{aligned} y &= g_\varphi(\boldsymbol{s}) = \Phi_\Omega(x), \\ g_\varphi &: \boldsymbol{s} \mapsto [0,1], \quad \Omega = \{\theta, \mu, \varphi, \boldsymbol{w}, \mathbf{V}, \boldsymbol{q}, \mathbf{R}, \mathbf{M}\}. \end{aligned} \tag{6.4}$$

Let $t \in \{0,1\}$ be the target class label for patient's data x, the network is trained by minimizing the binary cross-entropy loss function as:

$$L(t, \Phi_\Omega) = -t \log(\Phi_\Omega) - (1-t)\log(1-\Phi_\Omega) \tag{6.5}$$

The loss function is summed over all inputs from the training set and minimization is performed w.r.t. Ω parameters.

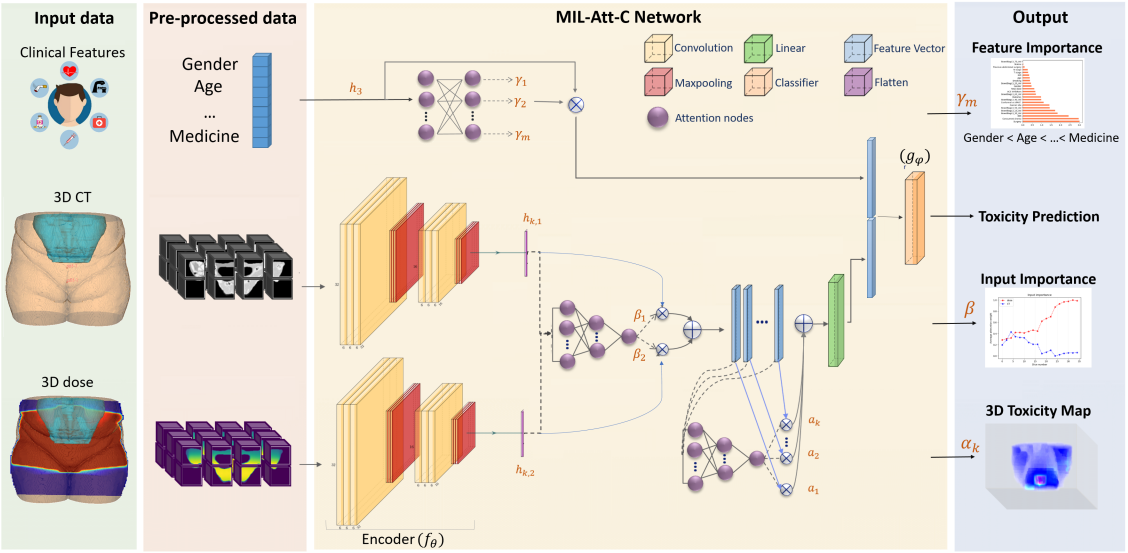


Figure 6.1: Architecture of multiple instance learning with attention modules (MIL-Att) network with three paths for analysing clinical data, 3D CT scans and 3D dose treatment plans.

6.4 Model training

To avoid overfitting, transfer learning with the same methodology as described in the previous chapter was applied to the image dataset. The two pre-trained autoencoders' weights were transferred to the model encoders. The clinical data were repeated for the newly generated (augmented) data.

In order to determine which input data are more informative for prediction, the model was trained using three different strategies as below:

- only clinical features (metadata): only the path with the clinical data was trained and CT and dose paths were disabled (MIL-Att-M).
- spatial features (images): two paths analysing CT and dose were trained (MIL-Att-I)
- combination of clinical and spatial paths were trained (MIl-Att-C)

6.5 Traditional machine learning models

Conventional methods for modeling radiotherapy outcomes solely rely on clinical data to predict toxicity. Recently, deep learning has been widely used in researches with the aim of improving the prediction performance. In order to investigate the impact of deep learning, a separate analysis of the 22 clinical features using common machine learning models (LR, SVM, RF) was conducted (see Chapter 4). Since one of the primary objective of this chapter was to identify clinical risk factors, the three machine learning models were evaluated in terms of their ability to detect risk factors and compared to the proposed model.

6.6 Experimental results

6.6.1 Prediction performance

Fig.6.2 shows the experimental results for accuracy and AUC. When examining bowel urgency and faecal incontinence, it is seen that combining CT imaging, dose distributions, and clinical information improved the prediction accuracy and AUC; as training the model with just CT scans and dose distributions for bowel urgency and faecal incontinence prediction achieved 80% and 70% accuracy, respectively, while this increased to 85% and 75% when clinical features were also added (with corresponding results for AUC).

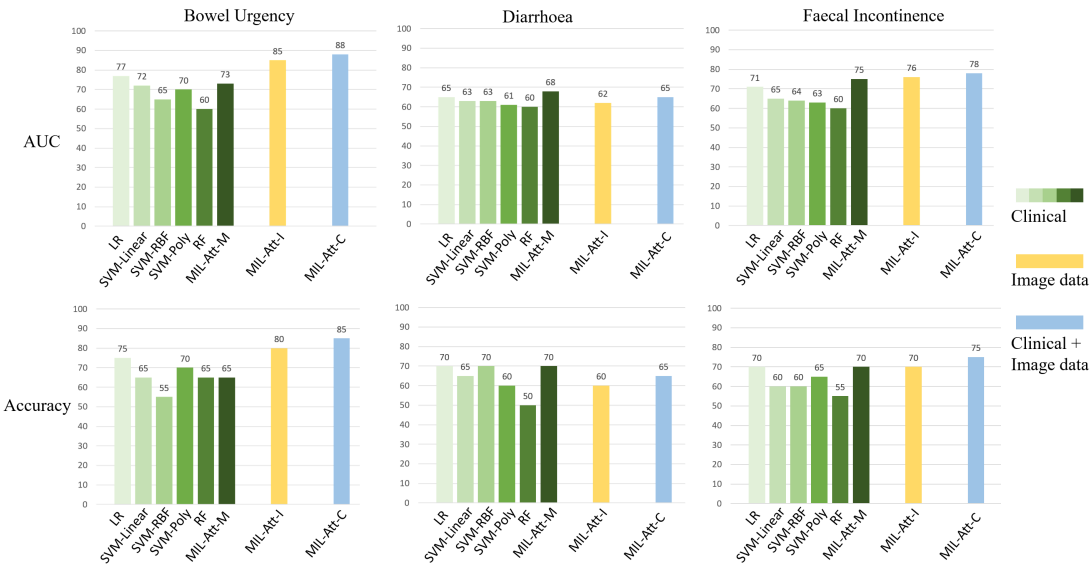


Figure 6.2: Comparison of prediction performance for various models for different toxicities. **Abbreviations**: AUC: area under the receiver operating characteristic curve; LR: logistic regression; RBF: radial basis function; MIL-Att: multiple instance learning network with attention; MIL-Att-M: network trained with clinical data; MIL-Att-I: network trained on CT and dose data; MIL-Att-C: network trained with combination of clinical data, CT scans and dose plans.

In comparison, training both paths of the model for diarrhoea produced a lower AUC and accuracy than just training with the clinical data. Bowel urgency had the highest values for accuracy and AUC, while diarrhoea had the lowest values for both.

When the same clinical data were assessed using different classification models (as shown by the green bars in Fig.6.2), the results indicated that logistic regression and the neural network (MIL-Att-C) achieved slightly higher values than SVM and random forest models.

To assess the significance of performance improvement, we conducted the DeLong test (see Fig.6.3).The combination of the model significantly enhanced predictions for bowel urgency and faecal incontinence in comparison to traditional ML models (p value < 0.05). However, for diarrhoea, the statistical improvement between LR and deep learning models was not significant.

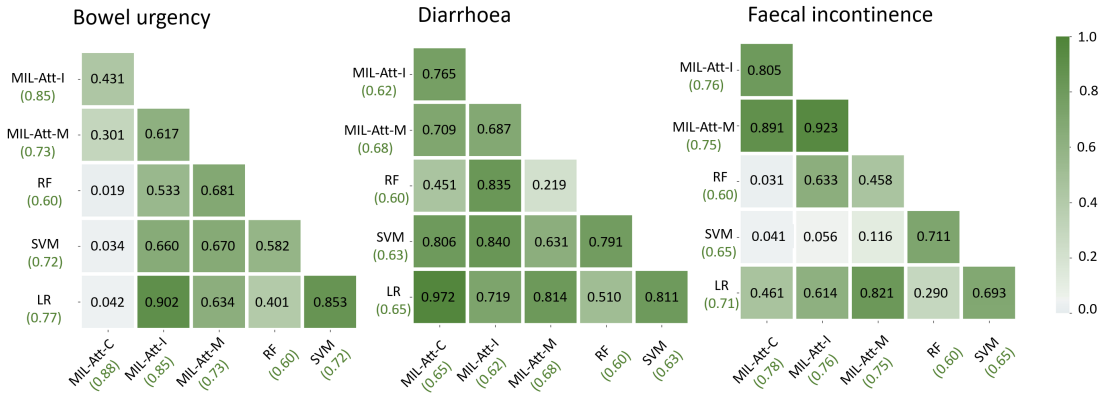


Figure 6.3: p value map of DeLong test between the different models in the test sets. AUC values are in parentheses.

6.6.2 Toxicity risk map - analysis of α weights

The results of the α attention module show the importance of different cubes in the bowel bag (the higher the weight, the more important the cube is). For each patient, the attention weights α were used to construct a toxicity risk map.

Fig.6.4 illustrates the toxicity risk map for two randomly selected patients with bowel urgency toxicity. Similar results to the previous model analysing only image data were acquired; the attention weights are concentrated on the right iliac fossa of the bowel bag for the prediction of bowel urgency, even when taking clinical factors into account.

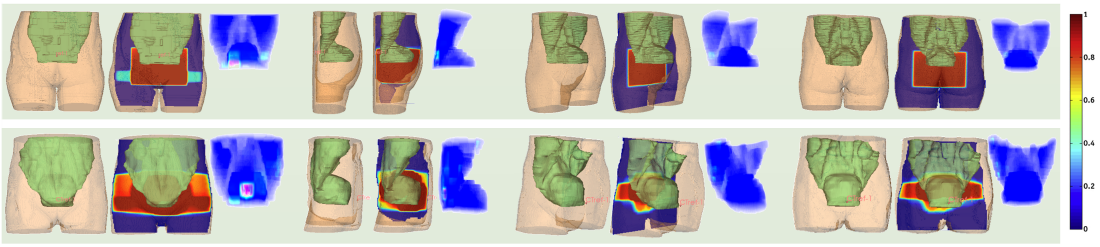


Figure 6.4: Examples of the toxicity risk map generated for bowel urgency. For each view, from left to right, the first and second images are the patient's bowel bag structure and the radiation dose distribution, respectively. The third image is the toxicity map generated by the model. The higher value of the toxicity map indicates a higher importance for the risk of developing toxicity.

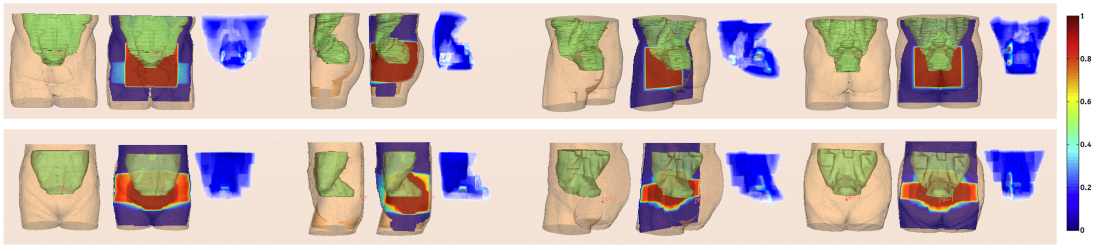


Figure 6.5: Examples of the toxicity risk map for diarrhoea. For each view, from left to right, the first and second images are the patient's bowel bag structure and the radiation dose distribution, respectively. The third image is the toxicity map generated by the model. The higher value of the toxicity map indicates a higher importance for the risk of developing toxicity.

In contrast, no clear anatomical region could be identified from the attention weights for prediction of diarrhoea (see Fig.6.5), while for faecal incontinence the toxicity was predicted by attention weights in the postero-inferior region (i.e. corresponding to the anorectum; see Fig.6.6).

The toxicity maps observed in patients vary significantly, reflecting the real-world diversity of the study cohort, including different anatomical variations such as size, BMI, and gender. To further understand the correlation between toxicity and anatomical region, a toxicity atlas was created using the maps generated by the proposed network with the same approach as 5.3.6. Fig.6.7 shows the generated atlas for each type of toxicity. The attention atlas for bowel urgency revealed that cubes located in the anterior and right iliac fossa of the bowel gained the highest

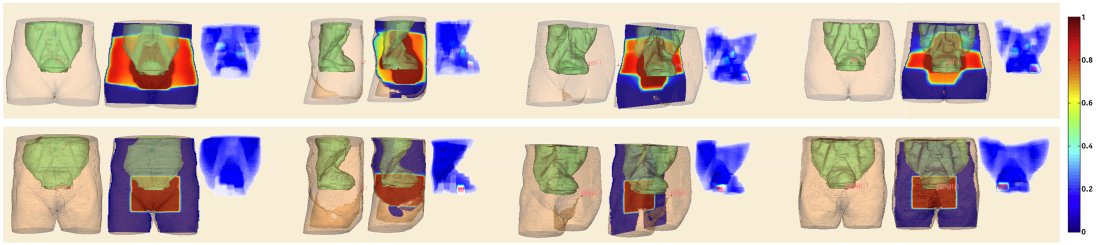


Figure 6.6: Examples of the toxicity risk map for faecal incontinence. For each view, from left to right, the first and second images are the patient's bowel bag structure and the radiation dose distribution, respectively. The third image is the toxicity map generated by the model. The higher value of the toxicity map indicates a higher risk of toxicity.

attention, which matches the findings of initial model when only CT scans and dose plans were analysed to predict mild/moderate bowel urgency toxicity. Atlas for diarrhoea shows that the attention weights were scattered throughout the bowel bag, while for faecal incontinence, the postero-inferior area stood out as the location with the most attention.

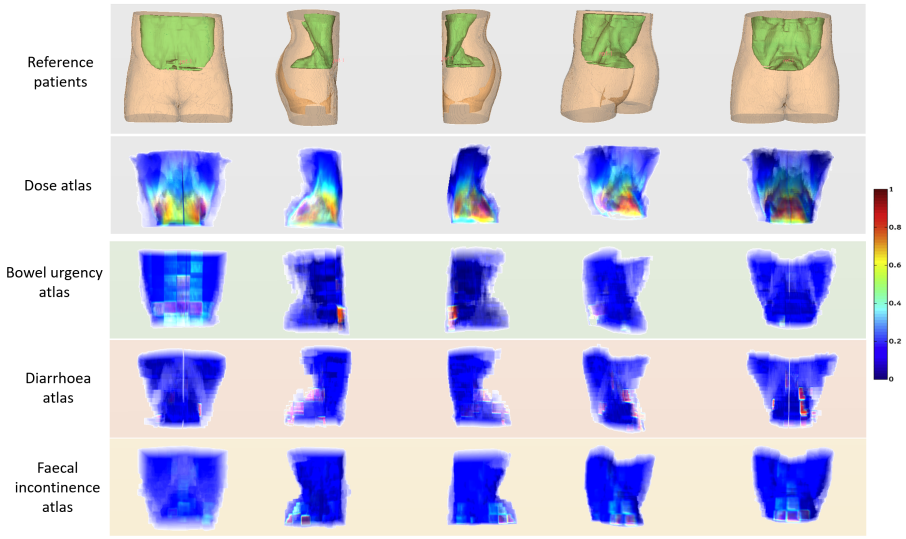


Figure 6.7: Toxicity atlas generated by the proposed model.

6.6.3 Importance of CT and dose - analysis of β weights

The β weights provide insight into how much impact each image modality has on toxicity prediction. By computing the average weight for each slice in both CT and dose plan, a better understanding of their importance can be achieved. Fig.6.8 visualises the relative significance of CT and dose for their respective 2D slices. In general, for bowel urgency and diarrhoea, dose had more of an effect on the final prediction, while for faecal incontinence, the caudal parts of the pelvis (slice number < 15) had higher weights for dose data and the cranial (upper) parts had higher values for CT data.

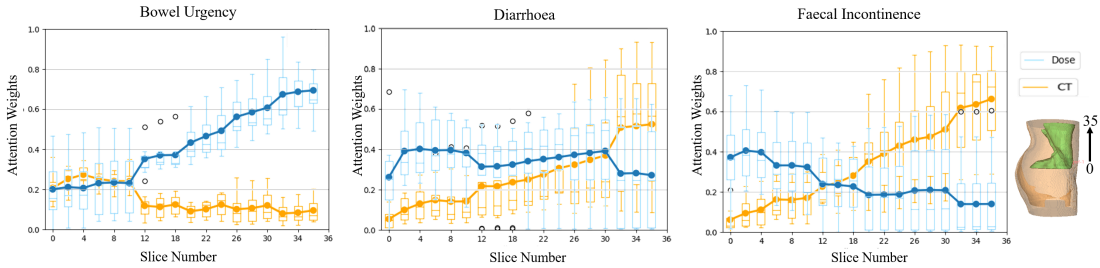


Figure 6.8: Quantitative evaluation of image association. Higher value of attention weight shows higher importance for toxicity prediction.

6.6.4 Detecting risk factors - analysis of γ weights

To gain a better understanding of how each clinical variable is correlated with toxicity and to provide more information for the outcome prediction, the 22 clinical variables were jointly trained with the spatial data. Additionally, three ML models were trained with the clinical features alone as a comparison to the CNN model, which incorporates spatial information. The results of this comparison are shown in Fig. 6.9 (for a better representation, only the top fifteen features are presented).

When exploring bowel urgency, various models identified distinct sets or orders of features as the most significant variables. Nevertheless, the top ten features shared several elements, including BMI and cancer type which ranked among the first five important features in all

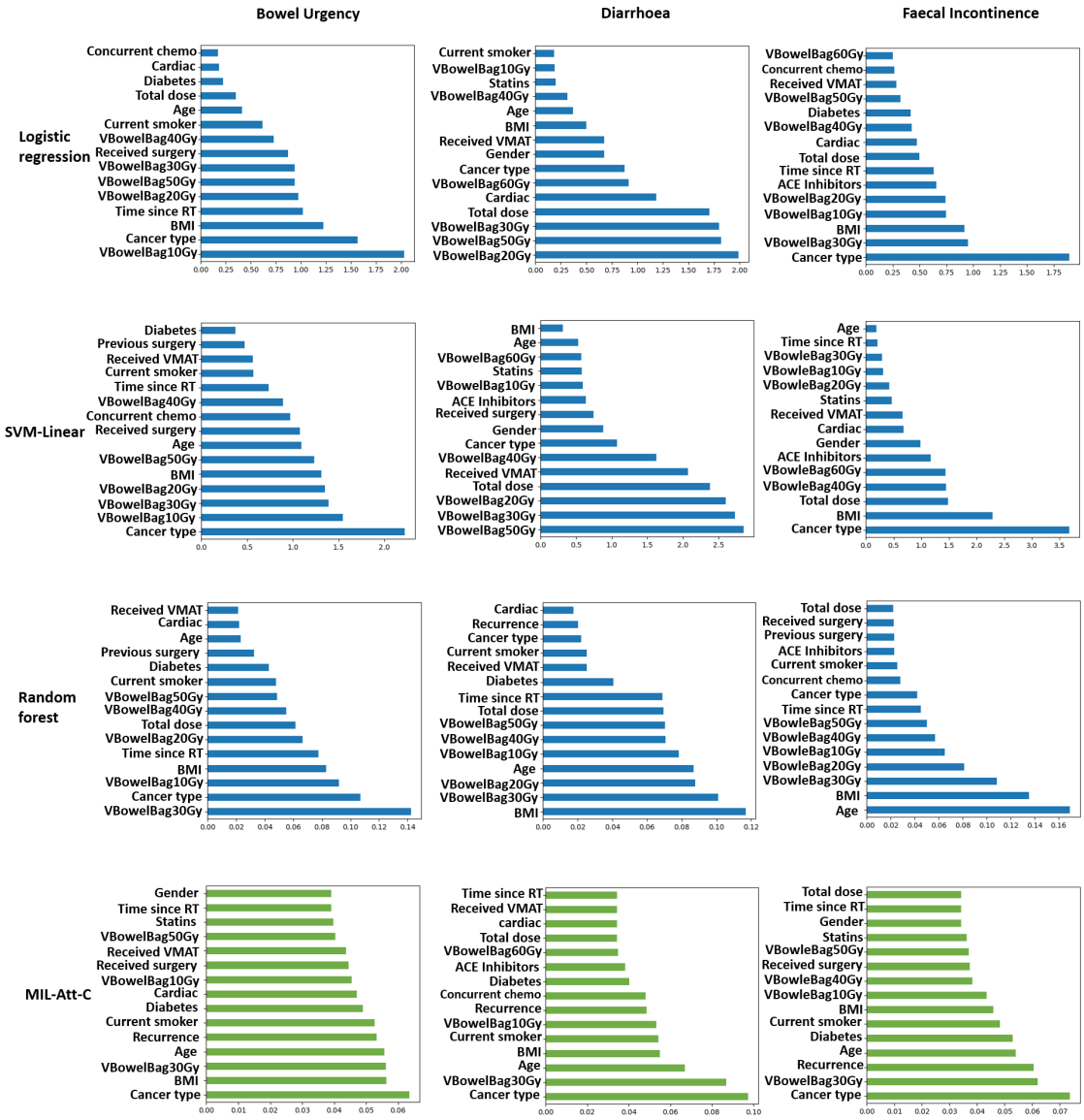


Figure 6.9: Analysis of risk factors. Fifteen most important features for LR, SVM, RF and the proposed model are extracted. The x axis presents the importance of features: for LR and SVM the coefficients of the model, for RF the mean decrease in impurity and for MIL-Att-C the gamma weights present the importance of each feature.
Abbreviations: BMI, body mass index; VMAT, Volumetric modulated arc therapy; RT, radiotherapy; ACE, Angiotensin-converting enzyme; SVM, support vector machine. **Note:** Total dose denotes the total prescribed dose.

models. Additionally, VBowelBag10Gy, VBowelBag30Gy, time since RT, smoking status, and diabetes were also recurrent across the models.

Regarding diarrhoea, dosimetric features, which capture the total dose or volume of the irradiated bowel bag (VBowelBagXGy), were significant for the ML models. In contrast, MIL-Att-M only included two dosimetric features (VBowelBag30Gy, VBowelBag10Gy) in the top 10 list. Cancer type, age, BMI, and VMAT were all included in the list for all models as non-dosimetric factors.

The plots for faecal incontinence indicate that almost half of the features were shared among the models. Cancer type was one of the top five features (except for RF, where it ranked ninth), similar to bowel urgency, and BMI had relatively high weights in all models. Other features such as total dose, VBowelBag30Gy, VBowleBag10Gy, and time since RT were also among the top 15 factors for all models. Moreover, ACE inhibitors were assigned high values in LR, SVM, and RF, but not in MIL-Att-C.

6.7 Discussion

Toxicity prediction models in conventional radiotherapy primarily rely on 1D data extracted from dose-volume histogram data (sometimes supplemented with clinical factors), and these models do not consider spatial information. In this study, a 3D CNN that analyses CT imaging, dose distributions, and clinical features together to predict RT-induced toxicities has been explored. By combining attention mechanisms with multiple instance learning, the model can visually explain toxicity distribution and identify potential risk factors. The experimental results for prediction performance show that models integrating spatial data (dose and CT) outperformed "traditional" machine learning models for two of the three toxicities. Specifically, the model reveals a possible differential spatial dose dependence for various bowel symptoms, likely reflecting different underlying pathophysiologies.

As reviewed in Chapter2, machine learning and deep learning techniques have demonstrated promising potential in toxicity prediction issue. Most published studies have utilised CNNs to analyse the inherent patterns among spatial variables in CT and dose data ([75],[172], [112]), and there are few that consider other imaging modalities (e.g., ventilation scintigraphy [15], positron emission tomography [163], and MRI [162]). Although these studies suggest that incorporating additional imaging data (beyond dose information) can enhance predictive accuracy, they do not provide a clear explanation of how these inputs can improve predictions; or quantify the importance of each input in relation to toxicity.

Moreover, predictive performance can be improved not only by adding imaging data but also by incorporating clinical features. Although there are a few studies that have reported deep learning models combining dose distribution and imaging data with clinical features ([77], [166]), the specific significance of each clinical feature in these models remains unclear.

To address these issues, a 3D CNN that simultaneously analyses CT scans, dose distribution, and patient clinical data has been explored to predict bowel toxicities. This proposed approach offers a solution to the issue of interpretability by employing attention layers that help explain how the network arrives at its final output.

Regarding bowel urgency, the results of α weights showed that the anterior and right regions of the bowel bag were associated with higher rates of toxicity. The generated atlas for faecal incontinence demonstrated that attention weights were concentrated in the postero-inferior regions. From a clinical standpoint, this pattern is intuitive and reflects the symptoms of anorectal toxicity.

Regarding diarrhoea, the atlas produced did not reveal a distinct association with any specific anatomical region. This could be due to the fact that diarrhoea is a complex concept, consisting of various bowel symptoms including stool consistency, frequency, urgency, and incontinence. These multifaceted symptoms may have diverse underlying mechanisms, and thus could be linked with different parts of the bowel rather than being confined to a particular region.

The analysis of β weights for diarrhoea showed that the dose distribution gained slightly higher attention for most slices, indicating that diarrhoea symptoms are mostly dependent on the dose received by the bowel. However, for the last slices (cranial parts), CT gained slightly higher attention, with considerable variation across the patient cohort, potentially indicating that neither dose nor CT are of particular importance in this part of the bowel volume. Regarding faecal incontinence, the analysis of β weights suggests that the dose delivered to the lower anorectum area is highly related to toxicity, with limited dependence on dose to other regions. For bowel urgency, the the analysis for β is the same in previous chapter.

By incorporating the attention module γ into the MIL-Att network, the model was able to determine which clinical variables were most significant in predicting the outcome. The comparison of γ weights with the feature importance obtained by traditional ML models indicated that they largely agreed on the top 15 important features. However, the ML models placed greater importance on dosimetric features than the MIL-Att-C network did. This suggests that dosimetric features are generally strongly associated with toxicity, and the neural network may be extracting these features independently from the dose distribution, resulting in lower γ weights for these features in the MIL-Att-C model.

Combining clinical data with CT and dose plans resulted in improved prediction performance for bowel urgency and faecal incontinence, as evidenced by higher AUC and accuracy scores. However, for diarrhoea, the model trained solely on clinical data outperformed the one trained on combined data. Perhaps while training with CT, dose and clinical data, the network stuck in a local minimum which is not as good as the optimum point in the cost function of training network with clinical data.

Comparing different input data for prediction performance in diarrhoea shows that clinical variables may have a stronger association with toxicity compared to dose or CT scans. However, the network's prediction accuracy for diarrhoea symptoms was the lowest among the three types of toxicities. This may be due to the fact that patient-reported diarrhoea describes multi-

ple symptoms as previously described, reflecting a complex underlying cause that depends on several factors. As a result, the relationship between dose distribution and the risk of diarrhoea may be less reliable, as observed in this experiment.

In general, the results showed that combining imaging, dose and clinical data mainly outperformed training on dose and CT. In an additional experiment, the impact of using only dose images in predicting toxicity compared to using dose and CT images was investigated. A model was trained with a single path using only dose images as input. The results showed that incorporating both dose and CT images also improved prediction performance. The details of this experiment are provided in Appendix A, as it falls outside the scope of this chapter.

The model achieved the highest accuracy in predicting bowel urgency, which is attributed to two factors: firstly, the relationship between anatomical dose distribution and bowel urgency is more straightforward; secondly, the dataset for diarrhoea and faecal incontinence was less balanced than that of bowel urgency; the number of patients with bowel urgency toxicity in the dataset is higher than those with either diarrhoea or faecal incontinence. This likely impacted the efficiency of network training and consequently model performance for these symptoms.

6.8 Summary and conclusions

In this chapter, a model to jointly analyse image data (3D CT scans and 3D dose distribution plans) and clinical features for the prediction of radiation-related toxicities has been presented. The main contribution of this is the ability to explain the network's behaviour in three different ways:

- analysis of α attention weights to identify the distribution of toxicity in different anatomical regions.
- analysis of β weights to determine the importance of CT and dose for prediction.
- analysis of γ weights to investigate the possible risk factors for patients.

Furthermore, experiments were conducted on different input data to analyse their predictive performance for three types of toxicities.

The model described in this chapter could in the future help clinicians in gaining a more comprehensive understanding of the factors that affect treatment outcomes as well as identifying potential risk factors for toxicity in patients. By analysing various data inputs and identifying key features associated with toxicity, clinicians can design an optimised dose treatment plan and minimise the risks of toxicity for each patient.

A comprehensive summary of these contributions, as well as the overall summary of this thesis, can be found in the next chapter. In that chapter, I will elaborate on the research itself, its objectives, limitations, and areas for potential improvement. Additionally, I will outline the direction of this research and discuss future works that can build upon the findings of this thesis. This chapter will provide a holistic view of the research, encapsulating its significance and paving the way for further advancements in the field.

Chapter 7

Conclusions

7.1 Summary and Achievements

This thesis presented a novel deep learning model for prediction of toxicity following external beam radiotherapy. The main objective of the study was to develop a workflow that could tackle some of the challenges involved in the current approaches for predicting toxicity outcomes.

The current conventional approaches for toxicity prediction mainly analyse uni-dimensional (1D) dosimetric data, which does not account for spatial information. In recent years, a number of deep learning models have been proposed as an alternative approach for toxicity prediction [103]. Unlike conventional methods, these models utilise 3D dosimetric data that include spatial information, providing a more comprehensive representation of the radiation dose distribution. However, the complexity of deep learning models poses a significant challenge for their application in real-world clinical problems. Several of the recent models used gradient-based methods to identify the most significant features to overcome this limitation. However, these methods often generate approximations of the patient's anatomy and lack clarity in their association with specific anatomical structures. To address this issue, a deep learning model that could explore spatial information in 3D CT scans and dose distributions was proposed.

The novelty of this work is that it provides an explanation for the final decision of the network by highlighting specific anatomical regions associated with a higher risk of toxicity.

Moreover, the predictive capability is not limited to spatial information alone; clinical metadata can also influence it. While there have been a few works that have addressed combining 3D imaging and dose data with clinical metadata, they still lack explainability. To address this, the proposed model was expanded with a new path, in which it incorporated clinical data along with CT and dose data. This increased the prediction power of the model. Employing an attention module along with the new path, the model could identify the risk factors that are associated with toxicity outcomes.

The predictive performance of the proposed network was evaluated by comparing it with three different analyses. Also, the attention risk map generated by the proposed model was compared with the Grad-CAM map to assess its ability to detect high-risk anatomical regions. In addition, a comprehensive analysis of three common machine learning models, LR, SVM, and RF, was conducted to evaluate the proposed deep learning model ability to identify clinical risk factors and prediction. Comparing different analysis techniques with and without clinical data was helpful because it allowed to identify the contribution of clinical data in improving the performance of the model for toxicity prediction. Through this comparison, the relative importance of dosimetric and clinical features in predicting toxicity was determined. Overall, the comparative experimental results demonstrated the proposed frameworks have great potential for assisting clinicians with outcome prediction. In the following, the major contributions of the thesis are outlined.

In Chapter 2, several published work that employed deep learning techniques for radiotherapy toxicity prediction were reviewed. By carefully analysing the findings of these studies, valuable insights into the current state-of-the-art methods for toxicity prediction as well as the key challenges and opportunities, including generating toxicity map or combining and analysing different data type, in this field were gained. In particular, the various approaches taken by re-

searchers to address issues such as data imbalance and model interpretability were examined.

In Chapter 5, a framework to analyse image and dose data for moderate/severe bowel urgency prediction was proposed. This work has been the first published attempt that can explain the network's decision procedure in terms of detecting anatomical regions associated with toxicity and how informative the different kinds of 3D information (here CT and dose) are for outcome prediction.

In Chapter 6, the framework was expanded to combine 3D imaging and dose data with clinical features to predict three symptomatic side effects after pelvic RT: bowel urgency, diarrhoea, and faecal incontinence. The new architecture included a path for clinical data, along with a separate attention module that detects the importance of each feature for the predicted toxicity. This study has been the first attempt to explicitly provide an estimate of the importance of each clinical factors while combining 3D imaging data with clinical features.

From a clinical perspective, this study concludes that for bowel urgency toxicity, the anterior and right iliac fossa regions of the bowel bag are correlated with the risk of toxicity. For faecal incontinence, the postero-inferior regions may be associated with risk of toxicity. However, for diarrhoea, no specific anatomical regions were identified. The analysis of beta weights revealed that CT scans can provide information for predicting toxicity and the informative slices of the CT scan varied depending on the specific type of toxicity being predicted. Moreover, the analysis of the gamma weights indicated that cancer type, BMI, and dosimetric features are clinical risk factors for bowel urgency and faecal incontinence. For diarrhoea, while summary dosimetric features played a crucial role in correlating with toxicity, cancer type, age, BMI, and VMAT also emerged as significant factors.

7.2 Publications

The following publications were derived from work presented in this thesis:

Journal papers:

- Ane L Appelt*, **Behnaz Elhaminia***, Ali Gooya, A lexandra Gilbert, Mike Nix, "Deep learning for radiotherapy outcome prediction using dose data–a review", Clinical Oncology (2022), doi: 10.1016/j.clon.2021.12.002.
  *Joint first author
- **Behnaz Elhaminia**, Alexandra Gilbert, John Lilley, Moloud Abdar, Alejandro F Frangi, Andrew Scarsbrook, Ane Appelt, Ali Gooya. "Toxicity Prediction in Pelvic Radiotherapy Using Multiple Instance Learning and Cascaded Attention Layers." in IEEE Journal of Biomedical and Health Informatics (2023), doi: 10.1109/JBHI.2023.3238825.
- **Behnaz Elhaminia**, Alexandra Gilbert, Andrew Scarsbrook, John Lilley, Ane Appelt, and Ali Gooya. "Deep learning combining imaging, dose and clinical data for predicting bowel toxicity after pelvic radiotherapy" submitted to Medical Physics (under review).

Conference paper:

- **Behnaz Elhaminia**, Alexandra Gilbert, Alejandro F Frangi, Andrew Scarsbrook, John Lilley, Ane Appelt, Ali Gooya."Deep learning with visual explanation for radiotherapy-induced toxicity prediction", SPIE Medical Imaging, February 2023, doi:10.1117/12.2652481.

7.3 Limitations and areas for improvement

Despite the progress made by the work presented in this thesis, there are still some concerns that need to be addressed. Firstly, it should be acknowledged that the data availability in the

current dataset are limited. This can significantly impact the training of the network as well as the generalisability of the model and its ability to perform well on unseen datasets. To address this limitation, further training on a larger dataset can potentially increase the model's performance and enhance its generalisability. In the future works, testing the model on a completely independent and (preferably) external dataset - i.e. test performance and the generated attention maps on an independent dataset, without further training can be performed.

Secondly, there is no established ground-truth for the toxicity risk map, which makes it difficult to quantify the effectiveness of the attention risk map in terms of localisation. However, at this stage, the only possible option was to seek expert opinions to validate the clinical relevance of the findings. This can provide valuable insights and help to determine how the model's predictions align with clinical information. In the future, it may be possible for clinicians to establish a standard approach to create ground-truth labelling. This would provide a more objective measure to evaluate the performance of the model.

Thirdly, the γ attention weights revealed the relationship between each clinical feature and toxicity, but it remains unclear how different classes within these features impact the outcome. For instance, the network recognised BMI as a risk factor, but it is not clear if a high BMI carries more influence in causing side effects than a low BMI. To address this issue, the average γ weights for each class in each clinical feature (see Appendix B) is computed. Nonetheless, more analysis is necessary to provide a conclusive answer. This issue can be the focus of the future work.

Fourthly, cross-validation was not feasible for our dataset for two reason; (i) for the purpose of transfer learning, we trained an autoencoder on our dataset. Then, we transferred its learned weights to the encoders of a MIL-Att network. In cross-validation, the dataset is divided into n folds and, one fold, is selected as the test set. Then in each iteration, the model is evaluated based on that test set. This means that cross-validation tests the model on all the data in the dataset. Now, if we performed cross-validation on our dataset, the data that are in the

test fold have been already seen by the autoencoders. This would make the cross-validation results inaccurate. One solution would be to remove the data in the selected fold and train the autoencoder with the rest of the data, then transfer the weights from the autoencoder to the MIL-Att network and train the MIL-Att network. since this process should be repeated in each iteration, it is computationally expensive and not feasible. Instead, we selected 40 patients from the beginning and trained the autoencoders without those 40 patients. We then tested the model on those 40 patients, which were completely unseen data. (ii) the number of patients with positive labels is generally small in our dataset. If we divide the dataset into different folds for cross-validation, we must have at least one positive label in each fold to be able to evaluate the performance. Considering that we performed data augmentation, it is possible that the positive data in the test fold is an augmented version of real data in the training folds. This evaluation is also not accurate because the data in the test is already in the train set. However, to provide more comparison, the DeLong test was examined and reported.

Regarding the Delong test, it is important to acknowledge that its primary purpose is to compare empirical AUC values for different models. However, interpreting the results poses challenges. The DeLong's test is specifically designed to compare the AUC of two or more correlated models, employing calculations to conduct a statistical analysis and ascertain if observed differences in AUC are statistically significant. An issue that arises is the potential for the multiple comparisons problem; since the statistical test is conducted on a sample rather than the entire population, there is a risk of finding something statistically significant purely by chance with a sufficient number of tests. This phenomenon is known as the type I error, where the chance of incorrectly declaring an effect due to random error in the sample is present. To address this, the p-value of each individual test should be adjusted upwards, ensuring that the overall error rate for all tests remains at 0.05. This precaution is crucial to mitigate the risk of incorrectly identifying a significant difference. Despite the acknowledged flaws in reporting Delong's p-values and the inevitability of certain issues, it remains a standard and convenient approach in association studies. Various methods, such as Bonferroni–Holm correction [70], exist to ad-

just p-values, but careful and thorough assessment is necessary. More sophisticated statistical analysis of the results could be in the future work to further strengthen the findings.

There are also some limitations regarding the data used in this study. Firstly, most cervical cancer patients in the dataset also received brachytherapy. However, the dose delivered during this therapy was not captured in the external beam dose distribution. This limitation could affect the accuracy of the model's predictions, and it is important to find ways to integrate brachytherapy data into the model to improve its accuracy.

Secondly, typical of medical datasets, the current approach is also challenged by data missing and time to event. The dataset utilised in this study contained missing data, primarily due to that data couldn't be found in the patient's clinical records; and for toxicity data was missing as patients may not have responded to the relevant questions on the questionnaires. These missing data instances posed limitations in the analysis and interpretation of the results. In an attempt to address the missing variables, including BMI, imputation methods was employed. However, it is important to note that imputation involves estimating the missing values and it does not precisely reflect the exact values that were missing.

Thirdly, the patient grouping in the dataset was adjusted to address the imbalance issue, although for clinical purposes, it might be more beneficial if the model could predict ordinal labels or different groupings. Unfortunately, due to the limited and imbalanced nature of the data, analysing the model with ordinal and alternative groupings was not feasible. It is worth mentioning that in Chapter 5, the grouping was based on grades 0 and 1 for cases without toxicity and grades 2 and 3 for those with toxicity. Despite the inability to extend this grouping strategy for Chapter 6, the attention results remained consistent. Both grouping models indicated that toxicity was localized in the anterior and right iliac fossa.

Additionally, the toxicity grade was determined based on patient-reported grading, presenting both advantages and disadvantages. On the positive side, relying on patient self-reporting can enhance accuracy by capturing subtle symptoms not immediately apparent to healthcare

professionals. However, a drawback is the potential for variability in symptom reporting among patients, and also some patients may not complete the questionnaires consistently or accurately.

Finally, fairness and bias are also significant challenges; medical datasets are often collected from specific populations or patient groups, which can lead to biases in the data. For example, the dataset employed in this study includes more female patients than male patients. Therefore, it may not be representative of the general population. This can lead to biased results and may limit the generalisability of the findings. Additionally, medical datasets may also be biased due to factors such as unequal access to healthcare, differences in treatment protocols, and diagnostic errors. Several studies have been conducted to explore fairness and address associated issues [42], [111]. Additionally, there are various tools such as Python libraries like AIFair360 [11] and FairLearn [16] that have been developed specifically for bias detection within datasets. While the detection of dataset fairness was not within the scope of this PhD thesis, it presents an avenue for future research and investigation.

There are also some limitation that relates to the scope of the study. For example, the focus of study is on only three commonly reported bowel symptoms. There are other potential toxicity outcomes that could have been considered, such as urinary and sexual symptoms, which are also common following radiotherapy. Further research could potentially enhance the overall understanding of the relationship between anatomical regions and toxicity.

Another criticism that may arise regarding this study is the issue of causality. Currently, the research examines associations, specifically identifying which parts (spatial/anatomical) of the dose distributions and CT scans appear to be most strongly linked to toxicity based on their weighted importance in the model. However, it is crucial to acknowledge that these associations do not establish causation. The study should recognise this limitation and the fact that further investigations would be necessary to establish causal relationships. There are casual methods that are used to infer causal relationships between variables or factors in a system [8]. Causal methods aim to uncover the cause-and-effect relationships and understand the un-

derlying mechanisms that drive those correlations (In contrast to traditional machine learning methods that focus on predicting outcomes based on observed correlations in the data). The objective of this research did not encompass the investigation of causality, and it is a potential avenue for future work.

Overall, it is essential to address all these challenges to ensure that the model's predictions are as accurate and reliable as possible. Tackling these concerns will improve the robustness and accuracy of the model and pave the way for more effective treatment and management of radiotherapy. This can be accomplished for the future works.

A note on Chapter 2:

In Chapter 2, a review paper that summarised the published works on toxicity prediction using deep learning models was presented. However, since the publication of that paper, there have been several other studies that have employed deep learning techniques for toxicity prediction or reviewed these models in the context of radiotherapy [117], [94], [130], [145] and[175]. Therefore, it is important to acknowledge the more recent works that have contributed to the field and have extended understanding of the potential of deep learning for toxicity prediction in radiotherapy.

7.4 Further research directions

For future work, there are several potential directions to pursue. Firstly, expanding this study to other datasets could validate the findings in different populations. Secondly, a segmentation approach could replace the use of cubes in the bags, where different segments of the bowel are extracted, and multiple instance learning can be performed on these segments. This has the potential to improve the localisation of the risk map. Thirdly and the most crucial focus of future work could be to automatically generating an optimal dose plan that minimises the risk of toxicity; developing more sophisticated machine learning algorithms, such as Bayesian networks

and graph convolutional networks, along with this proposed model can create a framework to optimise the dose distribution which minimise the toxicity risk. Furthermore, the area of interest for predicting bowel related toxicities was the bowel bag, a decision made by Dr. Alexandra Gilbert, an expert in patient-reported outcomes following pelvic radiation therapy. However, a potential avenue for future research could involve analysing additional regions and areas outside the bowel bag, such as the spinal cord. Ultimately, the goal for future work can be to create a more personalised and effective approach to radiation therapy that minimises the risk of side effects for patients, and this research can have the potential to make a significant impact in the field of cancer treatment.

Appendix A

Analysis of dose data for toxicity prediction

In several studies, toxicity prediction has been analysed solely based on dose data, and traditional NTCP modeling has also focused on dose data. Thus, in this experiment, we aim to evaluate whether incorporating CT data can improve the performance of toxicity prediction compared to using only dose data.

In this experiment, we trained the model only with one path and dose distribution input data.

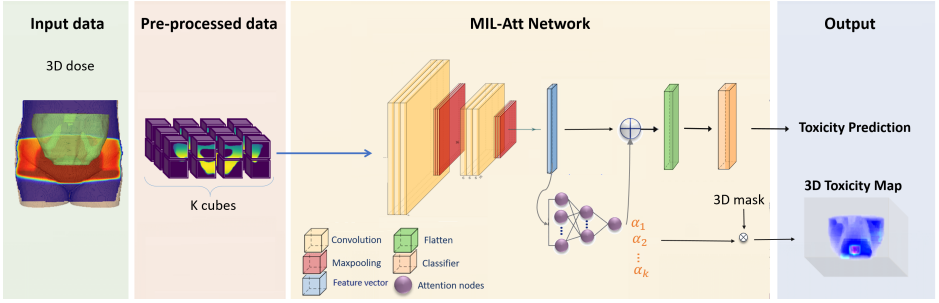


Figure A.1: The schematic illustration of the trained model. 3D input image is pre-processed and fed into the Attention-MIL network. The output of the network is a binary variable defining toxicity prediction.

Fig.A.1 shows the architecture of the trained model. The training process follows the methodology outlined in Chapter 5 and the experimental results are conducted on unseen data (test set). Table A.1 shows the results of comparison for three types of toxicities.

Table A.1: Comparison of prediction performance across different input analysis. Best performance in each metric is shown in bold.

	Bowel urgency		Diarrhoea		Faecal incontinence	
Method	Acc	AUC	ACC	AUC	ACC	AUC
MIL-Att-dose	0.70	0.76	0.60	**0.65**	0.67	0.72
MIL-Att-I	**0.80**	**0.85**	0.60	0.62	**0.70**	**0.76**

***Abbreviation:**: Acc, accuracy; AUC, area under receiver operating characteristic curve.

The comparison of the experimental results revealed that the incorporation of CT scans in addition to dose images led to a noticeable improvement in prediction performance for bowel urgency and faecal incontinence, but not for diarrhoea. These findings are consistent with the results obtained from combining imaging data with clinical variables; the observed lack of improvement in the prediction accuracy for diarrhoea after incorporating CT scans in addition to dose images could be attributed to the complexity of the symptom. Diarrhoea is a multi-dimensional symptom that includes several bowel symptoms such as stool consistency, frequency, and urgency, which makes it more challenging to identify the specific factors contributing to its development. However, it is known that the dosimetric data is closely related to the risk of diarrhoea toxicity. As a result, the inclusion of CT scans may not be as informative in identifying the factors contributing to diarrhoea toxicity as dosimetric data. Therefore, based on the findings from this comparison experiment and previous experiments, it can be concluded that incorporating CT scans and clinical variables can improve the prediction performance for bowel urgency and faecal incontinence, while the prediction of diarrhoea symptoms is mostly dependent on the dosimetric data.

Appendix B

Average of γ Weights

The attention module γ computes importance weights that reveal the neural network's focus on clinical factors in predicting toxicity. For example, for all types of toxicities, the cancer type received high attention weights. Nevertheless, it remains unclear which specific cancer subtype (rectal, cervical, endometrial, or anal) has the most substantial impact on the final output. To investigate this, we computed the average γ weights for each subcategory within each feature. The top 15 weights are presented in Figure B.1.

We can see that for bowel urgency the highest attention weight gained by rectal cancer, followed by endometrial, cervical, and anal cancers in descending order. Let's say we're looking at the relationship between rectal cancer and bowel urgency toxicity. We might assume that rectal cancer increase the probability of a patients developing bowel urgency toxicity. But this is not completely true, as the relationship between variables and the outcome are complex. The relationship between different variables and toxicity follows a structural causal model, which means that there may be hidden variables with various pathways, such as chains or forks, between them. Investigating these relationships requires a thorough study beyond the scope of our study.

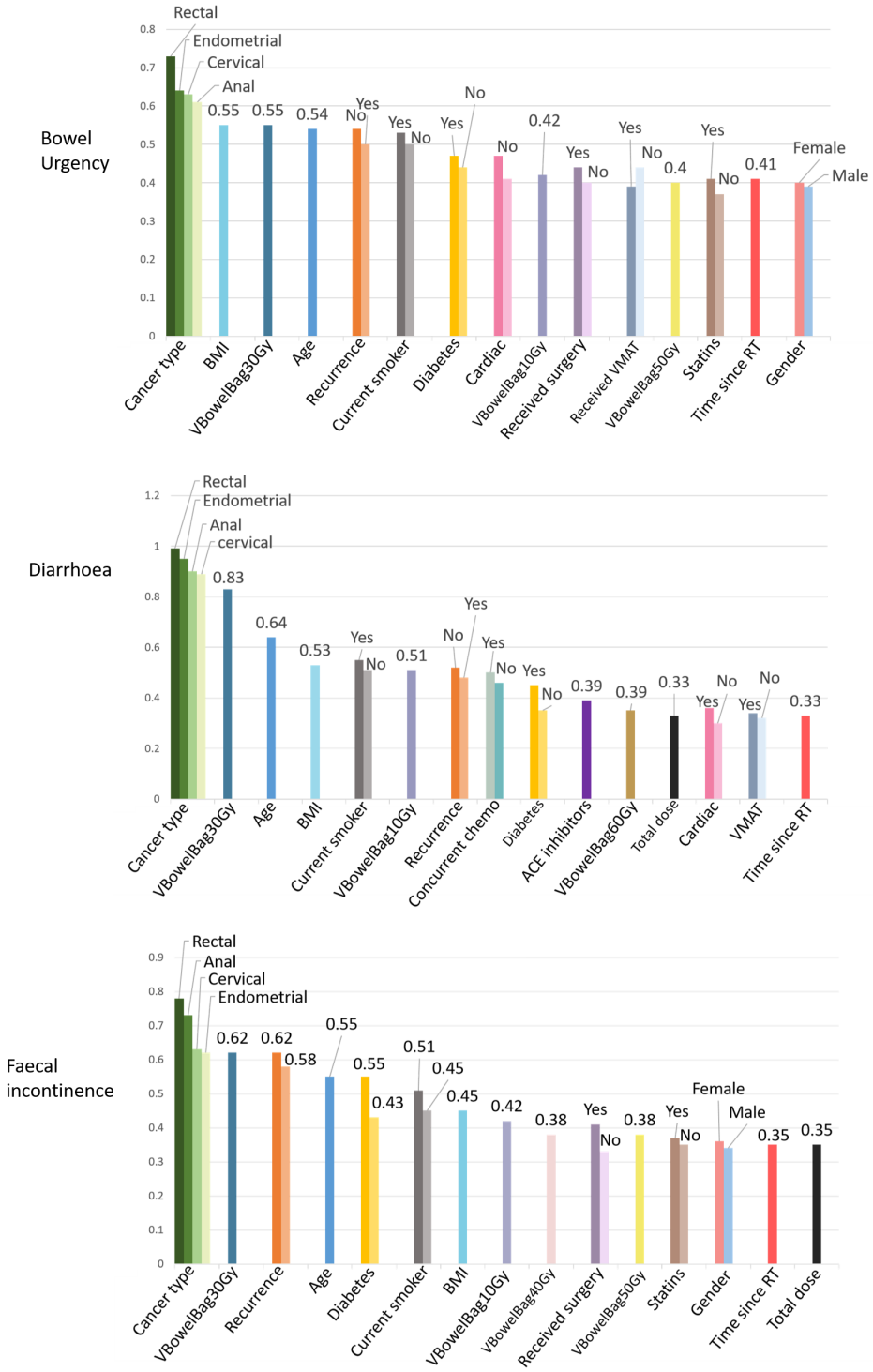


Figure B.1: The average of γ weights for top 15 important features. **Abbreviations:** BMI, body mass index; VMAT, Volumetric modulated arc therapy; RT, radiotherapy; ACE, Angiotensin-converting enzyme. **Note:** Total dose denotes the total prescribed dose.

References

[1] Neil K Aaronson et al. “The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology”. In: JNCI: Journal of the National Cancer Institute 85.5 (1993), pp. 365–376.

[2] X Allen Li et al. “The use and QA of biologically related models for treatment planning: Short report of the TG-166 of the therapy physics committee of the AAPM”. In: Medical physics 39.3 (2012), pp. 1386–1409.

[3] Jervoise Andreyev. “Gastrointestinal symptoms after pelvic radiotherapy: a new understanding to improve management of symptomatic patients”. In: Lancet. Oncol. 8.11 (2007), pp. 1007–1017.

[4] S Aneja et al. “Deep neural network to predict local failure following stereotactic body radiation therapy: integrating imaging and clinical data to predict outcomes”. In: International Journal of Radiation Oncology, Biology, Physics 99.2 (2017), S47.

[5] Antreas Antoniou, Amos Storkey, and Harrison Edwards. “Data augmentation generative adversarial networks”. In: arXiv preprint arXiv:1711.04340 (2017).

[6] AL Appelt et al. “Deep learning for radiotherapy outcome prediction using dose data–a review”. In: Clinical Oncology 34.2 (2022), e87–e96.

[7] Samuel G Armato III et al. “The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans”. In: Medical physics 38.2 (2011), pp. 915–931.

[8] Kellyn F Arnold et al. “Reflection on modern methods: generalized linear models for prognosis and intervention—theory, practice and implications for machine learning”. In: International journal of epidemiology 49.6 (2020), pp. 2074–2082.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: arXiv preprint arXiv:1409.0473 (2014).

[10] Yiftach Barash et al. “Ulcer severity grading in video capsule images of patients with Crohn’s disease: An ordinal neural network solution”. In: Gastrointestinal Endoscopy 93.1 (2021), pp. 187–192.

[11] Rachel K. E. Bellamy et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, ar Oct. 2018. URL: https://arxiv.org/abs/1810.01943.

[12] Mohamed Amine Benadjaoud et al. “Functional data analysis in NTCP modeling: a new method to explore the radiation dose-volume effects”. In: International Journal of Radiation Onc 90.3 (2014), pp. 654–663.

[13] Søren M Bentzen et al. “Bioeffect modeling and equieffective dose concepts in radiation oncology–terminology, quantities and units”. In: Radiotherapy and Oncology 105.2 (2012), pp. 266–268.

[14] Søren M Bentzen et al. “Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues”. In: International Journal of Radiation Oncolog 76.3 (2010), S3–S9.

[15] Liang Bin et al. “A deep learning-based dual-omics prediction model for radiation pneumonitis”. In: Medical Physics 48.10 (2021), pp. 6247–6256.

[16] Sarah Bird et al. Fairlearn: A toolkit for assessing and improving fairness in AI. Tech. rep. MSR-TR-2020-32. Microsoft, May 2020. URL: https://www.microsoft.com/

en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[17] Christopher M Bishop. Maximum margin classifiers. In Pattern recognition and machine learning. Vol. 4. 4. Springer, 2006.

[18] Dacian V Bonta et al. "A variable critical-volume model for normal tissue complication probability". In: Medical Physics 28.7 (2001), pp. 1338–1343.

[19] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. "Rank consistent ordinal regression for neural networks with application to age estimation". In: Pattern Recognition Letters 140 (2020), pp. 325–331.

[20] Marc-André Carbonneau et al. "Multiple instance learning: A survey of problem characteristics and applications". In: Pattern Recognition 77 (2018), pp. 329–353.

[21] Mauro Carrara et al. "Development of a ready-to-use graphical tool based on artificial neural network classification: application for the prediction of late fecal incontinence after prostate cancer radiation therapy". In: International Journal of Radiation Oncology* Biology* P 102.5 (2018), pp. 1533–1542.

[22] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: Journal of artificial intelligence research 16 (2002), pp. 321–357.

[23] Shifeng Chen et al. "Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis". In: Medical physics 34.10 (2007), pp. 3808–3814.

[24] Xinyuan Chen et al. "A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning". In: Medical physics 46.1 (2019), pp. 56–64.

[25] Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. "A neural network approach to ordinal regression". In: 2008 IEEE international joint conference on neural networks (IEEE world con IEEE. 2008, pp. 1279–1284.

[26] Alexander Chi et al. “Correlation of three different approaches of small bowel delineation and acute lower gastrointestinal toxicity in adjuvant pelvic intensity-modulated radiation therapy for endometrial cancer”. In: Technology in cancer research & treatment 11.4 (2012), pp. 353–359.

[27] Philip Chikontwe et al. “Multiple instance learning with center embeddings for histopathology classification”. In: International Conference on Medical Image Computing and Computer-Assis Springer. 2020, pp. 519–528.

[28] Convolutional Neural Network. wikimedia. 2021. URL: https://commons.wikimedia.org/wiki/File:Convolutional_Neural_Network_with_Color_Image_Filter.gif.

[29] Joaquim Pinto da Costa and Jaime S Cardoso. “Classification of ordinal data using neural networks”. In: European conference on machine learning. Springer. 2005, pp. 690–697.

[30] Sunan Cui et al. “Artificial neural network with composite architectures for prediction of local control in radiotherapy”. In: IEEE transactions on radiation and plasma medical sciences 3.2 (2018), pp. 242–249.

[31] Vittoria D’Avino et al. “Prediction of gastrointestinal toxicity after external beam radiotherapy for localized prostate cancer”. In: Radiation Oncology 10.1 (2015), pp. 1–9.

[32] Raj M Dalsania et al. “Management of long-term toxicity from pelvic radiation therapy”. In: American Society of Clinical Oncology Educational Book 41 (2021), pp. 147–157.

[33] Laura A Dawson et al. “Analysis of radiation-induced liver disease using the Lyman NTCP model”. In: International Journal of Radiation Oncology* Biology* Physics 53.4 (2002), pp. 810–821.

[34] Laura A Dawson et al. “Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation”. In: International Journal of Radiation Oncology Biolog 62.3 (2005), pp. 829–837.

[35] Dirk De Ruysscher et al. “Radiotherapy toxicity”. In: Nature Reviews Disease Primers 5.1 (2019), pp. 1–20.

[36] Jamie A Dean et al. “Normal tissue complication probability (NTCP) modelling using spatial dose metrics and machine learning methods for severe acute oral mucositis resulting from head and neck radiotherapy”. In: Radiotherapy and Oncology 120.1 (2016), pp. 21–27.

[37] Gilles Defraene et al. “The benefits of including clinical factors in rectal normal tissue complication probability modeling after radiotherapy for prostate cancer”. In: International Journal o 82.3 (2012), pp. 1233–1242.

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

[39] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: 2009 IEEE conference on c Ieee. 2009, pp. 248–255.

[40] Jeff Donahue et al. “Decaf: A deep convolutional activation feature for generic visual recognition”. In: International conference on machine learning. PMLR. 2014, pp. 647–655.

[41] Getao Du et al. “Medical image segmentation based on u-net: A review”. In: Journal of Imaging Scie 64.2 (2020), pp. 20508–1.

[42] Mengnan Du et al. “Fairness in deep learning: A computational perspective”. In: IEEE Intelligent Sys 36.4 (2020), pp. 25–34.

[43] Issam El Naqa et al. “Multivariable modeling of radiotherapy outcomes, including dose–volume and clinical factors”. In: International Journal of Radiation Oncology* Biology* Physi 64.4 (2006), pp. 1275–1286.

[44] Issam El Naqa et al. “On the fuzziness of machine learning, neural networks, and artificial intelligence in radiation oncology”. In: International journal of radiation oncology, biology, phys 100.1 (2018), pp. 1–4.

[45] Issam El Naqa et al. “Predicting radiotherapy outcomes using statistical learning techniques”. In: Physics in Medicine & Biology 54.18 (2009), S9.

[46] B Elhaminia et al. “Deep learning with visual explanation for radiotherapy-induced toxicity prediction”. In: Medical Imaging 2023: Computer-Aided Diagnosis. Vol. 12465. SPIE, 2023, p. 124651V. DOI: 10.1117/12.2652481. URL: https://doi.org/10.1117/12.2652481.

[47] Behnaz Elhaminia et al. “Toxicity Prediction in Pelvic Radiotherapy Using Multiple Instance Learning and Cascaded Attention Layers”. In: IEEE Journal of Biomedical and Health Inform 27.4 (2023), pp. 1958–1966.

[48] Bahman Emami et al. “Tolerance of normal tissue to therapeutic irradiation”. In: International Journa 21.1 (1991), pp. 109–122.

[49] Fajardo. Pyreadstat. Version 1.2.1. Sept. 24, 2018. URL: https://github.com/Roche/pyreadstat.

[50] Jiawei Fan et al. “Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique”. In: Medical physics 46.1 (2019), pp. 370–381.

[51] Abolfazl Farahani et al. “A brief review of domain adaptation”. In: Advances in data science and info (2021), pp. 877–894.

[52] Andriy Fedorov et al. “3D Slicer as an image computing platform for the Quantitative Imaging Network”. In: Magnetic resonance imaging 30.9 (2012), pp. 1323–1341.

[53] Francisco Fernández-Navarro, Annalisa Riccardi, and Sante Carloni. “Ordinal neural networks without iterative tuning”. In: IEEE transactions on neural networks and learning systems 25.11 (2014), pp. 2075–2085.

[54] Claudio Fiorino et al. “Dose–volume effects for normal tissues in external radiotherapy: pelvis”. In: Radiotherapy and Oncology 93.2 (2009), pp. 153–167.

[55] Bernd Fischer and Jan Modersitzki. “A unified approach to fast image registration and a new curvature based registration technique”. In: Linear Algebra and its applications 380 (2004), pp. 107–124.

[56] Hiram A Gay et al. “Pelvic normal tissue contouring guidelines for radiation therapy: a Radiation Therapy Oncology Group consensus panel atlas”. In: Int. J. Radiat. Oncol. Biol. Phys 83.3 (2012), e353–e362.

[57] Jacob Gildenblat and contributors. PyTorch library for CAM methods. url: https://github.com/jacobg grad-cam. 2021.

[58] Kyle J Godfrey and Michael Kazim. “Radiotherapy for active thyroid eye disease”. In: Ophthalmic Plastic & Reconstructive Surgery 34.4S (2018), S98–S104.

[59] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[60] Ian Goodfellow et al. “Generative adversarial nets”. In: Advances in neural information processing s 27 (2014).

[61] Elfriede Greimel et al. “Psychometric validation of the European organisation for research and treatment of cancer quality of life questionnaire-endometrial cancer module (EORTC QLQ-EN24)”. In: European journal of cancer 47.2 (2011), pp. 183–190.

[62] Elfriede R Greimel et al. “The European Organization for Research and Treatment of Cancer (EORTC) Quality-of-Life questionnaire cervical cancer module: EORTC QLQ-CX24”. In: Cancer 107.8 (2006), pp. 1812–1822.

[63] Sarah L Gulliford et al. “Use of artificial neural networks to predict biological outcomes for patients receiving radical radiotherapy of the prostate”. In: Radiotherapy and oncology 71.1 (2004), pp. 3–12.

[64] Demis Hassabis. Transfer learning is key to AGI. https://youtu.be/YofMOh6_WKo. [Online; accessed Mar 17, 2018]. 2018.

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

[66] Haibo He et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: 2008 IEEE international joint conference on neural networks (IEEE world congress o
IEEE. 2008, pp. 1322–1328.

[67] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. "Support vector learning for ordinal regression". In: (1999).

[68] Francisco Herrera et al. Multiple instance learning. Springer, 2016.

[69] Mohammad Hesam Hesamian et al. "Deep learning techniques for medical image segmentation: achievements and challenges". In: Journal of digital imaging 32.4 (2019), pp. 582–596.

[70] Sture Holm. "A simple sequentially rejective multiple test procedure". In: Scandinavian journal of st
(1979), pp. 65–70.

[71] Ahmed Hosny et al. "Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study". In: PLoS medicine 15.11 (2018), e1002711.

[72] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks". In: Proceedings of the IEEE co
2018, pp. 7132–7141.

[73] Geoffrey D Hugo et al. "A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer". In: Medical physics 44.2 (2017), pp. 762–771.

[74] Elizabeth Huynh et al. "Artificial intelligence in radiation oncology". In: Nature Reviews Clinical On
17.12 (2020), pp. 771–781.

[75] Bulat Ibragimov et al. "Deep learning for identification of critical regions associated with toxicities after liver stereotactic body radiation therapy". In: Medical physics 47.8 (2020), pp. 3721–3731.

[76] Bulat Ibragimov et al. “Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT”. In: Medical physics 45.10 (2018), pp. 4763–4774.

[77] Bulat Ibragimov et al. “Neural networks for deep radiotherapy dose analysis and prediction of liver SBRT outcomes”. In: IEEE journal of biomedical and health informatics 23.5 (2019), pp. 1821–1833.

[78] Rashmi Jadon et al. “A systematic review of dose-volume predictors and constraints for late bowel toxicity following pelvic radiotherapy”. In: Radiation Oncology 14.1 (2019), pp. 1–14.

[79] Andrew Jaegle et al. “Perceiver: General perception with iterative attention”. In: International confere PMLR. 2021, pp. 4651–4664.

[80] Alan M Kalet, Samuel MH Luk, and Mark H Phillips. “Radiation therapy quality assurance tasks and tools: the many roles of machine learning”. In: Medical physics 47.5 (2020), e168–e177.

[81] John Kang et al. “Machine learning approaches for predicting radiation therapy outcomes: a clinician’s perspective”. In: International Journal of Radiation Oncology* Biology* Physic 93.5 (2015), pp. 1127–1135.

[82] Benjamin H Kann et al. “Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks”. In: Scientific reports 8.1 (2018), pp. 1–11.

[83] Brian D Kavanagh et al. “Radiation dose–volume effects in the stomach and small bowel”. In: International Journal of Radiation Oncology* Biology* Physics 76.3 (2010), S101–S107.

[84] Jack Kiefer and Jacob Wolfowitz. “Stochastic estimation of the maximum of a regression function”. In: The Annals of Mathematical Statistics (1952), pp. 462–466.

[85] Roel GJ Kierkels et al. “Direct use of multivariable normal tissue complication probability models in treatment plan optimisation for individualised head and neck cancer ra-

diotherapy produces clinically acceptable treatment plans". In: Radiotherapy and Oncology 112.3 (2014), pp. 430–436.

[86] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: arXiv preprint arXiv:1412.6980 (2014).

[87] Rainer J Klement et al. "Support vector machine-based prediction of local tumor control after stereotactic body radiation therapy for early-stage non-small cell lung cancer". In: International Journal of Radiation Oncology* Biology* Physics 88.3 (2014), pp. 732–738.

[88] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. "Handling imbalanced datasets: A review". In: GESTS international transactions on computer science and engineerin 30.1 (2006), pp. 25–36.

[89] GJ Kutcher et al. "Histogram reduction method for calculating complication probabilities for three-dimensional treatment planning evaluations". In: International Journal of Radiation On 21.1 (1991), pp. 137–146.

[90] Yann LeCun et al. "Backpropagation applied to handwritten zip code recognition". In: Neural computation 1.4 (1989), pp. 541–551.

[91] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: Proceedings of the IEEE 86.11 (1998), pp. 2278–2324.

[92] Sangkyu Lee et al. "Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk". In: Medical physics 42.5 (2015), pp. 2421–2430.

[93] Tsair-Fwu Lee et al. "Using multivariate regression model with least absolute shrinkage and selection operator (LASSO) to predict the incidence of xerostomia after intensity-modulated radiotherapy for head and neck cancer". In: PloS one 9.2 (2014), e89700.

[94] Guangqi Li, Xin Wu, and Xuelei Ma. "Artificial Intelligence in Radiotherapy". In: Seminars in Cancer Biology. Elsevier. 2022.

[95] Hongming Li, Mohamad Habes, and Yong Fan. "Deep ordinal ranking for multi-category diagnosis of Alzheimer's disease using hippocampal MRI data". In: arXiv preprint arXiv:1709.01599 (2017).

[96] Hongming Li et al. "Deep convolutional neural networks for imaging data based survival analysis of rectal cancer". In: 2019 IEEE 16th International Symposium on Biomedical Imagin IEEE. 2019, pp. 846–849.

[97] Jiayun Li et al. "A multi-resolution model for histopathology image classification and localization with multiple instance learning". In: Computers in biology and medicine 131 (2021), p. 104253.

[98] Na Li et al. "Analysis of related factors of radiation pneumonia caused by precise radiotherapy of esophageal cancer based on random forest algorithm". In: Mathematical Biosciences and I 18.4 (2021), pp. 4477–4490.

[99] Bin Liang et al. "Prediction of radiation pneumonitis with dose distribution: a convolutional neural network (CNN) based model". In: Frontiers in oncology (2020), p. 1500.

[100] Hui Lin et al. "A super-learner model for tumor motion prediction and management in radiation therapy: development and feasibility evaluation". In: Scientific reports 9.1 (2019), pp. 1–11.

[101] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: Medical image analysi 42 (2017), pp. 60–88.

[102] Roderick JA Little and Donald B Rubin. "Statistical analysis with missing data. John Wiley & Sons". In: New York (2002).

[103] Bin Lou et al. "An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction". In: The Lancet Digital Health 1.3 (2019), e136–e147.

[104] Bradley Christopher Lowekamp et al. "The design of SimpleITK". In: Front. Neuroinform. 7 (2013), p. 45.

[105] Dhruv Mahajan et al. “Exploring the limits of weakly supervised pretraining”. In: Proceedings of the European conference on computer vision (ECCV). 2018, pp. 181–196.

[106] Seied Rabie Mahdavi et al. “Use of artificial neural network for pretreatment verification of intensity modulation radiation therapy fields”. In: The British Journal of Radiology 92.1102 (2019), p. 20190355.

[107] Rafid Mahmood et al. “Automated treatment planning in radiation therapy using generative adversarial networks”. In: Machine Learning for Healthcare Conference. PMLR. 2018, pp. 484–499.

[108] Lawrence B Marks et al. “Use of normal tissue complication probability models in the clinic”. In: International Journal of Radiation Oncology* Biology* Physics 76.3 (2010), S10–S19.

[109] Darcy Mason. “SU-E-T-33: pydicom: an open source DICOM library”. In: Medical Physics 38.6Part10 (2011), pp. 3493–3493.

[110] James L McClelland, David E Rumelhart, PDP Research Group, et al. Parallel distributed processing Vol. 2. MIT press, 1987.

[111] Ninareh Mehrabi et al. “A survey on bias and fairness in machine learning”. In: ACM Computing Sur 54.6 (2021), pp. 1–35.

[112] Kuo Men et al. “A deep learning model for predicting xerostomia due to radiation therapy for head and neck squamous cell carcinoma in the RTOG 0522 clinical trial”. In: International Journal of Radiation Oncology* Biology* Physics 105.2 (2019), pp. 440–447.

[113] Philippe Meyer et al. “Survey on deep learning for radiotherapy”. In: Computers in biology and medi 98 (2018), pp. 126–146.

[114] Daniele Micci-Barreca. “A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems”. In: ACM SIGKDD Explorations Newsletter 3.1 (2001), pp. 27–32.

[115] Dan Nguyen et al. “3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture”. In: Physics in medicine & Biology 64.6 (2019), p. 065020.

[116] Stanislav Nikolov et al. “Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy”. In: arXiv preprint arXiv:1809.04430 (2018).

[117] Dipesh Niraula et al. “Current status and future developments in predicting outcomes in radiation oncology”. In: The British Journal of Radiology 95.1139 (2022), p. 20220239.

[118] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: IEEE Transactions on knowled 22.10 (2009), pp. 1345–1359.

[119] Adam Paszke et al. “Automatic differentiation in PyTorch”. In: (2017).

[120] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[121] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: Journal of Machine Learning Res 12 (2011), pp. 2825–2830.

[122] Andrea Pella et al. “Use of machine learning methods for prediction of acute toxicity in organs at risk following prostate radiotherapy”. In: Medical physics 38.6Part1 (2011), pp. 2859–2867.

[123] Eduardo HP Pooch, Pedro Ballester, and Rodrigo C Barros. “Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification”. In: Thoracic Image Analysis: Second International Workshop, TIA 2020, Held in Conjunction with MI Springer. 2020, pp. 74–83.

[124] X Qi, J Neylon, and A Santhanam. “Dosimetric predictors for quality of life after prostate stereotactic body radiation therapy via deep learning network”. In: International Journal of R 99.2 (2017), S167.

[125] Ning Qian. “On the momentum term in gradient descent learning algorithms”. In: Neural networks 12.1 (1999), pp. 145–151.

[126] Gwenolé Quellec et al. “Multiple-instance learning for medical image and video analysis”. In: IEEE Rev. Biomed. Eng. 10 (2017), pp. 213–234.

[127] Shyamsundar Rajaram et al. “Classification approach towards ranking and sorting problems”. In: European conference on machine learning. Springer. 2003, pp. 301–312.

[128] Prajit Ramachandran et al. “Stand-alone self-attention in vision models”. In: Advances in neural info 32 (2019).

[129] Sebastian Raschka and Vahid Mirjalili. Python machine learning: Machine learning and deep learnir Packt Publishing Ltd, 2019.

[130] Brandon Reber et al. “Comparison of Machine-Learning and Deep-Learning Methods for the Prediction of Osteoradionecrosis Resulting From Head and Neck Cancer Radiation Therapy”. In: Advances in Radiation Oncology 8.4 (2023), p. 101163.

[131] John M Robertson, Matthias Söhn, and Di Yan. “Predicting grade 3 acute diarrhea during radiation therapy for rectal cancer using a cutoff-dose logistic regression normal tissue complication probability model”. In: International Journal of Radiation Oncology* Biology* Ph 77.1 (2010), pp. 66–72.

[132] Berkman Sahiner et al. “Deep learning in medical imaging and radiation therapy”. In: Medical physics 46.1 (2019), e1–e36.

[133] Veit Sandfort et al. “Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks”. In: Scientific reports 9.1 (2019), pp. 1–9.

[134] Todd W Schiller et al. “Modeling radiation-induced lung injury risk with an ensemble of support vector machines”. In: Neurocomputing 73.10-12 (2010), pp. 1861–1867.

[135] Jürgen Schmidhuber. “Learning to control fast-weight memories: An alternative to dynamic recurrent networks”. In: Neural Computation 4.1 (1992), pp. 131–139.

[136] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: Proceedings of the IEEE international conference on computer visi 2017, pp. 618–626.

[137] Amnon Shashua and Anat Levin. “Ranking with large margin principle: Two approaches”. In: Advances in neural information processing systems 15 (2002).

[138] Ke Sheng. “Artificial intelligence in radiotherapy: a technological review”. In: Frontiers of Medicine 14.4 (2020), pp. 431–449.

[139] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: arXiv preprint arXiv:1409.1556 (2014).

[140] Carla Sini et al. “Patient-reported intestinal toxicity from whole pelvis intensity-modulated radiotherapy: first quantification of bowel dose–volume effects”. In: Radiotherapy and Oncology 124.2 (2017), pp. 296–301.

[141] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. “UCF101: A dataset of 101 human actions classes from videos in the wild”. In: arXiv preprint arXiv:1212.0402 (2012).

[142] Niemierko A Stavreva and Goitein M Stavrev. “Modelling the dose-volume response of the spinal cord based on the idea of damage to contiguous functional subunits”. In: International journal of radiation biology 77.6 (2001), pp. 695–702.

[143] Min Su et al. “An artificial neural network for predicting the incidence of radiation pneumonitis”. In: Medical physics 32.2 (2005), pp. 318–325.

[144] Xu Sun and Weichao Xu. “Fast implementation of DeLong’s algorithm for comparing the areas under correlated receiver operating characteristic curves”. In: IEEE Signal Processing Lette 21.11 (2014), pp. 1389–1393.

[145] Shohei Tanaka et al. “A deep learning-based radiomics approach to predict head and neck tumor regression for adaptive radiotherapy”. In: Scientific reports 12.1 (2022), p. 8899.

[146] MTW Teo, D Sebag-Montefiore, and CF Donnellan. "Prevention and management of radiation-induced late gastrointestinal toxicity". In: Clinical oncology 27.11 (2015), pp. 656–667.

[147] Jonas Teuwen, Zeno AR Gouw, and Jan-Jakob Sonke. "Artificial Intelligence for Image Registration in Radiation Oncology". In: Seminars in Radiation Oncology. Vol. 32. 4. Elsevier. 2022, pp. 330–342.

[148] S Tomatis et al. "Late rectal bleeding after 3D-CRT for prostate cancer: development of a neural-network-based predictive model". In: Physics in Medicine & Biology 57.5 (2012), p. 1399.

[149] Seiji Tomori et al. "A deep learning-based prediction model for gamma evaluation in patient-specific quality assurance". In: Medical physics 45.9 (2018), pp. 4055–4065.

[150] Tong Tong et al. "Multiple instance learning for classification of dementia in brain MRI". In: Medical image analysis 18.5 (2014), pp. 808–818.

[151] Du Tran et al. "Learning spatiotemporal features with 3d convolutional networks". In: Proceedings of the IEEE international conference on computer vision. 2015, pp. 4489–4497.

[152] Susan L Tucker et al. "Dose–volume response analyses of late rectal bleeding after radiotherapy for prostate cancer". In: International Journal of Radiation Oncology* Biology* Physics 59.2 (2004), pp. 353–365.

[153] Cancer Research UK. Stages, type and grades. 2022. URL: https://www.cancerresearchuk.org/about-cancer/cervical-cancer/stages-types-grades/stage-3.

[154] San José State University. URL: https://www.sjsu.edu/faculty/gerstman/StatPrimer/z-two-tails.pdf.

[155] George Van Andel et al. "An international field study of the EORTC QLQ-PR25: a questionnaire for assessing the health-related quality of life of patients with prostate cancer". In: European journal of cancer 44.16 (2008), pp. 2418–2424.

[156] Stef Van Buuren and Karin Groothuis-Oudshoorn. “mice: Multivariate imputation by chained equations in R”. In: Journal of statistical software 45 (2011), pp. 1–67.

[157] Liesbeth Vandewinckele et al. “Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance”. In: Radiotherapy and On 153 (2020), pp. 55–66.

[158] Ashish Vaswani et al. “Attention is all you need”. In: Advances in neural information processing syst 30 (2017).

[159] Tom Vercauteren et al. “Diffeomorphic demons: Efficient non-parametric image registration”. In: NeuroImage 45.1 (2009), S61–S72.

[160] Ivan R Vogelius et al. “Failure-probability driven dose painting”. In: Medical Physics 40.8 (2013), p. 081717.

[161] Xiang-Bo Wan et al. “Molecular prognostic prediction for locally advanced nasopharyngeal carcinoma by support vector machine integrated approach”. In: PloS one 7.3 (2012), e31989.

[162] Chuang Wang et al. “Predicting spatial esophageal changes in a multimodal longitudinal imaging study via a convolutional recurrent neural network”. In: Physics in Medicine & Biology 65.23 (2020), p. 235027.

[163] Chunhao Wang et al. “Dose-distribution-driven PET image-based outcome prediction (DDD-PIOP): a deep learning study for oropharyngeal cancer IMRT application”. In: Frontiers in Oncology (2020), p. 1592.

[164] Huina Wang, Yihui Liu, and Wei Huang. “Random forest and Bayesian prediction for Hepatitis B virus reactivation”. In: 2017 13th International Conference on Natural Computation, Fuz IEEE. 2017, pp. 2060–2064.

[165] Linda J Wedlake et al. “Evaluating the efficacy of statins and ACE-inhibitors in reducing gastrointestinal toxicity in patients receiving radiotherapy for pelvic malignancies”. In: European journal of cancer 48.14 (2012), pp. 2117–2124.

[166] Mattea L Welch et al. “User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions”. In: Physica Medica 70 (2020), pp. 145–152.

[167] RN Whistance et al. “Clinical and psychometric validation of the EORTC QLQ-CR29 questionnaire module to assess health-related quality of life in patients with colorectal cancer”. In: European journal of cancer 45.17 (2009), pp. 3017–3026.

[168] Lei Xing, Elizabeth A Krupinski, and Jing Cai. “Artificial intelligence will soon change the landscape of medical physics research and practice”. In: Medical physics 45.5 (2018), pp. 1791–1793.

[169] Yan Xu et al. “Deep learning of feature representation with multiple instance learning for medical image analysis”. In: 2014 IEEE international conference on acoustics, speech and signal IEEE. 2014, pp. 1626–1630.

[170] Noorazrul Yahya et al. “Dosimetry, clinical factors and medication intake influencing urinary symptoms after prostate radiotherapy: An analysis of data from the RADAR prostate radiotherapy trial”. In: Radiotherapy and Oncology 116.1 (2015), pp. 112–118.

[171] Noorazrul Yahya et al. “Statistical-learning strategies generate only modestly performing predictive models for urinary symptoms following external beam radiotherapy of the prostate: A comparison of conventional and machine-learning methods”. In: Medical Physics 43.5 (2016), pp. 2040–2052.

[172] Zhijian Yang et al. “Machine learning and statistical prediction of patient quality-of-life after prostate radiation therapy”. In: Computers in Biology and Medicine 129 (2021), p. 104127.

[173] Ziv Yaniv et al. “SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research”. In: Journal of digital imaging 31.3 (2018), pp. 290–303.

[174] Jason Yosinski et al. “How transferable are features in deep neural networks?” In: Advances in neural information processing systems 27 (2014).

[175] Meiyan Yue et al. “Dose prediction via distance-guided deep learning: Initial development for nasopharyngeal carcinoma radiotherapy”. In: Radiotherapy and Oncology 170 (2022), pp. 198–204.

[176] Marvin Zhang et al. “Adaptive risk minimization: Learning to adapt to domain shift”. In: Advances in Neural Information Processing Systems 34 (2021), pp. 23664–23678.

[177] Xin Zhen et al. “Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study”. In: Physics in Medicine & Bi 62.21 (2017), p. 8246.

[178] Chengju Zhou, Meiqing Wu, and Siew-Kei Lam. “SSA-CNN: Semantic self-attention CNN for pedestrian detection”. In: arXiv preprint arXiv:1902.09080 (2019).

[179] Xizhou Zhu et al. “An empirical study of spatial attention mechanisms in deep networks”. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, pp. 6688–6697.

[180] Fuzhen Zhuang et al. “A comprehensive survey on transfer learning”. In: Proceedings of the IEEE 109.1 (2020), pp. 43–76.