Use of Machine Learning Methods for the Prediction of Mandibular Osteoradionecrosis in Head and Neck Cancer Cases Treated with Radiotherapy

Humbert-Vidan, Laia

Awarding institution:
King's College London

KING'S COLLEGE LONDON

DOCTORAL THESIS

Use of Machine Learning Methods for the Prediction of Mandibular Osteoradionecrosis in Head and Neck Cancer Cases Treated with Radiotherapy

Author:
Laia Humbert-Vidan

Supervisors:
Dr Teresa Guerrero Urbano
Dr Andrew P King

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in the

School of Cancer & Pharmaceutical Sciences
King's College London

I would like to dedicate this thesis to my husband Daniel and my children Mark and Aina, for their unconditional patience, encouragement and love even during the most challenging moments of this journey. Thank you for always being there.

Acknowledgements

This thesis is the result of four years of hard work and persistence but also enjoyment and satisfaction. None of it would have been possible without the support from many academics, colleagues, friends and family.

I would like to express my deepest gratitude to my supervisors, Dr Teresa Guerrero Urbano and Dr Andrew King, for their invaluable guidance and continuous support and encouragement throughout my learning process in this journey. Not only have they been the key to the successful completion of my PhD but they have provided me with a wide range of skills that will serve me for the rest of my research career. I will never be able to thank them enough for this opportunity. I am also extremely grateful to the thesis progression committee members for contributing to a better research by questioning my methods and always providing useful advice. Many thanks should also go to Dr Vinod Patel for sharing his knowledge and ORN data, without which this thesis would not have been possible.

Thanks must be extended to my colleagues from the School of Biomedical Engineering and Imaging Sciences at King's College London for their support and inspiration and for making me feel part of the team. In particular, I would like to thank Dr Ilkay Oksuz and Dr Esther Puyol Antón for their technical support at the start of my thesis, Robin Andlauer for his advice on the multimodality fusion experiments and Shaheim Ogbomo-Harmitt for his advice on the interpretability experiments. I would also like to thank my colleagues from the Radiotherapy Physics department at Guy's and St Thomas' Hospitals for their support and friendship during all these years. Special thanks should go to the late Dr David Convery for his countless support on the technical issues encountered with the radiotherapy treatment planning systems.

I would also like to thank my family and friends for their love and support throughout my life. My Mum has taught me to never give up and always follow my passions. My Dad inspired me from a very young age with his dedication to his career and patients. He later encouraged me to pursue this journey into cancer treatment and research through his own

illness. My brother has helped me always put things in perspective with his priceless sense of humor.

Most of all, I would like to thank my husband and friend Daniel for undoubtedly encouraging me to pursue my passion despite the added challenge of doing so while starting our own family. None of this would have been possible without his unconditional trust, patience and love. Thank you for your deep listening and your selfless advice.

Lastly, I would like to thank my children, Mark and Aina, for helping me develop my multitasking skills and teaching me to be present not matter how long my to-do list might be.

Laia Humbert Vidan, March 2023

Abstract

Mandibular osteoradionecrosis (ORN) in patients with head and neck cancer undergoing radiotherapy (RT) is a rare radiation-induced toxicity but can highly compromise patients' quality of life and result in costly clinical interventions.

In addition to clinical and demographic risk factors, radiation dose plays an important role in the development of mandibular ORN. Existing ORN prediction models use the dosimetric information extracted from the dose-volume histogram (DVH) of the mandible. In a DVH, the clinical radiation dose distribution map of the mandible volume is reduced to a 2D representation that omits any spatial dose information. Because the anatomy and the radiosensitivity varies across the mandible, this spatial dose information is clinically relevant. In this thesis I hypothesise that the incidence of mandibular ORN can be predicted based on the clinical radiation dose distribution maps as the dosimetric factor combined with the clinical and demographic factors.

A class-balanced cohort of up to 92 ORN cases and 92 matched controls treated with intensity-modulated radiotherapy (IMRT) between 2011 and 2022 were retrospectively selected from the clinical database. The clinical and demographic data was retrieved from the clinical notes and the DVH and RT DICOM files were exported from the clinical treatment planning system. To facilitate subsequent ORN prediction model development, a pipeline was developed that involves a number of image pre-processing steps. First, the computed tomography (CT), RT dose and mandible structure volumes were registered to a common space to compensate for inter-patient positioning variations. Then the RT dose map was masked by the mandible structure, thus resulting in the mandible dose map.

The first part of this thesis focuses on exploring machine learning (ML) and deep learning (DL) classification models for the prediction of ORN incidence. A first experiment was performed to initiate the transition from traditional ORN risk factor analysis to ML-based case-by-case ORN incidence prediction. The performance of different ML methods was compared on the task of predicting ORN incidence based on DVH metrics and clinical and demographic data. Although no statistically significant difference was

observed between models, the artificial neural network (ANN) model showed the highest prediction accuracy (71%). This study was followed up with a DL-based approach that used the mandible radiation dose map as the input into a 3D deep convolutional neural network (CNN) for binary classification (ORN vs. non-ORN). The predictive performances (AUROC) of three different CNN architectures were compared, including a DenseNet121 (0.64), a DenseNet40 (0.69) and a ShuffleNet (0.65). The dose map-based deep CNN model prediction performance results were then compared to a DVH-based Random Forest (RF) model (0.61 AUROC). This DL-based ORN prediction approach was expanded to include other non-dosimetric risk factors (clinical variables). This was done following early and late multimodality fusion strategies, which resulted in similar prediction performances (0.68 and 0.70 AUROC, respectively) to the single-modality DL model (0.69), but had a statistically significantly higher performance than a RF model trained on clinical variables only (0.60).

The second part of this thesis focuses on exploring the interpretability of the DL-based ORN prediction model using the 3D Grad-CAM pixel-attribution method and quantitatively analysing its results to draw clinically relevant conclusions. The results obtained were in alignment to existing clinical knowledge derived from the more traditional statistical approaches, which represents an important step towards gaining trust for the clinical implementation of a DL-based ORN prediction model.

Finally, this thesis includes a description of the PREDMORN multi-institutional study, which I designed and developed to obtain the largest and most diverse mandibular ORN dataset worldwide that will allow for the development of robust and generalisable ORN prediction models and further subsequent studies.

Overall, I expect that the work included in this thesis will represent a significant step towards a more individualised treatment of head and neck cancer that will potentially result in an incidence reduction or better prognosis of mandibular ORN.

Publications

- L. Humbert-Vidan, V. Patel, A.P. King and T. Guerrero Urbano. Letter to the editor regarding the paper entitled "Comparison of Machine Learning and Deep Learning Methods for the Prediction of Osteoradionecrosis Resulting from Head and Neck Cancer Radiation Therapy" by Reben et al. (2022). Accepted for publication in *Advances in Radiation Oncology*. 2023

- L. Humbert-Vidan, C.R. Hansen, C.D. Fuller, S. Petit, A. van der Schaaf, L.V. van Dijk, G.M. Verduijn, H.Langendijk, C.Muñoz-Montplet, W.Heemsbergen, M.Witjes, A.S.R. Mohamed, A.A. Khan, J. Marruecos Querol, I. Oliveras Cancio, V. Patel, A.P. King, J. Johansen, and T. Guerrero Urbano. Protocol letter: A multi-institutional retrospective case-control cohort investigating PREDiction models for Mandibular OsteoRadioNecrosis in head and neck cancer (PREDMORN). In *Radiotherapy and Oncology*, 176:99-100, 11, November 2022

- L. Humbert-Vidan, V. Patel, R. Andlauer, A.P. King and T. Guerrero Urbano. PO-1770 Prediction of mandibular ORN with DL-based classification of 3D radiation dose distribution maps. In *Radiotherapy and Oncology, Poster (Digital): Radiomics, Modelling and Statistical Methods*, Volume 170, Supplement 1, S1574, May 2022

- L. Humbert-Vidan, V. Patel, R. Andlauer, A.P. King and T. Guerrero Urbano. Prediction of mandibular ORN incidence from 3D radiation dose distribution maps using deep learning. In *Applications of Medical Artificial Intelligence, AMAI 2022. Lecture Notes in Computer Science*, 13540:49-58, December 2021

- L. Humbert-Vidan, V. Patel, R.H. Begum, M. McGovern, D. Eaton, A. Kong, I. Petkar, M. Reis Ferreira, M. Lei, A.P. King, and T. Guerrero Urbano. PH-0387 Mandible osteoradionecrosis: a dosimetric study. In *Radiotherapy and Oncology*, vol. 161:285-286, August 2021

- L. Humbert-Vidan, V. Patel, I. Oksuz, A.P. King and T. Guerrero Urbano. Comparison of machine learning methods for prediction of osteoradionecrosis incidence in pa-

tients with head and neck cancer. In *The British Journal of Radiology*, 94:20200026, 4, March 2021

- V. Patel, L. Humbert-Vidan, C. Thomas, I. Sassoon, M. McGurk, M.R. Fenlon and T. Guerrero Urbano. Dentoalveolar radiation dose following IMRT in oropharyngeal cancer - An observational study. In *Special Care in Dentistry*, 41(3), 2021

- V. Patel, L. Humbert-Vidan, C. Thomas, I. Sassoon, M. McGurk, M.R. Fenlon and T. Guerrero Urbano. Radiotherapy quadrant doses in oropharyngeal cancer treated with intensity modulated radiotherapy. In *Faculty Dental Journal*, Volume 11, Issue 4, 2020

- L. Humbert-Vidan, I. Oksuz, V. Patel, A.P. King and T. Guerrero Urbano. EP-1929 Prediction of voxelwise mandibular osteoradionecrosis maps in HNC patients using deep learning. In *Radiotherapy and Oncology*, 133:S1050, April 2019

- L. Humbert-Vidan, S.L. Gulliford, V. Patel, C. Thomas and T. Guerrero Urbano. EP-1603: Atlas of complication incidence to explore dosimetric contributions to osteoradionecrosis. In *Radiotherapy and Oncology*, 123:S864-S865, May 2017.

Contents

List of figures

List of tables

Nomenclature

Symbols

a	Learning rate constant in the gradient descent optimisation algorithm
α/β	alpha/beta ratio in the LQ survival curve
b	Bias term in an ANN
$C, l2$	LR penalty and regularisation hyperparamters
f	Activation function in an ANN
γ	SVC gamma hyperparameter
$L_{3DGradCAM}$	3D GradCAM localisation map
$\cap$	Overlap between two voxel regions
$P(y\|x)$	Predicted probability of class y given the input x
S, K, I	Convolution operation output, kernel and image
$\sigma(z)$	Sigmoid or logistic function
θ_i	Model parameters in logistic regression
w_i	Ensemble model weights
x_i	Model input variables
y	Model output variables

Abbreviations

3DCRT	3D Conformal Radiotherapy

AAA Analytical Anisotropic Algorithm

AdaBoost Adaptive Boosting

AI Artificial Intelligence

ANN Artificial Neural Network

AUROC Area Under the Receiver-Operator Characteristics curve

BIC Bayesian Information Criterion

BNN Bayesian neural network

CNN Convolutional Neural Networks

CRT Chemoradiotherapy

CT Computerised Tomography

CTV Clinical Tumour Volume

D_{max} Maximum Dose

D_{mean} Mean Dose

D Absorbed Radiation Dose

DICOM Digital Imaging and Communications in Medicine

DL Deep Learning

Dmin Min Dose

DOB Date Of Birth

DSC Dice Similarity Coefficient

DTA Data Transfer Agreement

DVH Dose-Volume Histogram

DX% Dose received by the X% of the anatomical volume

EBRT External Beam Radiotherapy

EMC Erasmus Medical Centre

EQD2 Equivalent Dose in 2Gy fractions

FcNN Fully Connected Neural Network

FN False Negative

FP False Positive

FPR False Positive Rate

fx fraction

GDPR General Data Protection Regulation

GPU Graphics Processing Unit

Grad-CAM Gradient-weighted Class Activation Mapping

GSTT Guy's and St Thomas' NHS Foundation Trust

GTV Gross Tumour Volume

Gy Grays

HNC Head and Neck Cancer

HN Head and Neck

HNSCC HN Squamous Cell Carcinoma

HPV Human Papilloma Virus

HU Hounsfield Units

ICO Catalan Institute of Oncology

IMRT Intensity-Modulated Radiotherapy

IQR Interquartile Range

KCL King's College London

linac Linear Accelerator

LKB Lyman-Kutcher-Burman

LQ Linear-Quadratic Model

LR Logistic Regression

MC Monte Carlo Algorithm

MDACC MD Anderson Cancer Centre

MeV Mega electron Volts

MLC Multi-Leaf Collimators

ML Machine Learning

MLP Multilayer Perceptron

MONAI Medical Open Network for Artificial Intelligence

MWU Mann-Whitney U statistical test

CI CTCAE The National Cancer Institute Common Terminology Criteria for Adverse Events

NICE National Institute for Health and Care Excellence

NTCP Normal Tissue Complication Probability

OAR Organ At Risk

OAR Organ at Risk

OCC Oral Cavity Cancer

OPC Oropharyngeal Cancer

ORN Osteoradionecrosis

OUH Odense University Hospital

PCA Principal Component Analysis

POCRT Postoperative Chemoradiotherapy

PORT Postoperative Radiotherapy

PREDMORN Prediction Models in Mandibular Osteoradionecrosis

PTV Planning Tumour Volume

PV Pixel Value

QUANTEC Quantitative Analyses of Normal Tissue Effects in the Clinic

RBF Radial Basis Function

ReLU Rectified Linear Unit

RF Random Forest

RGF Regularised Greedy Forest

RT Radiotherapy

SVM Support Vector Machine

TCP Tumour Control Probability

TNM Tumour-Node-Metastasis

TN True Negative

TPR True Positive Rate

TPS Treatment Planning System

TP True Positive

TRIPOD Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

UMCG University Medical Centre Groningen

VMAT Volumetric Arc Therapy

VXGy Anatomical volume receiving XGy of radiation dose

Chapter 1

Introduction

1.1 Motivation

A priori knowledge of which patients are more likely to develop mandibular osteoradionecrosis (ORN) may inform the clinical decision of closer follow-up schedules and prophylactic measures. Existing ORN normal tissue complication probability (NTCP) models [1] rely solely on dose-volume histogram (DVH) parameters (in addition to other clinical and demographic risk factors). The current gold standard radiotherapy (RT) plan optimisation processes also rely on the DVH-based constraints for the target volumes and the organs at risk such as the mandible. The DVH of a structure is a 2D reduction of the simulated 3D radiation dose distribution within that structure and it discards any spatial dose information. Moreover, the anatomy of the mandible is not homogeneous, with varying bone composition and vascularisation. The radiobiological response across the mandible also varies, with some regions more prone to ORN development. NTCP models for mandibular ORN incorporating spatial dose information would enhance treatment personalisation by taking into account the anatomical and radiobiological heterogeneities within the mandible.

In comparison to manual image feature extraction methods, deep learning (DL) offers the opportunity for an automated NTCP prediction pipeline that extracts features from 3D radiation dose distribution maps and combines them with relevant non-image data such as clinical, demographic and patient variables.

The first part of this thesis explores the implementation of DL methods in ORN NTCP models using 3D radiation dose distribution maps. In the second part of this thesis, I explore DL interpretability methods and perform quantitative analyis on the results

to obtain clinically explainable conclusions from the DL-based ORN prediction model predictions. The need for larger and more diverse datasets to train more generalisable ORN prediction models has led to the design and development of the PREDMORN multi-institutional study to investigate prediction models for mandibular ORN in head and neck cancer (HNC), which is the focus of the final part of this thesis.

The next section summarises the original contributions of the work included in this thesis to the field of mandibular ORN prediction.

1.2 Contributions

The use of ML and DL methods in radiotherapy toxicity prediction is still in its early stages. The original contributions of this thesis are mostly in the use of ML and DL for radiation toxicity prediction in head and neck cancer, with a focus on mandibular ORN.

- **Use of ML methods for prediction of mandibular ORN incidence.** Mandibular ORN has a multifactorial aetiology with radiation dose, clinical and demographic information as potential risk factors. Existing work has explored correlations between these factors using traditional statistical methods. In comparison to these methods, ML methods enable the analysis on large datasets without a priori knowledge on how these are related with the potential of achieving predictions on a case-by-case basis. The use of ML methods in the context of ORN prediction is a novel approach that represents a first step towards AI-based ORN NTCP models.
- **Transition from the DVH to the 3D radiation dose distribution map as the dosimetric input in ORN prediction models.** Existing work on ORN prediction is based on DVH metrics. The proposed DL-based pipeline uses clinical 3D radiation dose distribution maps instead, which include spatial dose information. This novel approach has the potential of including localisation knowledge that can be linked to the existing anatomical and radiobiological organ heterogeneities for the prediction of ORN, thus resulting in more realistic and clinically relevant NTCP models.
- **Multimodality DL-based ORN NTCP modelling.** Multimodal fusion DL strategies have the potential of including different data modalities (e.g. dose maps and clinical data), learning the interactions between the risk factors and even adaptively fusing the different modalities based on their respective informativeness. This novel application of multimodal fusion in the context of ORN prediction has the clear advantage of an enhanced and more comprehensive clinical decision-making tool.

- **Quantitative analysis of pixel attribution interpretability results.** This is the first study that analyses spatial dose associations with mandibular ORN incidence using the DL-based 3D Grad-CAM voxel attribution method. Moreover, existing RT toxicity prediction studies with successfully implemented interpretability methods provide a qualitative analysis of their results. We propose a comprehensive quantitative analysis of the 3D Grad-CAM results that we hope will further contribute to gaining users' trust.
- **The PREDMORN study.** Due to the low prevalence rate of mandibular ORN, low patient numbers represent a statistical limitation to the existing work. External validation of models using data from multiple centres is an essential feature of effective model evaluation with a view to clinical translation but is often lacking. The PREDMORN (PREDiction models for Mandibular OsteoRadioNecrosis) study is a multi-institutional effort involving six teaching hospitals. It will enable the largest datasets worldwide to be used to develop, train and validate robust and generalisable NTCP models for mandibular ORN.

1.3 Outline

This section provides an outline of the thesis structure, which is composed of the following chapters:

Chapter 2 provides the theoretical context to the key clinical concepts used in this thesis. It is structured in four main parts, focusing on the most relevant concepts for this thesis in HNC, external beam RT, radiobiology and mandibular ORN, respectively.

Chapter 3 first focuses on the technical background to the ML and DL methods used in this thesis. It then reviews the most relevant literature on the use of DL methods for the prediction of radiation-induced toxicities.

Chapter 4 first describes the patient selection process and how the different cohorts were constructed. Next, it specifies which data was used in each experiment of this thesis, including the details of how the data was obtained. It then describes the processing steps followed for the imaging data, including a section on the data protection and anonymisation measures considered. Finally, it discusses some of the decisions made in the data collection and processing steps with respect to other relevant published work.

Chapter 5 presents the results from a comparison between five different supervised classification ML methods for the task of predicting mandibular ORN incidence based on DVH metrics and clinical and demographic variables.

Chapter 6 introduces a novel approach to ORN prediction based on 3D dose distribution maps of the mandible rather than the traditionally used DVH metrics. The performance of DL models trained on the dose maps is compared to that of a DVH-based ML model. Finally, an analysis is included on the effect of factors such as the choice of classification probability threshold or minimum follow-up time requirements for the control group on the model performance results.

Chapter 7 expands on the ORN prediction DL pipeline by including clinical and demographic variables using early and late multimodality fusion methods and compares its prediction performance to that of the single modality ML and DL models.

Chapter 8 aims to provide an insight into how the DL-based decisions are made by including pixel-attribution interpretability methods into the ORN prediction DL pipeline.

Chapter 9 introduces the PREDMORN (PREDiction models for Mandibular OsteoRadioNecrosis) multi-institutional study and describes its published protocol.

Chapter 10 summarises the clinical impact of the scientific contributions of this thesis, discusses the limitations of this work and proposes future directions to address them.

Chapter 2

Clinical Background

This Chapter aims to provide the theoretical context to the most relevant clinical concepts used in this thesis. Section 2.1 provides a clinical background on head and neck cancers, Section 2.2 describes the key concepts of external beam radiotherapy, Section 2.3 contains the most relevant radiobiology concepts and Section 2.4 focuses on mandibular osteoradionecrosis. The technical machine and deep learning concepts, some of them mentioned here, are covered in Chapter 3.

2.1 Head and neck cancers

HNC accounts for 3% of all cancers in the UK, with an average of 12,422 new cases each year [2] and is the seventh most common cancer worldwide [3]. Most primary HNCs start in squamous cells, which are cells that line the mouth, nose and throat. Thus, the most common primary tumour sites for HN squamous cell carcinoma (HNSCC) are the oral cavity (lips, buccal mucosa, hard palate, anterior tongue, floor of mouth and retromolar trigone), the pharynx - which includes the nasopharynx, oropharynx (palatine and lingual tonsils, base of tongue, soft palate, uvula and posterior pharyngeal wall) and the hypopharynx -, the larynx, the paranasal sinuses and nasal cavity and the salivary glands (Figure 2.1). Sometimes, however, the primary tumour site is unknown, with the cancerous cells found away from any of the main HN sites, most commonly in the regional lymph nodes.

2.1.1 HNC risk factors

Tobacco and alcohol consumption, infection with the human papillomavirus (HPV), age, gender and poor oral and dental hygiene are amongst the factors that may increase the probability of developing HNC [4]. Most HPV-related HNSCCs arise in the oropharynx. HPV-related oropharynx cancer (OPC) cases are mostly younger patients with a generally good prognosis and improved overall survival.

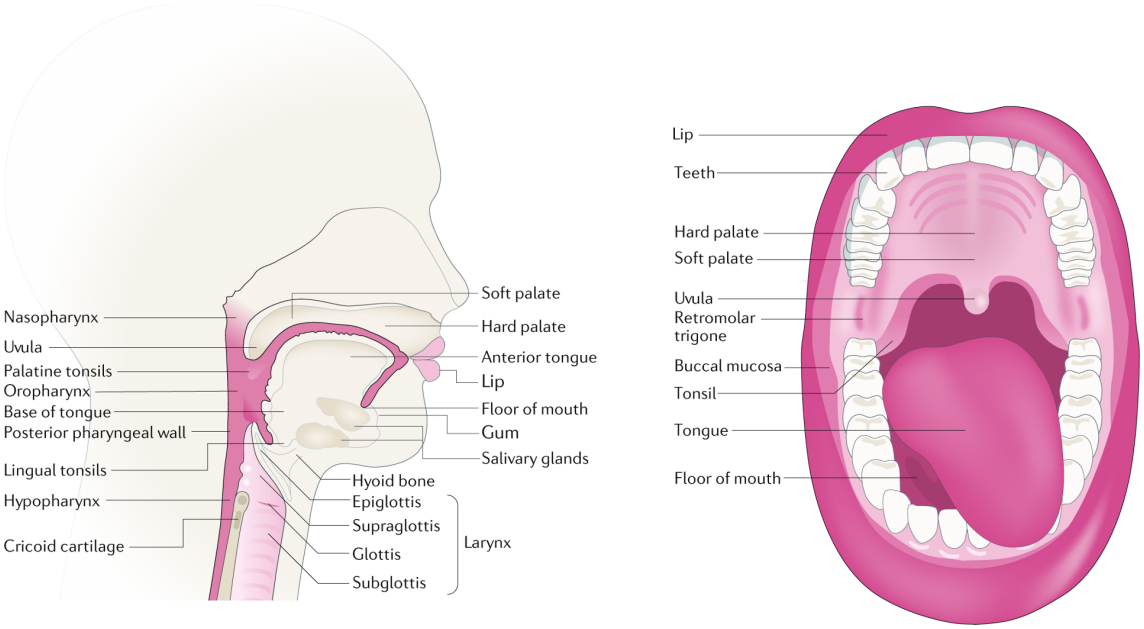


Fig. 2.1 Anatomical sites of head and neck squamous cell carcinoma (HNSCC) development. Figure source: Johnson et al. [5].

2.1.2 HNC staging

The extent or spread of the tumour is an important factor in cancer prognosis and treatment decisions. Based on the histopathologic analysis, the tumour-node-metastasis (TNM) staging system evaluates the characteristics of the tumour (T) at the primary site (based on size, location or both), the degree of involvement of regional lymph nodes (N) and whether the tumour has spread or metastasized (M) to other anatomical sites [6, 7]. Based on the combination of patient-specific T, N and M status, a stage I, II, III or IV cancer can be designated [8]. Early-stage disease is denoted as stage I or II whereas advanced disease is usually stage III or IV. Each anatomical subsite has its own TNM staging and, within this, sub-categories also exist; for instance, T4a (staged as IVa) and T4b (staged as IVb) depending on the local extent of disease [6].

2.1.3 HNC treatment options

Treatment for HNC is defined based on tumour location and TNM staging as well as the patient's performance status and medical history. The curative modalities of treatment for HNC are surgery and external beam radiotherapy (EBRT), used alone or in combination depending on primary site and extent of disease. In more advanced cases, EBRT is often delivered with concomitant chemotherapy (CRT). Postoperative RT (PORT) or CRT (POCRT) is used after primary surgery in cases with locally advanced disease. Patients with poor performance status, previous irradiation of the HN area or a poor prognosis might be prescribed a lower, palliative dose of radiation or systemic therapy such as chemotherapy or immunotherapy or, in some cases, be recommended for best supportive care.

2.1.4 Radiation-induced toxicities in HNC

With the introduction of modern imaging and RT techniques we are able to irradiate the tumour with high precision and conformity. However, due to the nature of energy deposition of photons in tissue, non-target organs inevitably absorb ionising radiation, resulting in normal tissue toxicity, which can lead to complications affecting the patient's quality of life [9, 10]. At Guy's and St Thomas' NHS Foundation Trust (GSTT) HNC patients are followed up closely during treatment and for at least five years post-RT. This provides an opportunity to not only assess treatment response but to quickly identify and act on potential side effects. Complications after radiotherapy fall into two main categories: early and late effects. Early effects develop during treatment or within a few weeks post-RT whereas late effects are observed months (>3) or years after treatment. Treatment adjustments (usually treatment breaks) are possible if early effects occur, although treatments are designed to minimise this as they are associated with poorer survival outcomes; however, by the time late effects develop it is too late to modify a treatment. Acute and late radiation-induced toxicities in HNC are dose limiting, have a significant effect on the patient's quality of life and can also jeopardise treatment compliance, potentially impacting outcome. The main radiation-induced toxicities in HNC include oral mucositis (inflammation of the oral mucosa), xerostomia (dry mouth), dysphagia (swallowing difficulty) and mandibular ORN (necrosis of the jaw) [11]. These complications are often associated, with patients experiencing more than one type as a result of the same course of RT [12, 13]. For instance, xerostomia may result in poor oral health which is in turn a risk factor for ORN [12]. ORN has also been associated to a higher prevalence and perceived symptom burden of

dysphagia [13]. The present work, however, has focused on mandibular ORN, which is a late radiation side effect.

2.2 External beam radiotherapy

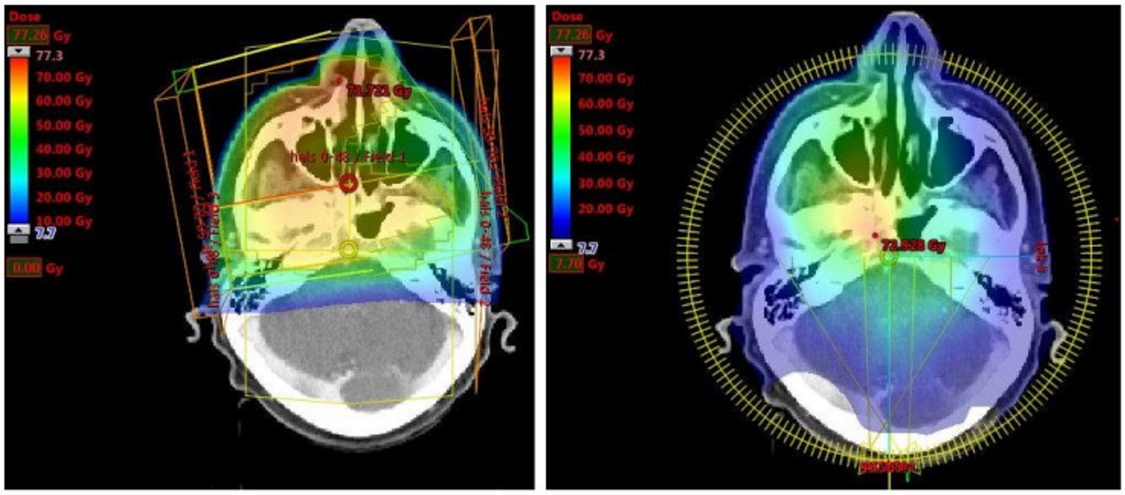
RT is the use of ionising radiation to produce radiation-induced damage to tumour cells with minimal damage to normal tissue cells. EBRT uses high-energy photons and electrons produced by a medical linear accelerator (linac). This section aims to describe the key concepts of radiation therapy that provide the theoretical background to the experimental work described in later chapters of this thesis. The interested reader is directed to [14] for more in-depth technical details on how radiation is produced by a clinical linac.

2.2.1 Intensity-modulated radiotherapy

One of the most significant advances in RT is the transition from 3D conformal RT (3DCRT) to intensity-modulated RT (IMRT), where geometrical beam shaping evolved to dynamic highly conformal intensity-modulated photon beams achieved with the multi-leaf collimators (MLC) in the treatment head. Thus, RT techniques such as IMRT and volumetric arc therapy (VMAT) – with the treatment head rotating around the patient during an IMRT beam – result in an improved target conformity and a reduction of high doses to the normal tissue. This is particularly relevant to HNC [15, 16] as often the tumour is in close proximity to critical organs of complex geometries. However, the larger number of beams required in the IMRT or VMAT techniques have resulted in a larger volume of the anatomy receiving a low-dose bath compared to 3DCRT [17, 18] (Figure 2.2). Maesschalck et al. [19] found no reduction in mandibular ORN incidence with IMRT in a cohort of OPC patients.

2.2.2 Absorbed radiation dose

The bremsstrahlung photons emitted from the linac treatment head interact with the patient [14]. In the interaction between electromagnetic radiation and matter, photons interact with the electrons of the atoms within the tissue cells. At 6 MeV, the typical beam energy for HNC EBRT, the main interaction process is the Compton effect. The incident photon transfers a fraction of its energy to an electron that is ionized from one of the outer energetic layers of the atom. The ejected electron will continue to interact with other atoms



a b

Fig. 2.2 Comparison of a 3DCRT HNC RT plan (a) and a VMAT HNC RT plan (b). Figure source: Van der Veen et al. [18].

in tissue, thus creating a cascade of ionizations with decreasing energy. It is this cascade of ionizations that causes most of the biological damage to the cells, especially to the DNA molecules within it [20]. Absorbed radiation dose (D) is the quantity that measures how much of this energy (E) is deposited in a finite volume of tissue with mass (m) as a result of ionizing radiation $(D = E/m)$ [14]. Absorbed dose is measured in Grays (Gy), where 1 Gy is equivalent to 1 Joule per Kg $(1Gy = 1J/Kg)$.

2.2.3 EBRT treatment planning

The EBRT treatment planning process consists of a number of steps [21]; the most relevant steps for this work are described in this section.

Delineation of volumes of interest. The gross tumour volume (GTV) is defined on the planning computed tomography (CT) images by a clinical oncologist. A clinical tumour volume (CTV) is then created by adding a margin to account for subclinical extension of the tumour. A further planning tumour volume (PTV) is defined with margins that account for internal organ motion and patient set up uncertainties, respectively [22]. In addition to the target volumes, a number of organs at risk (OAR) are also defined; Figure 2.3 shows the main OARs in the HN region.

Absorbed dose calculation. The radiation dose distribution for a patient is created at the treatment planning stage. A computerised treatment planning system (TPS) simulates absorbed dose distribution for a given radiation beam arrangement and beam intensity

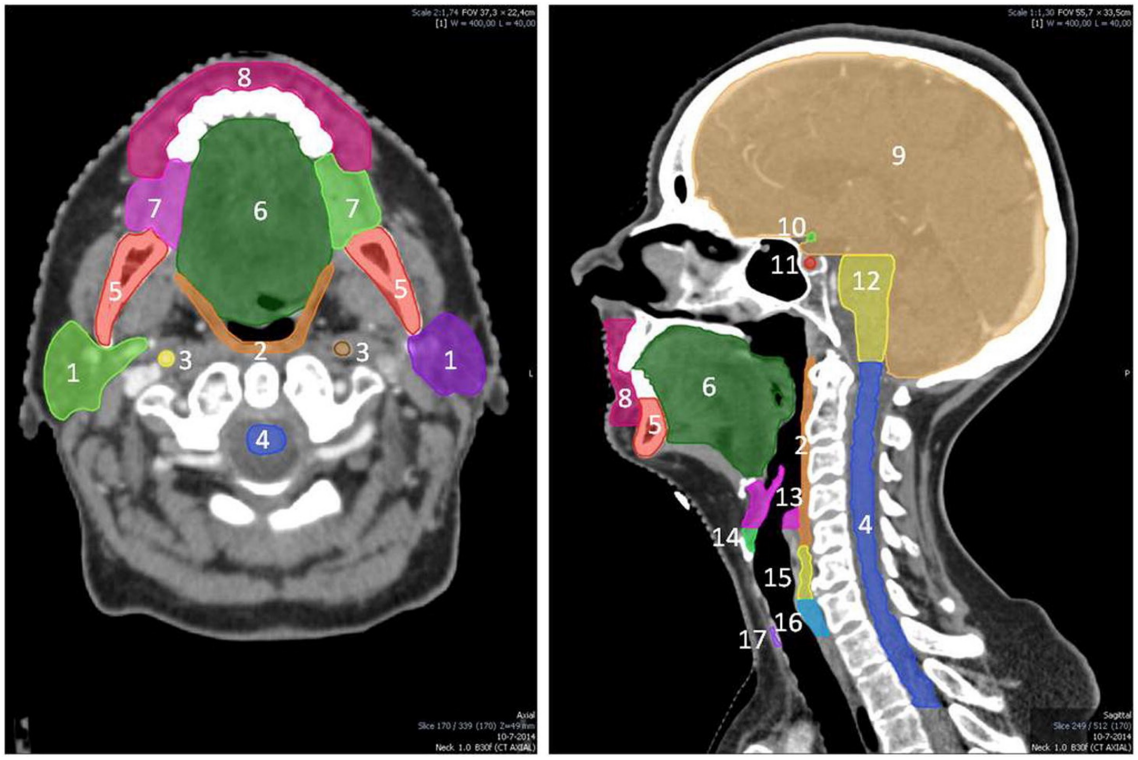


Fig. 2.3 Organs at risk in the HN region, where contour 5 corresponds to the mandible. Figure source: Brouwer et al. [23].

map, based on CT images of the treatment area. A CT scanner-specific calibration curve describing the relationship between electron density and tissue attenuation in the CT image is used by the TPS during dose calculations. The CT number (in Hounsfield units, HU) represents the level of attenuation at each voxel of a CT image, where $HU_{air} = -1000$ and $HU_{water} = 0$ [24]. Dose calculation algorithms are the basis of the radiation treatment plan optimisation process. The most commonly used dose calculation algorithms are the Monte Carlo-based (MC) and kernel-based algorithms [25].

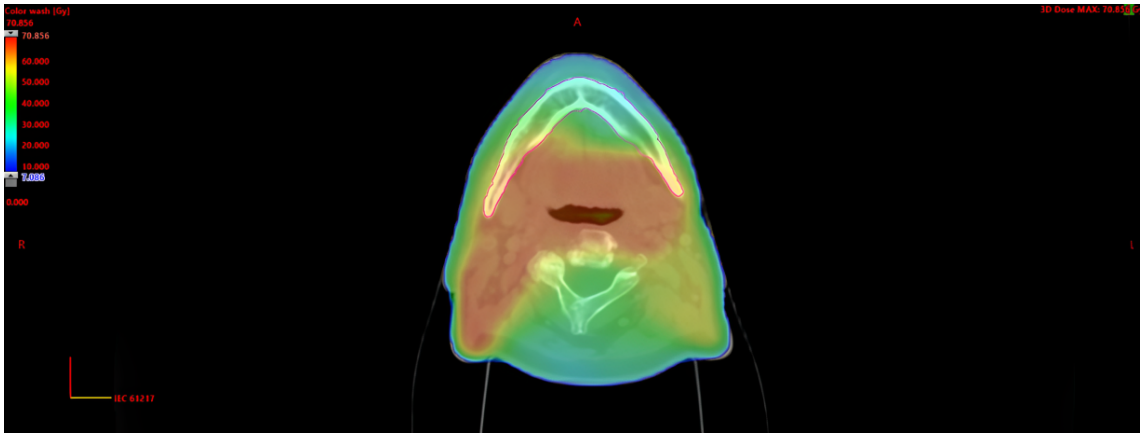
MC methods are highly accurate but very computationally expensive; they randomly simulate the interaction histories of a large number of particles through the treatment head and as they enter the different tissue types. At GSTT, HNC treatment plans used in this thesis were produced using the Monaco TPS (Elekta AB, Stockholm, Sweden), which uses a MC-based dose calculation algorithm, until 2016. Afterwards, the Eclipse TPS (Varian Medical Systems, Inc. Palo Alto, CA, US) with the Analytical Anisotropic Algorithm (AAA) was used instead. The AAA is a kernel-based 3D pencil beam convolution-superposition algorithm. A dose spread kernel is a representation of the energy spread from photons and electrons resulting from the interaction of the primary photons in tissue at a given point. The AAA convolves the kernel with the fluence of energy transferred from

the primary radiation beam to the secondary particles to calculate the absorbed dose [25]. Absorbed dose can be reported by the TPS as $D_{w,w}$, $D_{m,m}$ or $D_{w,m}$, where the first subscript refers to the medium in which radiation transport occurs and the second subscript refers to the medium in which the energy is deposited. At GSTT, $D_{w,m}$ was historically the choice for clinical treatment plans with the Monaco TPS. $D_{m,m}$ was later adopted and all clinical treatment plans created with the Eclipse TPS have been calculated in $D_{m,m}$. Systematic differences of up to 2.7% between $D_{m,m}$ and $D_{w,m}$ have been previously reported [26] for critical structures in the head and neck region. For the work described in this thesis, I re-calculated all the Monaco plans originally created in $D_{w,m}$ into $D_{m,m}$.

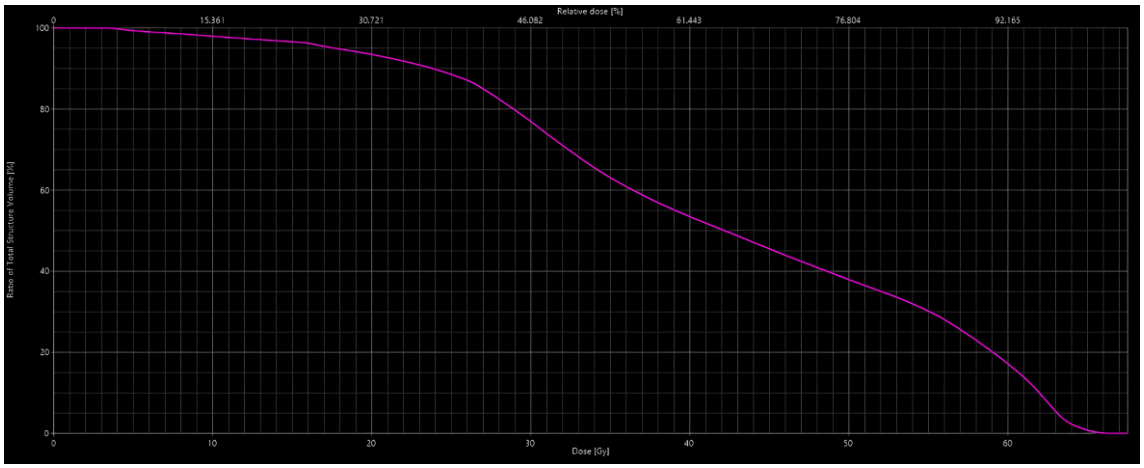
Treatment plan optimisation. A DVH is a 2D representation of the absorbed dose received by a given anatomical volume (Figure 2.4b). DVH-based dosimetric parameters such as maximum (D_{max} or D2%), minimum (D_{min} or D98%), mean (D_{mean}) and median (D50%) doses and other dose-volume levels are used as objectives or constraints to obtain a radiation dose distribution that maximises the coverage of the target volumes while minimising the irradiation of the OARs. At the treatment plan optimisation process the optimal beam intensities are determined by the TPS based on the previously defined target volumes and OARs with the corresponding DVH-based radiation dose objectives and constraints. The use of the DVH for radiotherapy dose optimisation has, however, some limitations [27], especially in non-uniform or partial irradiation of organs, because a DVH does not offer any spatial information and ignores that there may be regions of different functionality and dose response within a single organ. The reduction of a 3D dose map (Figure 2.4a) to a 2D DVH (Figure 2.4b) results in the loss of clinically relevant spatial localisation as well as dose gradient and direction information.

2.3 Radiobiology and therapeutic window

Radiation biology plays a key role in understanding the mechanisms of tumour and normal tissue response to radiation, which is the foundation for defining the treatment strategy and for the development of new approaches in RT. This section aims to describe the fundamental radiobiology concepts that underly the work in this thesis. The interested reader is directed to Joiner et al. [20] for more in-depth information on clinical radiobiology.



(a)



(b)

Fig. 2.4 (a) Axial slice of a 3D dose distribution map with the mandible segmentation for a GSTT HNC case and (b) corresponding mandible DVH, with percentage of mandible volume receiving a given dose level plotted against the dose level in Gy.

2.3.1 DNA damage and the linear-quadratic model

In the cell reproductive cycle, the mitosis or M phase is a process during which the genetic material of a cell is duplicated in order to reproduce itself and create a new identical cell. Radiation is most effective during the M phase: irradiated DNA that is not able to repair itself will fail to replicate during mitosis thus leading to cellular death. Cells can also die before or after attempting mitosis. Cancer cells go through the cell cycle faster than normal cells and the probability of radiation being delivered during the M phase is higher [20]. Radiotherapy takes advantage of this to maximise cancer cells' death while minimising normal tissue damage.

The linear-quadratic (LQ) model (Figure 2.5) is a mathematical representation of the radiation dose and cell survival relationship: $S = e^{(-\alpha D - \beta D^2)}$, where α and β are parameters describing the cell's radiosensitivity and D is the radiation dose absorbed by the cell. The shape of the LQ survival curve is determined by the α/β ratio, which is the dose level at which the linear (αD) and quadratic (βD^2) contributions to damage are equal. The linear contribution, dominant at low doses, is related to lethal cell damage caused by a single incident particle ('single hit' cell death) and the quadratic contribution, dominant at high doses, can be attributed to lethal cell damage from different radiation interactions ('multiple hit' cell death) [28]. The LQ model is widely used in the clinic to estimate equivalent RT fractionation schedules (e.g. EQD2) but also to predict tumour control probability (TCP) and normal tissue complication probability (NTCP) [29]. EQD2 (Equation 2.1) is the equivalent dose in 2 Gy fractions, as derived from the LQ model. Any fractionation schedule can be translated into EQD2 using this equation, where D is the total dose prescribed, d is the dose per fraction and the α/β ratio is the dose at which the linear and quadratic components of the LQ model are equal.

$$EQD2 = D\left(\frac{d + (\alpha/\beta)}{2 + (\alpha/\beta)}\right) \tag{2.1}$$

2.3.2 Normal tissue complication probability (NTCP) models

The escalation of the radiation dose to the tumour is limited by the complications that may develop from the irradiation of surrounding normal tissue [30, 17]. There is, however, a 'therapeutic window' (Figure 2.6) in which the TCP is larger than the NTCP, thus providing an opportunity for treatment optimisation. NTCP models aim at characterising the shape of response of normal tissue for a given radiation dose distribution. NTCP models are used

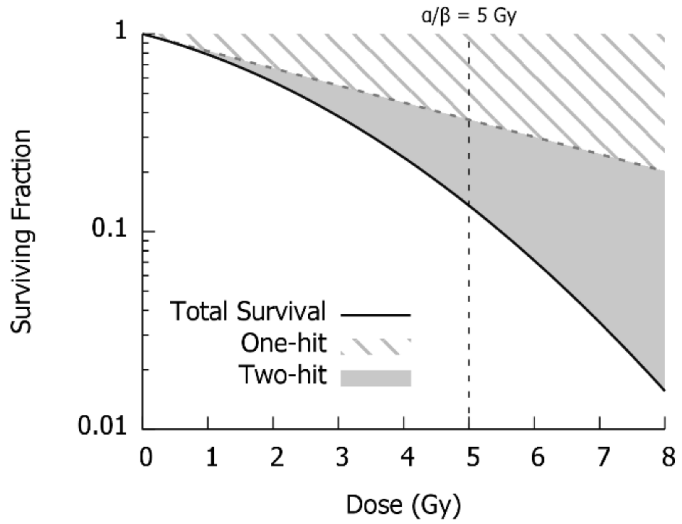


Fig. 2.5 Linear quadratic cell survival model. Figure source: Lyman [29]

as a clinical decision support system [31] to reduce the incidence of a given toxicity by identifying the patients who are at a higher risk of developing it [32, 33]. The two main types of NTCP modelling [34], analytical models and data-driven models, are described below.

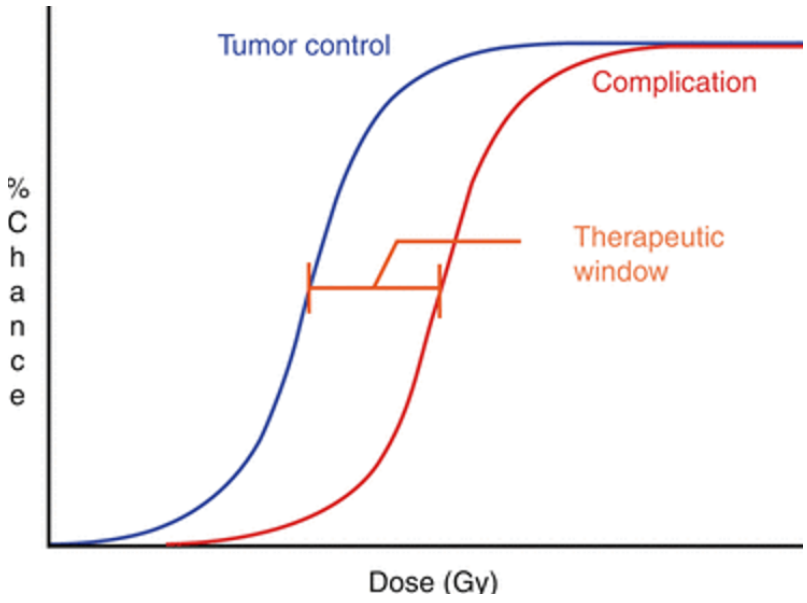


Fig. 2.6 The probabilities of tumour control (TCP) and of normal tissue complication (NTCP) increase with the amount of radiation dose delivered to a tumour (and inevitably to the surrounding normal tissue). The 'therapeutic window' gives a compromise between the two, with the TCP being larger than NTCP in the region. Figure source: Chang et al. [28].

Analytical NTCP models. Analytical NTCP models are based on mathematical equations that explain the theoretical assumptions made on the relationships between the toxicity outcomes and the input variables. The most widely known analytical NTCP model is the Lyman-Kutcher-Burman (LKB) model (Equation 2.2), which uses DVH-based dose metrics [29]. Although the LKB model does not consider the tissue heterogeneity and functional architecture aspects described above, it remains widely used in the clinic [35]. While the LKB model is based on a single-dose variable, the widely used multivariable logistic regression analytical model (Equation 2.3) can incorporate several variables [36].

$$NTCP_{m,D_{50}}(x) = 1/\sqrt{2\pi}\int_{-\infty}^{x} e^{(-u^2)/2}du \tag{2.2}$$

$$\text{where } x = (D - D_{50})/(mD_{50})$$

$$NTCP = 1/1 - e^{-s} \tag{2.3}$$

$$\text{where } s = \beta_0 + \beta_1 variable_1 + \ldots + \beta_n variable_n$$

Data-driven NTCP models. Data-driven methods, as opposed to the analytical models, learn non-linear relationships directly from the data without the need for (potentially biased) a priori assumptions of how the toxicity outcomes and input variables are related. Data-driven NTCP models can be classed into the more traditional ML methods and DL methods such as CNN, both described in more detail in Chapter 3.

2.3.3 Limitations of DVH-based NTCP models

The relationship between dose-volume and normal tissue toxicity is often complex [37, 38] because a) dose distributions to normal tissue surrounding the tumour are heterogeneous, b) organ function is not uniformly distributed within an organ [39], c) organs at risk often have sub-regions that may respond differently to radiation, d) in addition to radiation dose response, prediction of normal tissue response involves several other factors such as patient characteristics and clinical variables. DVH-based NTCP models are not able to capture these complexities [35]. Chapter 3 (Section 3.3) discusses methods to manually extract spatial dose features from radiation dose distribution maps as well as the more novel DL-based alternatives.

2.4 Mandibular osteoradionecrosis

Mandibular ORN is a rare but severe late radiation-induced toxicity observed in 4-8% [40] of HNC patients treated with RT. Incidence of ORN in the population of patients treated at GSTT is 5.5% overall in HNC [41]; however, for certain groups such as oropharyngeal or oral cavity cancer this can be around 10% [42, 43]. The most commonly used definition of ORN in the UK is [44] 'an area of exposed irradiated bone that fails to heal over a period of three months without any evidence of persisting or recurrent tumour' [45]. Necrosis of the bone can develop either spontaneously after irradiation or triggered by trauma to the irradiated mandible bone (e.g., dental extractions, surgery, implants). There is no clear consensus on a theory that explains how ORN develops. A first theory claimed that bone necrosis occurs as a result of reduced blood supply, hypoxia and hypo-cellularity caused by exposure to radiation [46]. A more recent theory suggests that radiation-induced fibrosis of the irradiated tissues extends to the blood vessel walls eventually resulting in a reduced blood supply and subsequent necrosis of the bone tissue [47, 48]. The intrinsic anatomical heterogeneity of the mandible may influence the localisation of the ORN regions. The lower jaw is more prone to ORN as its bone is more cortical and therefore has reduced vascularity with respect to the upper jaw. Moreover, ORN is also more likely in the posterior jaw as radiation doses tend to be higher posteriorly.

2.4.1 ORN staging

Mandibular ORN can be diagnosed by physical examination (e.g., Figure 2.7) and/or radiologically (based on 2D or 3D imaging). There are numerous systems to classify the severity of ORN [45]. The National Cancer Institute Common Terminology Criteria for Adverse Events (NCI CTCAE) [49] and the Notani [50] scales are commonly used at GSTT. Tables 2.1 and 2.2 describe these two ORN staging systems in detail.

2.4.2 Risk factors for mandibular ORN

ORN has a multifactorial aetiology with radiation dose, clinical and demographic information as risk factors [51–53, 40, 44]. A number of case-control studies have investigated the correlation between dosimetric, clinical and demographic factors and ORN [54, 55, 41, 56–61, 43, 62, 1, 63]. Radiation dose is a major risk factor for ORN and optimisation of the RT dose distribution to the mandible should be based on robust and well supported and validated dosimetric constraints. Table 2.3, adapted from Brodin et al. [32] and De Felice et al.

Fig. 2.7 Example of a mandibular ORN case at GSTT. This image shows exposed bone in line with ORN of the upper right second premolar tooth socket.

Table 2.1 The NCI CTCAE v5.0 ORN scale

Grade	Description
Grade 1	Asymptomatic; clinical or diagnostic observations only; intervention not indicated.
Grade 2	Symptomatic; medical intervention indicated (e.g. topical agents); limiting instrumental activity of daily living.
Grade 3	Severe symptoms; limiting self-care activity of daily living; elective operative intervention indicated.
Grade 4	Life-threatening consequences; urgent intervention indicated.
Grade 5	Death

Table 2.2 The Notani ORN scale

Grade	Description
Grade I	ORN confined to the alveolar bone.
Grade II	ORN limited to the alveolar bone and/or the mandible above the level of the mandibular alveolar canal.
Grade III	ORN that extended to the mandible under the level of the mandibular alveolar canal and ORN with a skin fistula and/or a pathologic fracture.

[53], summarises the findings on radiation dose correlations with ORN incidence by some of these studies. There is no clear consensus on a dose threshold for ORN development risk. Early ORN studies report significant differences in mean and high doses between the ORN and control groups. More recent studies have highlighted the role of intermediate and low radiation dose levels. Patel et al. [64] showed that microvasculature collapse in the mandible occurs even at radiation dose levels of 30 Gy. In a previously presented analysis [65] on a 70 ORN and 140 controls GSTT cohort, 10% of the ORN cases had a D_{mean} in the ORN region below 50 Gy and in 6% of the ORN cases it was below 45 Gy. Gomez et al. [54] similarly concluded that in some cases the ORN region did not correspond to the mandible area receiving the maximum dose. Aarup-Kirstensen et al. [61] found significant differences between the ORN and control groups at doses between 30 Gy and 60 Gy. A study by the MD Anderson HNC Symptom Working Group [59] with 68 ORN patients and 131 matched controls suggested that the volume of a lower dose level might have contributed to ORN more significantly than a maximum point dose due to the resulting larger volume of microvasculature damage. A study [1] with the largest ORN cohort to date found all dose-volume parameters, including the low dose levels, significantly higher in the ORN group.

Smoking, alcohol and poor dental hygiene are included in most published studies as lifestyle related risk factors. Neglected dentition can result in dental diseases, which has been associated with an increased risk of ORN [66, 44, 52]. For HNC patients, a pre-RT dental assessment is recommended by the National Institute for Health and Care Excellence (NICE) guidelines [49]. However, previous studies [55, 64] have discussed an increased incidence of ORN in the HPV-associated OPC group of patients, that are generally younger, with better dental status and without the lifestyle factors associated with ORN (e.g., smoking, alcohol). Table 2.4 provides a summary of the potential risk factors for ORN that have been considered clinically and in research.

Table 2.3 Review summary of published studies on dosimetric associations with ORN (adapted from Table 5 in [32] and [53]). Studies including treatment modalities other than IMRT (predominantly) were excluded from the original table and more recent studies were added.

Study cohort	**Dosimetric factors**	**Reference**
168 OCC, nasopharynx, larynx/hypopharynx, sinus and OPC patients, out of which only 2 ORN cases (OCC primary site)	Median $D_{max} = 67.98$ Gy, median $D_{mean} = 38.45$ Gy in ORN cases.	Gomez et al. [54]
36 HNC patients who developed ORN after RT or chemo-RT in 2 Gy/fx	69% of ORN cases had a mandible D2% $>$ 60 Gy and 52.8% of cases had a $D_{mean} > 60$ Gy.	De Felice et al. [41]
68 OPC patients with ORN and 131 matched controls, both groups treated with IMRT	Mandibular V44Gy $< 42\%$ and V58Gy $< 25\%$ were identified as significant cutoffs with 81% of ORN cases identified in patients with mandibular V44Gy $\geq 42\%$ and V58Gy $\geq 25\%$.	Mohamed et al. [59]
44 oropharyngeal or oral HNC patients with ORN and 78 matched controls with treated with IMRT at 1.6 – 2.12 Gy/fx	Mandible D_{max} 13.2 Gy and Dmean 14.6 Gy average differences between the ORN side and the contralateral non-ORN part	Owosho et al. [58]
14 ORN patients and a matched group of 14 controls out of a population of 252 OCC and OPC patients, all predominantly treated with IMRT	No significant dosimetric differences between the ORN and control groups.	Moon et al. [60]

Continued on next page

Table 2.3 – continued from previous page

Study cohort	Dosimetric factors	Reference
1196 oropharyngeal HNC patients (77 ORN cases) treated with IMRT or chemo-IMRT at 1.2-2.6 Gy/fx	Absolute mandibular V50Gy and V60Gy were significant indicators of ORN with average V50Gy $= 35.8cm^3$ vs. $30.8cm^3$ and V60Gy $= 18.9cm^3$ vs. $15.3cm^3$ in cases vs. controls, respectively.	Caparrotti et al. [57]
1:2 matching in a nested case-control study with 56 ORN cases out of 1224 HNC treated predominantly with IMRT	D_{mean} significantly higher in ORN group (41.7 Gy vs. 37.7 Gy). Significant differences between ORN and control groups for doses between 30 Gy and 60 Gy.	Aarup-Kirstensen et al. [61]
46 ORN cases out of a 616 HNSCC population	V30Gy-V70Gy significantly higher in ORN group with V60Gy > 14% as an independent risk factor.	Kubota et al. [62]
173 ORN cases out of 1259 HNC patients	All DVH parameters significantly associated with ORN in univariate models. D30% $<$ 42Gy and D30% $<$ 35Gy to achieve $<$ 5% risk of ORN_{I-IV} for patients without and with pre-RT extractions, respectively; D30% $<$ 25Gy and D30% $<$ 17Gy for $<$ 5% risk of ORN_{IV}.	Van Dijk et al. [1]

Continued on next page

Table 2.3 – continued from previous page

Study cohort	Dosimetric factors	Reference
46 ORN cases out of 227 OCC patients.	V60Gy significantly associated with ORN at univariate and multivariate analysis. Dmean significant at univariate analysis.	Möring et al. [63]

Table 2.4 Demographic, clinical and treatment ORN risk factors

Type of risk factor	Risk factor	Description
Patient-related factors	Gender	Male predominance (factor of 3:1) [51, 43]
	Age	Common at age of 55 ± 10 years [44]. HPV-associated OPC cases tend to be even younger.
	Smoking	Smoking status was found a significant risk factor [57, 60, 63]; 32% increased risk for patients who continue to smoke during RT [44].
	Alcohol	Increased risk of ORN with excessive alcohol consumption [58, 44].
	HPV	Increased ORN incidence in HPV-associated OPC cases [43].
Tumour-related factors	Hystopathology	Increased ORN incidence in squamous cell carcinoma (SCC) cases [52].
	Stage	Higher SCC stages have been associated with higher radiation doses to the mandible [64]. An increased risk of ORN has been reported in higher SCC stages [51, 52].

Continued on next page

Table 2.4 – continued from previous page

Type of risk factor	Risk factor	Description
	Size	Larger tumours have been associated with higher radiation doses to the mandible [44] and a higher ORN incidence [52].
	Primary tumour site	The majority of ORN cases are treated for OCC and OPC cancer [52]. Tumour-to-bone proximity results in higher risk of ORN [56]. Primary tumour site was found to be an independent risk factor for ORN in a study by Kubota et al. [62].
Treatment-related factors	Radiotherapy technique	With IMRT it is possible to conform the high dose to the target volume better than with 3DCRT. However, IMRT results in a more extensive low-dose bath that incorporates larger jaw volumes than with 3DCRT [44]. In Moon et al. [60], 3DCRT resulted in significantly higher ORN incidence than IMRT, with higher D_{max}, V60Gy and V70Gy dose-volume levels in the 3DCRT ORN cases.
	Radiotherapy dose	QUANTEC [35] does not provide recommendations on radiation dose-volume constraints for the mandible and there is no clear consensus on a dosimetric threshold as an ORN development risk. An increased risk at doses above 40 Gy is generally accepted [44]. Table 2.3 summarises the different dose or dose-volume levels proposed in existing literature.

Continued on next page

Table 2.4 – continued from previous page

Type of risk factor	Risk factor	Description
	RT-induced complications	RT may cause other complications which, in turn, may result in an increased risk of ORN. Xerostomia might affect the oral environment; trismus might result in limited oral access; dysphagia might require the intake of high caloric liquid food supplements. All of these may result in dental decay and, consequently, in an increased ORN risk.
	Chemotherapy	Chemotherapy has been associated with an increased risk of ORN [44].
	Surgery	Surgery has been associated with an increased risk of ORN [56, 61, 43].
Dental factors	Oral hygiene and dentition	Poor dental hygiene and neglected dentition can result in dental diseases, which has been associated with an increased risk of ORN [52, 56, 64, 44].
	Dental extractions	Dental extractions have been associated with an increased risk of ORN [52, 60, 61, 1, 43]. Nabil et al. [12] concluded that the highest ORN risk corresponds to extraction of mandibular teeth within a region that has received >60 Gy.

2.4.3 Prediction of mandibular ORN

The current clinical practice for HNC RT at GSTT is largely based on the Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) recommendations [35] for organ at risk radiation dose-volume constraints. However, in the QUANTEC report only a limited set of toxicities is included and no specific recommendations are made to prevent mandibular ORN [32]. In 2017, a thorough systematic review on HN NTCP models by Brodin et al. [32] could not include any reference to an existing NTCP model for ORN. As discussed in Section 2.4.2, published studies have largely focused on identifying the associations between risk factors and the development of ORN. Efforts on patient-specific prediction of ORN, however, are more limited. The need for personalised plan optimisation to reduce mandibular ORN has been acknowledged in a recent review by De Felice et al. [53]. Prediction of potential ORN in the treatment of HNC may lead to risk-reduction measures (e.g., reduced mandibular radiation dose near extraction sites when possible) and/or a more dedicated follow-up for early detection and intervention of ORN. We have published the first study [67] to explore patient-specific ORN prediction using ML methods; this work is further described in Chapter 5. Shortly after, an NTCP model for ORN was published by van Dijk et al. [1]. Their NTCP model was based on mandible dose-volume parameters and clinical variables using multivariable stepwise forward selection regression analysis. The resulting final model was based on the D30% of the mandible bone and pre-RT dental extraction and had a validation performance of AUROC=0.75 and AUROC=0.82 for the prediction of ORN_{I-IV} and ORN_{IV} stages, respectively.

2.4.4 ORN management

Mandibular ORN is not as common as other HNC toxicities and has consequently attracted less research attention. However, the treatment of ORN is often complex and costly [48]. There remains no agreed management approach [53] with varying modalities depending on its severity. Conservative treatment options such as observation, ultrasound therapy, hyperbaric oxygen therapy or medical management (pentoxifylline, tocopherol, clodronate) may be considered for less severe cases of bone exposure cases [53, 42]. More severe cases might require surgical intervention such as the mandibulectomy and bony free-flap reconstruction procedures [53], although there are concerns that surgery may worsen the condition and in advanced stages may be unable to improve quality of life [68].

2.5 Discussion

RT is the mainstay of curative options for HNC. ORN of the mandible is a rare but severe RT-induced side effect that not only has a detrimental impact on patients' quality of life but often also requires costly clinical interventions. NTCP models are a clinical decision support system to reduce the incidence of a given toxicity by identifying the patients who are at a higher risk of developing it. NTCP models have traditionally used DVH metrics, which are limited and lack the spatial information included in radiation dose maps. Chapter 6 explores the use of radiation dose maps as an alternative to DVH metrics in the prediction of ORN using DL methods. This Chapter has provided a review on existing ORN studies. However, this comparison is often difficult due to the large diversity of cohorts, treatment techniques and inclusion criteria considered. Moreover, each of these studies are based on very limited datasets due to the naturally low prevalence of ORN. Consequently, the PREDMORN study was developed, which will result in ORN prediction modelling with the largest and most diverse ORN cohort ever published before. More details on the published study protocol are provided in Chapter 9.

Chapter 3

Technical Background

This Chapter provides a technical background to the ML and DL methods used in the experiments presented in the subsequent chapters. Section 3.1 presents the main concepts in ML and DL, with Section 3.2 focusing on model evaluation with small datasets. Section 3.3 discusses existing work on the use of DL methods for the prediction of radiation-induced toxicities.

3.1 Machine and deep learning

As described by Tom Mitchell in 1997 [69], ML algorithms 'learn from experience (E) with respect to some class of tasks (T) and performance measure (P), if its performance at tasks T, as measured by P, improves with experience E'. Traditional ML methods learn a mapping from hand-crafted features of the input data to the desired output. In more complex ML methods such as deep CNNs, the feature extraction is embedded in the network processes. Collectively, ML and DL methods are commonly referred to as artificial intelligence (AI), although the correct use of the term AI also encompasses other, non-learning-based 'intelligent' techniques. This section describes the most relevant concepts of AI to this thesis.

3.1.1 Supervised, semi-supervised, unsupervised and reinforcement learning

ML algorithms can be classed into four types based on the nature of the training data and the way in which it is used. In supervised learning, the most common type in

radiation oncology [38], the algorithm is presented with a training set of input data and the corresponding output labels or 'ground truth'. The algorithm is expected to learn the mapping between the training inputs and outputs to then be able to predict the outputs on an unseen input dataset. The error between the predicted outputs and the ground truth is used to find the optimal model parameters. Further details on this optimisation process and on the training of supervised models is provided in this Chapter in Section 3.1.4. In unsupervised learning the algorithm is presented with unlabelled data and is expected to learn by itself about the structure (e.g., correlations, patterns, features) of the input dataset. Semi-supervised learning lies between the two algorithm types previously described, with only part of the training data being labelled. Finally, reinforcement learning [70] refers to the fourth type of algorithm, where the algorithm learns to map inputs to actions by maximising (minimising) a reward (punishment) action evaluation signal.

3.1.2 Regression, classification and segmentation algorithms

While ML algorithms can perform a variety of tasks [71], the most common ones in radiotherapy applications are segmentation, regression and classification. Regression algorithms are used to produce a function $f(x_i)$ that best describes the relationship between the input variables (x_i) and the continuous output variable $y = f(x_i)$ [71]. This type of algorithm has been used in classical data-driven TCP and NTCP modelling to predict the probability of the treatment and toxicity outcomes, respectively [72]. Classification algorithms are another type of machine learning algorithm that can predict which of k categories or discrete class labels an input belongs to. Both regression and classification algorithms can be trained on an image level or a pixel level using CNNs.

Segmentation tasks are achieved with pixel-wise regression algorithms that predict the probability of each pixel being in each class; these probabilities are then converted to a one-hot-encoded representation for the final classification. Segmentation algorithms are clinically used for automatic delineation of anatomical treatment target volumes and organs at risk. The interested reader is directed to the recent review by Harrison et al. [73] for a more in-depth discussion on the use of ML for auto-segmentation in RT.

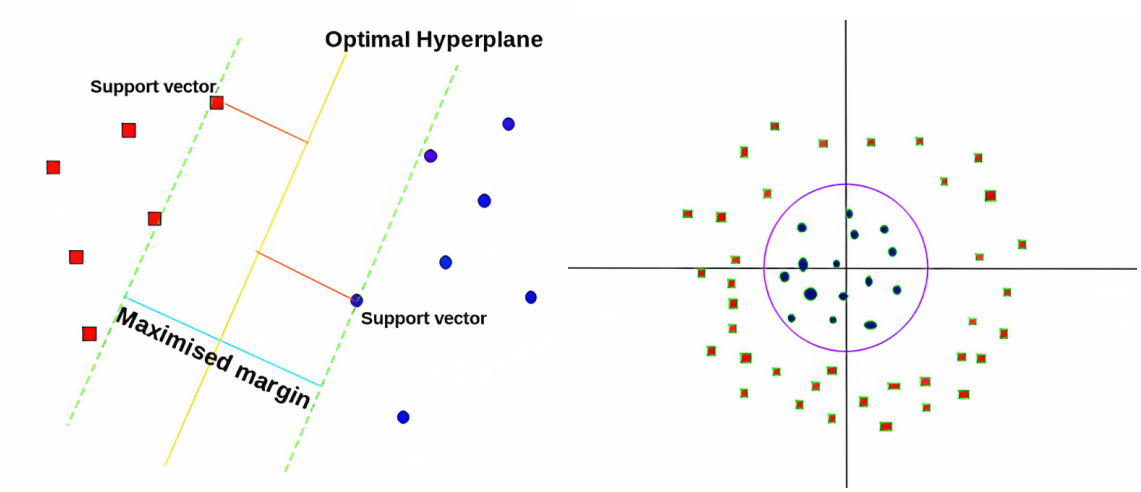
3.1.3 Supervised classification machine learning methods

Supervised classification algorithms are the most relevant type of model to this thesis. This type of algorithm learns to map the training input data to the output classification labels. The trained algorithm is then able to predict the class labels on an independent test dataset.

Below I describe some of the most commonly used algorithms for radiation-induced toxicity prediction [74], including logistic regression (LR), support vector machines (SVM) and two algorithms that use the ensemble learning technique: random forests (RF) and adaptive boosting (Adaboost). These algorithms have been used in the work described in Chapters 5 and 6 of this thesis.

Logistic regression [75] tries to fit a straight line in the input feature space that separates data according to its class. The probabilities of obtaining classes $y = 1$ and $y = 0$ given the input variables (i.e. features) x and model parameters θ is described by the sigmoid or logistic function $\sigma(z)$ as $P(y = 1|x) = \sigma(z)$ and $P(y = 0|x) = 1 - \sigma(z)$, where $\sigma(z) = 1/(1 + e^{(-z)})$ and $z = \theta_0 + \sum_{i=1}^{m} \theta_i x_i$. During training, the model aims to obtain the best prediction of the training set of labels by optimising its parameters θ.

Support vector machines [76] are a type of ML algorithm that uses a set of mathematical functions (a.k.a. kernels) to find the optimal hyperplanes that separate the input data into two (or more) classes. Like LR, SVM is a linear classifier. However, SVM can be extended to perform nonlinear classification (Figure 3.1) through the use of kernels, which can transform data which is not linearly separable into a higher-dimensional space where they are. The degree of acceptable misclassification is defined with the C penalty parameter, where smaller C values result in higher misclassification error rates. When using the radial basis function (RBF) kernel, the gamma hyperparameter can be tuned to find the optimal curvature of the decision boundary [77].



(a) (b)

Fig. 3.1 Diagram of a SVM optimal hyperplane in a linear (a) and non-linear (b) data distribution. However, the latter can be linearly classified by tuning the parameters of the SVM kernels. Image source: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989.

Ensemble learning [78] is a technique used to improve the overall classification accuracy of a ML model by training and combining the outputs of a set of T models to obtain an ensemble probability $P_{ensemble}(y|x) = \sum_{t=1}^{T} w_t P_t(y|x)$, where $P_t(y|x)$ is the predicted probability of class y given input x by model t. When the outputs are class labels rather than probabilities, these can be combined using majority voting, where the ensemble output class is defined as the class with the most votes. When the weights w_t are uniform, all the models are combined in simple uniform averaging or voting. Otherwise, the final ensemble model is a weighted sum of all the trained models with the highest weight given to the better performing one.

Ensemble learning algorithms can either use *bagging* or *boosting* as methods to introduce diversity amongst the ensemble members. In bagging, independent models are trained in parallel and combined at the end for the final decision. A **Random Forest** [79] is a bagging ensemble ML algorithm that combines decision trees that have been trained in parallel (Figure 3.2). Each tree is constructed based on a different randomly selected subset of the training data. In boosting, different models are trained sequentially, with each new model trained on an updated training dataset that gives more emphasis on the cases that have been misclassified in the previous round. The **Adaptive Boosting** algorithm [80] was the first boosting ensemble ML algorithm and it adaptively re-assigns the highest weights to the incorrectly classified data (Figure 3.3).

3.1.4 Artificial neural networks

An artificial neuron or perceptron (Figure 3.4), a concept first introduced by Rosenblatt et al. [81], is the simplest unit of an artificial neural network. An artificial neural network [82] consists of an input layer, an output layer and a hidden layer with multiple interacting artificial neurons. A deep artificial neural network (ANN) or multilayer perceptron (MLP) contains several hidden layers.

There are two main phases involved in the learning process of an ANN. During the forward propagation step or prediction phase, the ANN maps its inputs x_n to an output $y = f\sum_n(w_n x_n + b)$, where w_n are the learned coefficients or weights, b is the bias term and f is the activation function (Figure 3.4). Activation functions are used to nonlinearly transform the output, e.g., the softmax activation function [71] is used in the last layer to convert the raw output value into a probabilistic output per class with probability values between 0 and 1. The more common activation function in other layers is the Rectified Linear Unit (ReLU) (Figure 3.5). The interested reader is directed to Lederer et al. [84] for a review of the most commonly used activation functions.

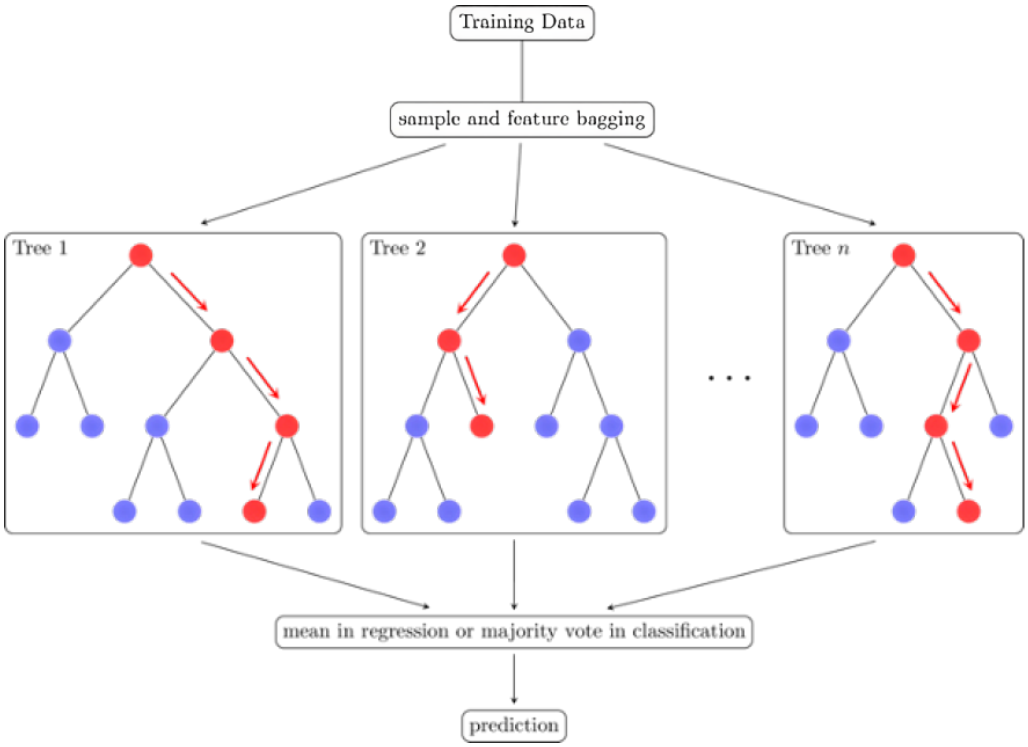


Fig. 3.2 Diagram of a random forest and how the predictions of all its decision trees are combined to obtain the final prediction. Image source: https://tikz.net/random-forest/.

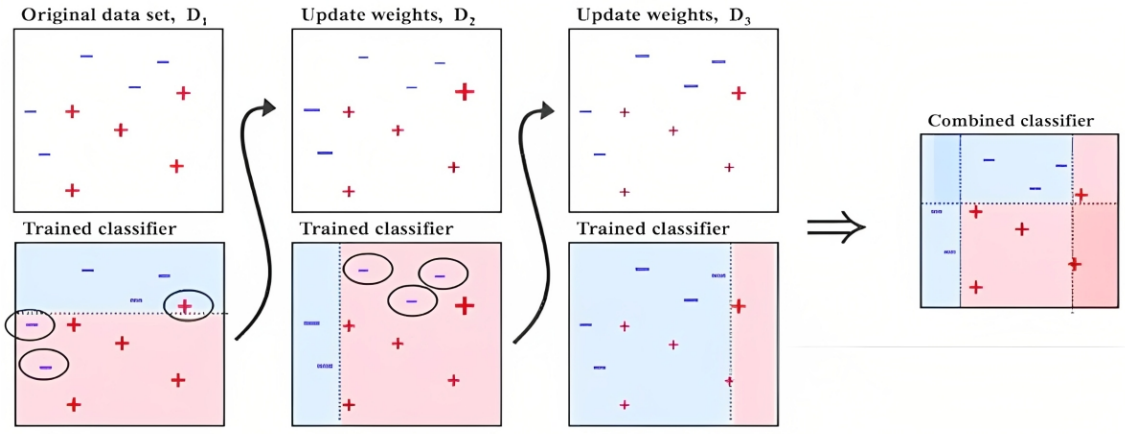


Fig. 3.3 Diagram of how the AdaBoost classifier re-weights the data at each model training until the final decision is obtained. Figure source: https://towardsdatascience.com/understanding-adaboost-for-decision-tree-ff8f07d2851.

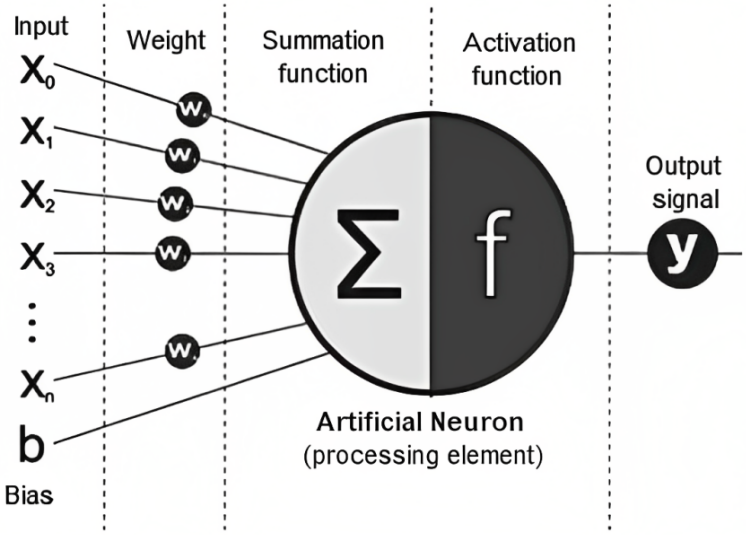


Fig. 3.4 Schematic representation of the mathematical model of an artificial neuron or perceptron. Figure source: Sarker et al [83].

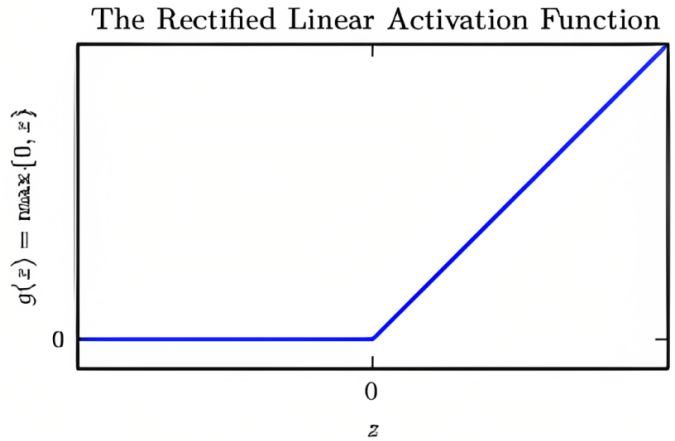


Fig. 3.5 The ReLU activation function. Image source: Goodfellow et al. [71].

The training phase (Figure 3.6) aims to obtain the optimal network weight and bias values. The predicted output is compared to the ground truth and the difference between the two, calculated with a cost function, is used to guide the process of adjusting the weights. The backpropagation process aims to minimise this cost function, i.e., to find the global cost minimum, by calculating the gradient (i.e., partial derivatives) of the cost function with respect to the weights and biases of the network [85]. The gradient descent optimisation algorithm uses this gradient to update the weight and bias values at each iteration as per $w' = w - a \times gradient$, where w' and w are the new and old weights, a is the learning rate constant at which the gradient is applied and *gradient* is calculated during backpropagation. The Adam (adaptive moment estimation) optimisation algorithm is widely used in training modern DL models. Instead of using a constant learning rate, the Adam optimiser updates the learning rate for each network weight individually during training. The interested reader is directed to Goodfellow et al. [71] for a description of other DL optimisation algorithms.

In addition to the learning rate, other model settings or hyperparameters that can be modified or tuned during the training phase include the number of epochs (i.e. number of times that the network sees the entire dataset) and the batch size (i.e. number of subsamples from the entire dataset that the model uses at each weight update or *iteration*. Furthermore, regularisation methods such as dropout and weight decay may be included in order to improve the model generalisation error [71], with the dropout rate and weight decay parameters included as hyperparameters. Section 3.2.1 in this Chapter describes hyperparameter tuning and model selection processes.

3.1.5 Binary classification with a deep CNN

DL is a subfield within ML that involves learning data features using ANNs with several layers of mathematical operations [86]. Deep CNNs are the most widely used DL method when the input data is high-dimensional (e.g. images). Transformer networks [87, 88] are a recent and promising attention-based alternative to CNNs to achieve DL with high-dimensional input data. This section, however, describes the operations involved in a CNN and describes the 3D DenseNet and 3D ShuffleNet architectures as examples of CNN models, both of which are used in the experiments detailed later in this thesis.

CNNs [89] include convolution and pooling layers to automatically extract the most useful features from input images to perform a specific task [38]. Both the convolution and pooling operations contribute to a reduction of the number of connections between layers, thus reducing the computational cost of the network, which is essential when dealing

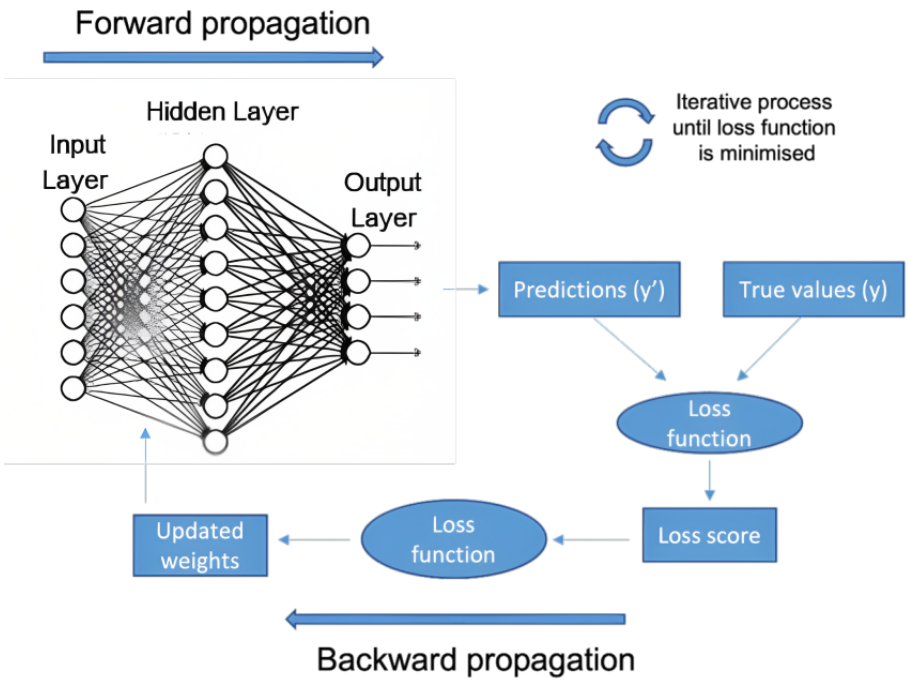
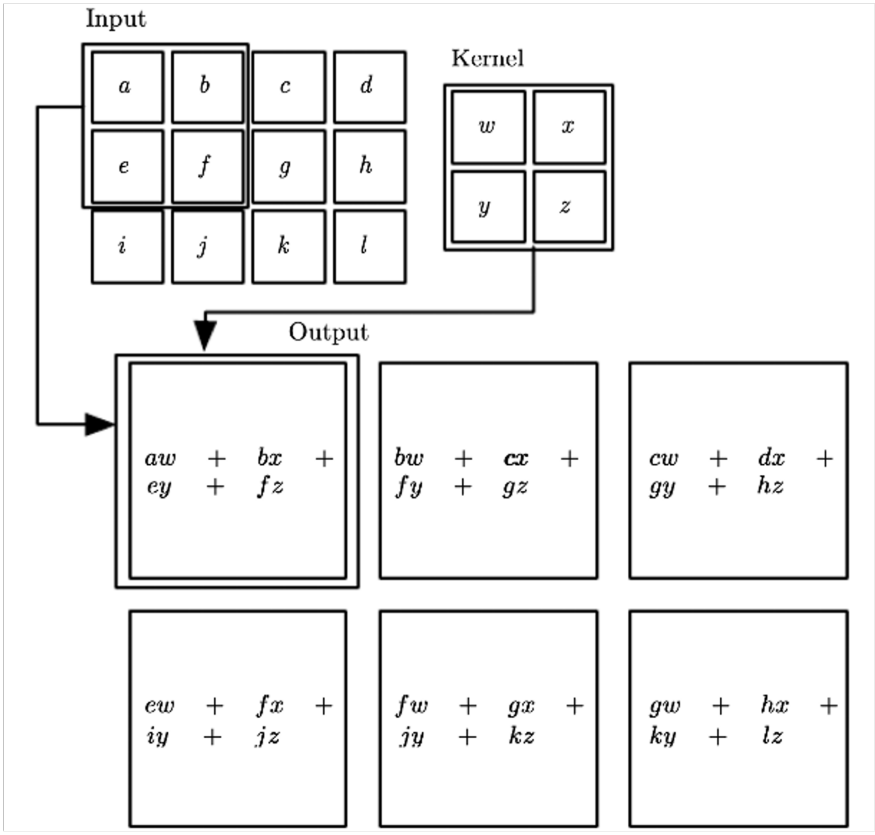


Fig. 3.6 Prediction and training processes in an ANN. Figure adapted from Sarker et al. [83].

with high-dimensional data such as images. In a convolution layer of a CNN, a filter (or convolution kernel) is scanned over the input tensor (image) with a tensor product operation (Figure 3.7) to produce an output tensor (or feature map) that represents a map of the presence of a pattern (described by the kernel) at different locations in the input image [86]. In a pooling layer the size of the feature map is downsampled, which reduces the total number of parameters that the network needs to train. Max pooling and average pooling are downsampling approaches that use the maximum and average values, respectively, in each patch of a feature map (Figure 3.8). A CNN can be trained for classification tasks by adding a fully connected layer that connects the output feature maps from the convolution and pooling layers to the classification outputs (classes) (Figure 3.9). Below is a description of the two CNN architectures that I used in this thesis: 3D the DenseNet and the 3D ShuffleNet.

3D DenseNet classification CNN. Deeper CNNs are more sensitive to small details and less sensitive to larger irrelevant variations in an image [85]. However, going deeper (more layers) has some drawbacks such as the partial derivative (gradient) of the loss function becoming so small after a long path between the input and the output that the network stops learning (a.k.a. vanishing gradient problem) or the issue of overfitting due to an increased number of model parameters [92]. Densely connected convolutional



$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n) K(m, n)$$

Fig. 3.7 Schematics of a 2D convolution and the corresponding convolution operation equation, where K corresponds to the kernel, I is the image and S is the output. Figure adapted from Goodfellow et al. [71].

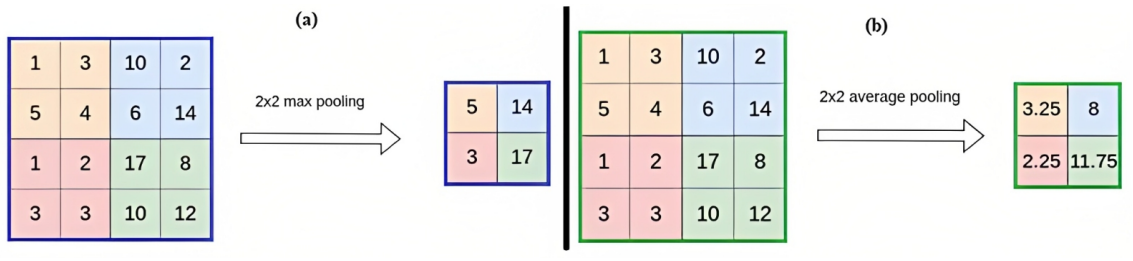


Fig. 3.8 Max and average pooling operations. Figure adapted from Vasilev et al. [90].

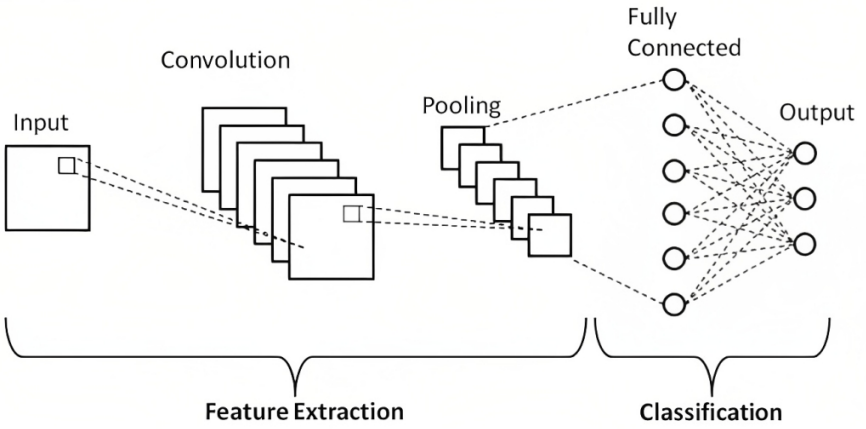


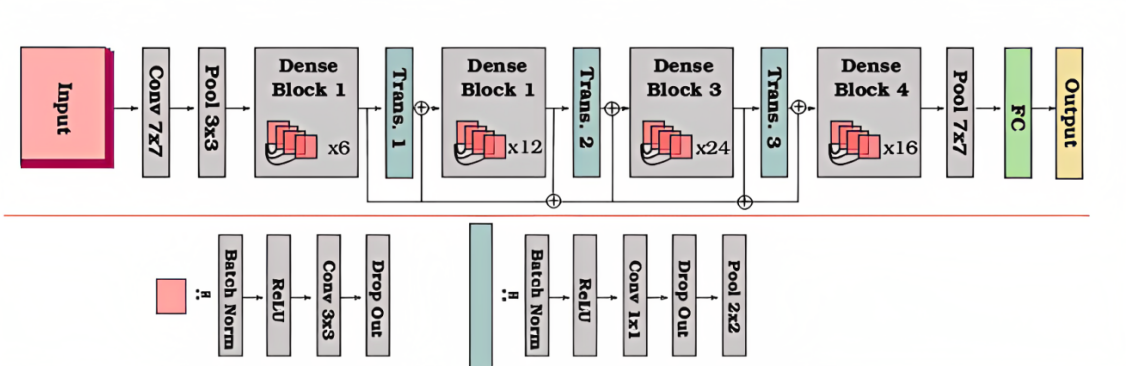
Fig. 3.9 Schematics of a classification CNN architecture. Image source: Phung et al. [91].

networks (DenseNets) were introduced by Huang et al. [93] as a novel approach to deeper CNNs while addressing these drawbacks. DenseNets consist of dense blocks connected via transition layers (Figure 3.10a). In a dense block, every layer is directly connected to each other (Figure 3.10b), thus ensuring maximum information and gradient flow across all layers as well as requiring fewer parameters as feature maps from other layers are reused in subsequent layers (e.g. the DenseNet-121 has around 7 million parameters whereas the widely used ResNet-18 and VGG-16 models have 11 million and just over 130 million parameters, respectively). In Chapter 6 I have used and compared the DenseNet121 and DenseNet40 versions, the latter being a much lighter version with just around 1 million parameters and a subsequently shorter training time (approximately 2 days vs. 5 days for a nested-CV ensemble training).

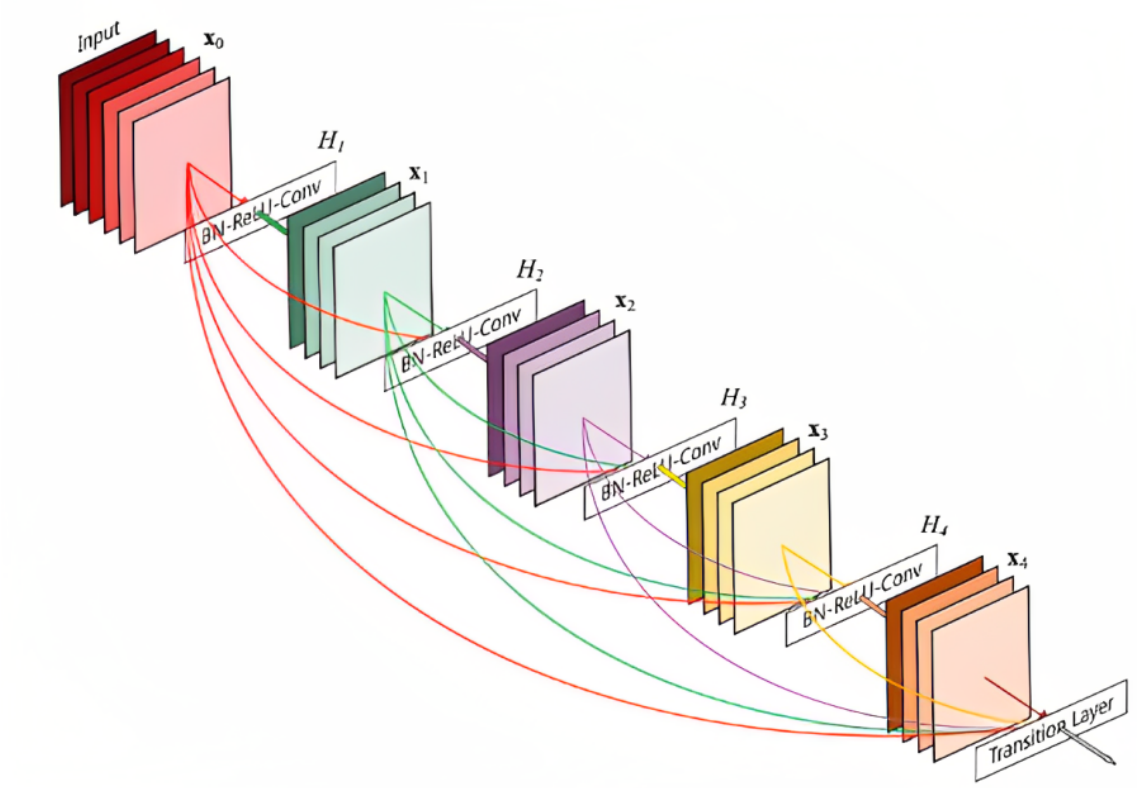
3D ShuffleNet classification CNN. The ShuffleNet CNN [94] is a 'lightweight model' that uses grouped convolutions and the channel shuffle operation (Figure 3.11) to accelerate the training process. In a study by Yang et al. [95] where different deep CNNs were compared with regards to their performance on small datasets, the ShuffleNet model, with only 343,842 parameters, showed similar performance to the DenseNet-121 model. In Chapter 6, I compare the performance of both networks in the prediction of ORN using radiation dose distribution maps.

3.2 Evaluation of ML models with limited data

Prediction models are often developed on limited datasets and an evaluation of their performance on independent datasets is required prior to clinical use. According to the TRIPOD (Transparent Reporting of a multivariable Prediction model for individual



(a)



(b)

Fig. 3.10 a) Schematics of a DenseNet-121 architecture, where dense blocks and transition layers are represented in red and green, respectively. b) Visual representation of a 5-layer dense block with the connections between each layer and its preceding feature maps. Image sources: https://www.pluralsight.com/guides/introduction-to-densenet-with-tensorflow and Huang et al. [93].

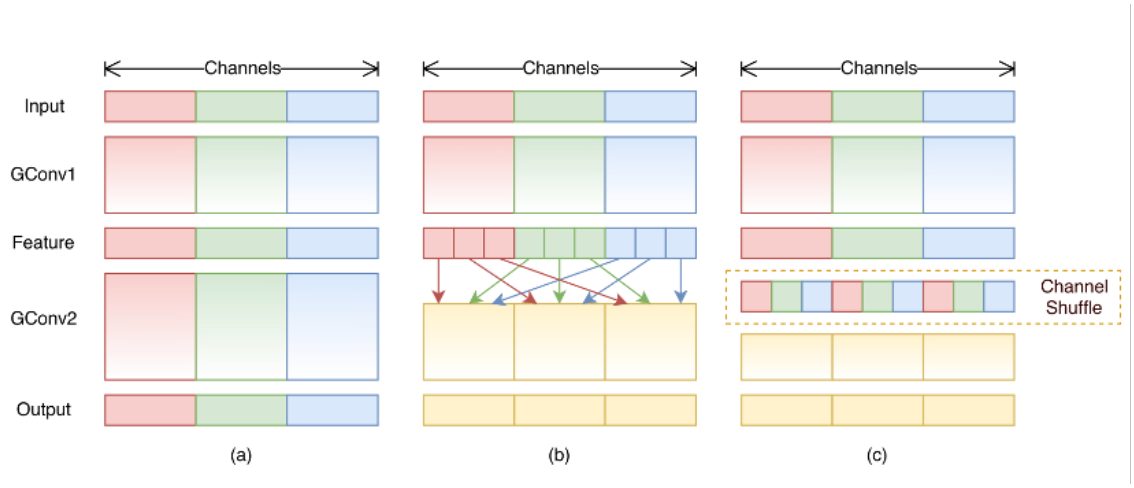


Fig. 3.11 Schematics of a ShuffleNet with two group convolutions (GConv) comparing the following scenarios: a) there is not cross talk between the groups, b) interaction between all input and output channels in GConv2 and c) equivalent interaction by applying the channel shuffle operation. Figure source: Zhang et al. [94].

prognosis Or Diagnosis) statement [96], prediction models may fall into different levels depending on the type of analysis performed. Our models fall into Type 1b, described as 'development of a prediction model using the entire data set, but then using resampling (e.g. bootstrapping or cross-validation) techniques to evaluate the performance and optimisms of the developed model'. Therefore, in this section an overview is provided of statistical techniques that can be used to perform this type of development/evaluation.

3.2.1 Hyperparameter tuning and model selection

In order to avoid overfitting, the process of learning the optimal hyperparameters (i.e. hyperparameter tuning) is carried out on a validation data subset unseen by the training algorithm [71]. To assess the overall model performance, the final model is trained with the optimal hyperparameters and tested on a further unseen test data subset. The k-fold cross validation procedure (Figure 3.12a) is an efficient resampling method for repeatedly splitting the data into non-overlapping training and validation subsets. Thus, each of the k-folds is efficiently used in both the hyperparameter optimisation and model validation steps. The overall model performance can then be calculated as the average from all folds. However, even if not simultaneously, in the k-fold cross validation method the same data is used for both the hyperparameter optimisation and model selection steps, which may introduce bias in the latter and result in an overoptimistic model performance [97, 98, 34].

3.2.2 Nested k-fold cross-validation

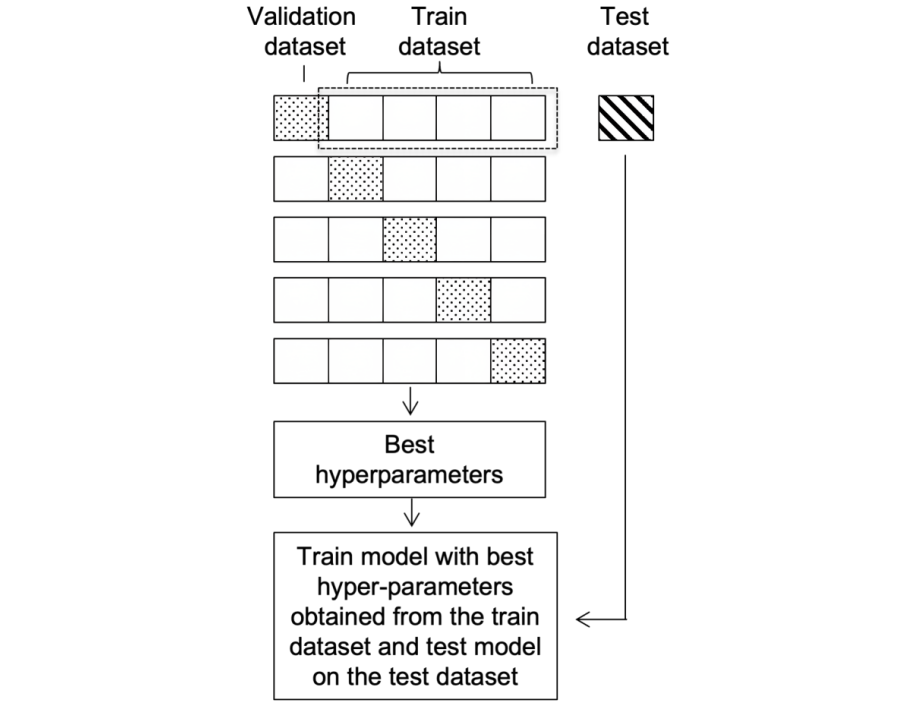
Nested cross validation [97] is an approach to hyperparameter optimisation and model selection that overcomes the problem of overoptimistic evaluation found in the regular k-fold cross-validation (CV) method. The nested CV uses a CV procedure inside the main CV (Figure 3.12b). In the outer CV, the data is randomly split into training and test sets following a k-fold CV approach, i.e. this split is repeated k times using k non-overlapping test sets. For each of the outer CV folds, hyperparameter optimisation is performed j times in an inner j-fold CV approach where the outer CV train dataset is further split into train and validation sets. Finally, for each of the outer CV folds, the entire training set is used for training using the optimised hyperparameters obtained from the inner CV and the prediction accuracy can be calculated on the held-out test set. In this way, the test set of each outer CV fold remains completely unseen, avoiding the bias introduced in traditional CV. A k-fold CV approach is *stratified* when the class balance is maintained in all CV folds.

3.2.3 Model discrimination performance metrics

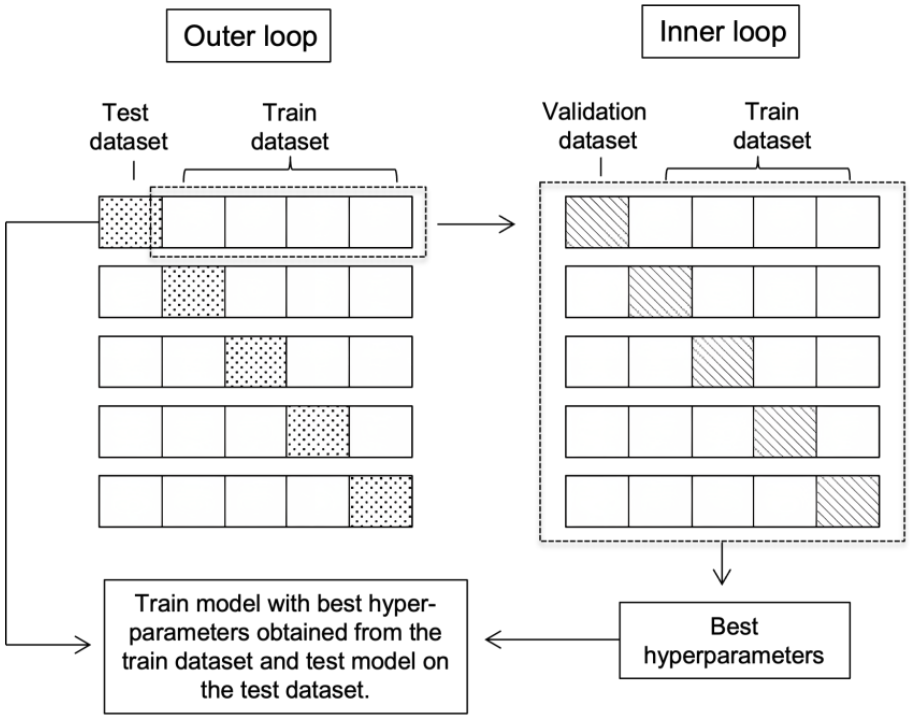
In the model selection process described in the previous section, the performance of the model is assessed based on a performance metric computed on the test data set and the best performing model is selected. Discrimination of a binary classification model is the ability of the model to correctly separate the subjects into the two classes considered. In this thesis, I have used a class-balanced cohort and assessed the discrimination ability of the models for the purpose of comparing different prediction models. Below is a description of the discrimination performance metrics used in this thesis.

A **Confusion Matrix** (Figure 3.13) is a tabular representation on the number of correctly and incorrectly predicted subjects for each class. Based on these numbers, the following metrics can be obtained:

- **Sensitivity, recall or true positive rate (TPR)** = TP/(TP+FN)
- **Specificity or true negative rate (TNR)** = TN/(TN+FP)
- **Precision or positive predictive value (PPV)** = TP/(TP+FP)
- **Negative predictive value (NPV)** = TN/(TN+FN)
- **Accuracy** = (TP+TN)/(TP+TN+FP+FN)



(a)



(b)

Fig. 3.12 (a) Standard 5-fold cross-validation vs. (b) nested 5-fold cross-validation workflows.

- **F1 score** = 2 x (precision x recall) / (precision + recall) = 2TP/(2TP+FP+FN)

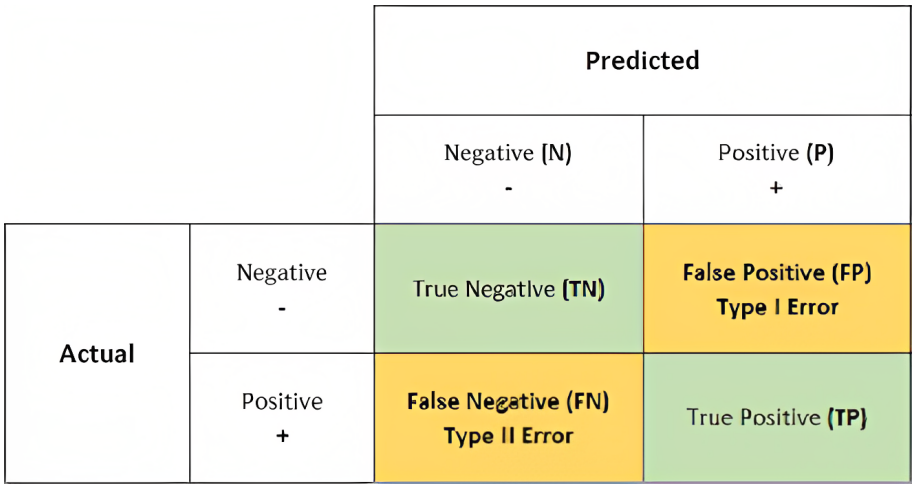


Fig. 3.13 Confusion matrix. Image source: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5.

It must be noted that while accuracy is a widely used discrimination metric in balanced datasets, it is not suitable for imbalanced datasets as it would over-represent the majority class. F1 score should be used instead in order to avoid misleading model performance conclusions [99]. As explained above, class balance has been maintained throughout this thesis. However, the effect of class imbalance on the robustness of model performance metrics should be taken into account when considering a dataset that reflects the actual prevalence of mandibular ORN, which is much lower than 50%.

The **Area Under the Receiver Operating Characteristics (AUROC) curve** is a plot (Figure 3.14) of the sensitivity against the false positive rate (FPR) (1-specificity) at various discrimination thresholds. This metric is not dependant on class balance.

3.3 Deep learning-based toxicity modelling

As discussed in Chapter 2, NTCP models have traditionally used DVH metrics as dosimetric variables. These metrics do not include clinically relevant spatial information of the dose distribution within the anatomical structures. Previous studies have shown that spatial information within a dose map is associated with radiation-induced toxicities in HNC [100] and that including such information into toxicity prediction models can result in improved prediction accuracies [34, 101–103]. Dosiomic features from radiation dose distribution maps can be handcrafted (e.g. extracted manually using conventional radiomics

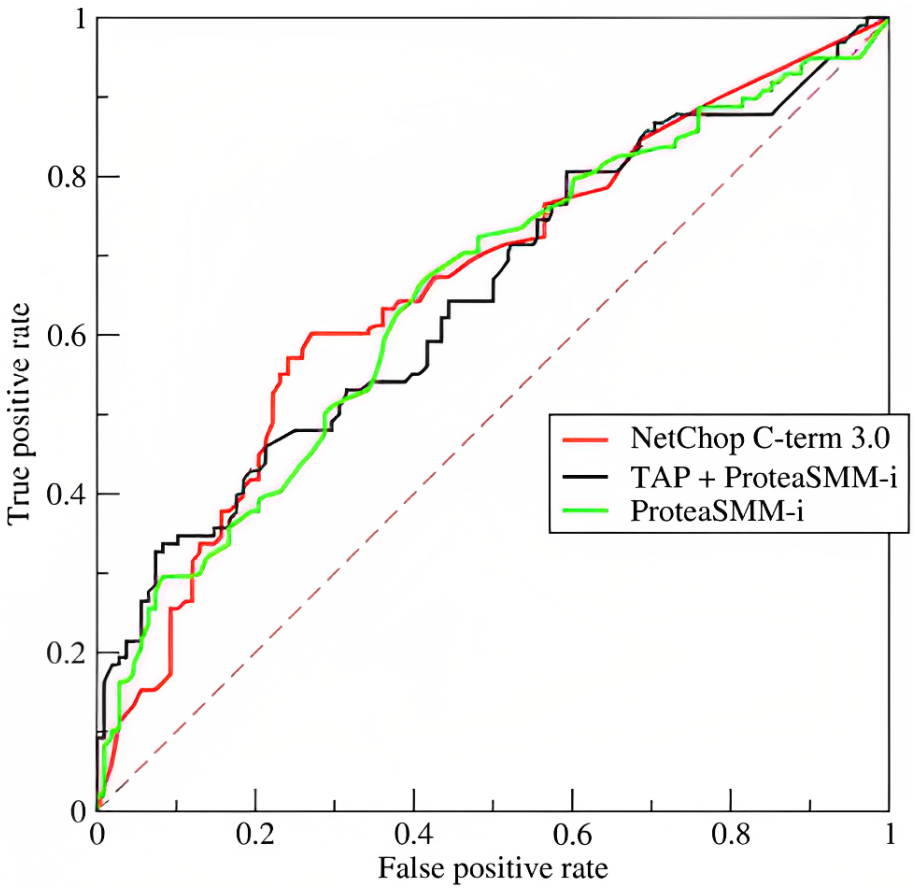


Fig. 3.14 Area Under the Receiver Operating Characteristics (AUROC) curve. Image source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic.

tools) or automatically learnt within a DL pipeline (Figure 3.15). These different possible approaches are discussed below.

3.3.1 Handcrafted dosiomic features

Several studies have used manual image feature extraction techniques to obtain features from dose distribution maps (i.e., dosiomics) in combination with ML classification methods. Dose-surface maps can be produced from tubular structures [104, 105]. However, manual extraction of spatial dose metrics becomes more complex with structures that cannot be 'unfolded' or flattened. Beasley et al. [106] applied image-based data mining to obtain a voxel-wise correlation map between radiation dose and radiation-induced trismus (difficulty in mouth opening) based on the Spearman's rank correlation coefficient between the toxicity outcome and the voxel value of the dose distribution to the mastication muscles. Dean et al. [107] included spatial dose information into their models for mucositis and dysphagia by describing the dose distribution with novel metrics based on the longitudinal and circumferential extents of the dose distributions to the oral mucosa and pharyngeal mucosa, respectively. Gabryś et al. [34] developed their own MATLAB-based software to handcraft features from the dose distribution volume such as spatial dose gradient, spread and skewness and combined these with demographic, DVH and radiomic features in NTCP models for xerostomia. Jiang et al. [108] applied ML methods directly to voxel dose values and other non-dosimetric features to predict xerostomia. Welch et al. [109] used the Python PyRadiomics package to extract statistical and shape features from the planned radiation dose distribution volumes in head and neck cancer.

3.3.2 Automated dosiomic features extraction

An alternative to these manual dosiomic feature extraction methods is the use of deep CNNs to automatically extract and use the most relevant features from the dose distribution volumes (Figure 3.13). A recent review by Appelt et al. [110] provides a thorough overview of the application of DL to RT outcome prediction and the resulting challenges of using radiation dose data. Some of the studies included in this review are mentioned in this section, however the interested reader is directed to the review article for additional examples. Zhen et al. [111] applied a pre-trained 2D VGG-16 to unfolded 2D rectum surface dose maps to predict rectum toxicity after cervical cancer radiotherapy. The 2D CNN performance (0.89 AUROC) was compared to a logistic regression (LR) model (0.70 AUROC) using DVH metrics and handcrafted features extracted from the rectum

surface dose maps. Ibragimov et al. [102, 112, 113] designed and successfully trained a 3D CNN to automatically identify patterns in the dose distributions to the portal veins of patients undergoing liver RT to predict 3+ acute and late hepatobiliary (HB) toxicities. They applied transfer learning from 3D CT images from a variety of human organs and data augmentation on the dose distribution maps to compensate for the small dataset size (125 subjects). Their predictions improved from 0.79 to 0.85 AUROC when the 3D CNNs for the dose maps were combined with a fully connected neural network (FcNN) for non-image data analysis. A study by Men et al. [103] with a total of 784 patients explored the combination of 3D dose maps with CT images and structure contours into a 3D residual CNN to predict radiation-induced xerostomia. Their CNN-based results (0.84 AUROC) were superior to a LR model that used DVH metrics and clinical variables (0.68 AUROC). Until my own work [114], which will be presented in Chapter 6, the use of DL for ORN prediction had not been investigated. A subsequent study [115] has been published and further discussion on the results of these two papers is included in Chapter 6.

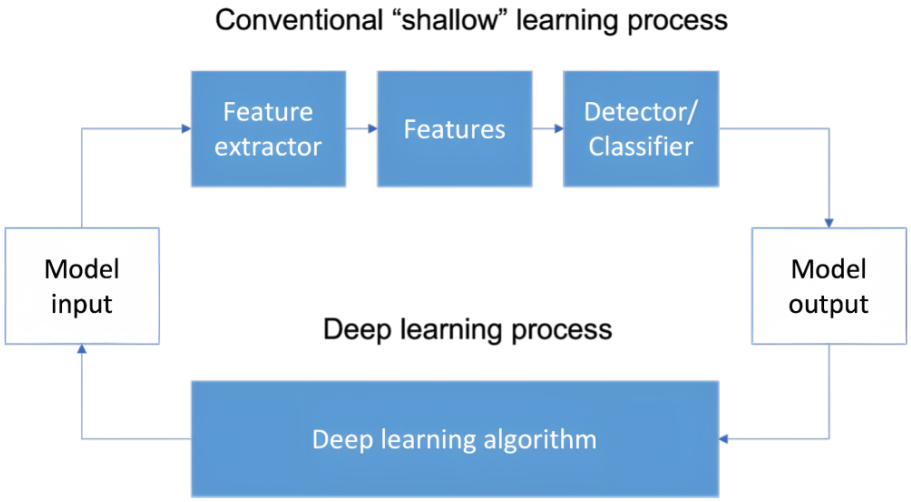


Fig. 3.15 Conventional vs. deep learning feature extraction and learning processes (adapted from El Naqa et al. [38]).

3.4 Discussion

The radiotherapy workflow consists of several complex and time-consuming steps, some of which have been described in Chapter 2. With AI, these processes have become more efficient and accurate [116]. While the clinical use of AI for automatic segmentation and treatment planning processes has been widely implemented and even commercialised [117], its use in the prediction of toxicity outcomes is still transitioning from the research

domain to the clinical setting [118, 119]. Mandibular ORN is a rare radiation-induced toxicity with naturally low case numbers and small datasets. Thus, the prediction of mandibular ORN using DL methods is particularly challenging. With the PREDMORN study (9), a larger dataset will be available as an expansion of the DL work presented in this thesis.

Radiation dose is a key risk factor for ORN but there are other clinical and demographic non-dosimetric risk factors. While Chapter 6 in this thesis describes the work carried out to develop a CNN-based method to predict mandibular ORN based on 3D radiation dose maps, Chapter 7 describes the work on developing a CNN-based prediction model that also includes non-image data (i.e., clinical and demographic variables) in combination with the radiation dose maps; the most relevant concepts of multimodality data fusion are also described. Chapter 8 explores the use of DL interpretability methods in the prediction of ORN, with detailed explanations of the most relevant concepts on DL interpretability.

Chapter 4

Materials

This Chapter describes the data used in this thesis and the different software employed for its processing. Section 4.1 describes the patient selection process that resulted in the three different cohorts used. Section 4.2 describes which data was used and how it was obtained. Section 4.3 describes how the image data was processed. Finally, Section 4.4 provides details on the data protection measures observed.

4.1 Patient selection

The patient selection criteria have evolved over the duration of my thesis and, as a result, three different cohorts have been used, which I have named Cohort 1, Cohort 2 and Cohort 3; the evolution into the three different cohorts is described below. Table 4.1 summarises each of the three cohorts as well as the experiments that these were used in.

4.1.1 Cohort description

Cohort 1. Originally, a total of 96 patients, 48 ORN cases and 48 controls, treated with radical IMRT between 2011 and 2015 were selected from the head and neck database maintained by the GSTT oncology team. The minimum follow-up time for the control group was 13.5 months. The median time from the end of RT to diagnosis of ORN was 11.8 months (IQR 20.8). Table 4.2 provides a summary of the demographic and clinical characteristics of this first cohort. Cohort 1 was used in a study [67] where I compared the performance of different ML methods in the prediction of ORN incidence using DVH metrics, clinical and demographic variables (Chapter 5).

Cohort 2. To increase the cohort size, additional ORN cases were included from the updated ORN list maintained by Dr Vinod Patel with patients treated between 2011 and 2019. Thus, the updated Cohort 2 consisted of 70 ORN cases and 70 controls. By the time the updated cohort was finalised, three control cases from the original cohort had developed ORN. Consequently, I decided to set a stricter minimum follow-up time of 3 years for the control group of the updated cohort. The median time from the end of RT to diagnosis of ORN was 12.5 months (IQR 21.6). Table 4.3 provides a summary of the demographic and clinical characteristics for Cohort 2. Primary tumour groups in Cohort 1 were updated to more generic ones in Cohort 2. This cohort was used in a study [114], where DL methods were used to predict ORN incidence based on 3D radiation map distributions and their performance compared to predictions made with a RF model based on DVH metrics.

Cohort 3. In parallel to this thesis, I have developed the PREDMORN multi-centre study (Chapter 9) in order to build robust ORN prediction models with the largest and most diverse cohort ever used in published studies. The PREDMORN study design and protocol was developed with contributions from all participating centres, with a control-case matching based on primary tumour site and treatment year and well-defined inclusion/exclusion criteria (Table 9.1 in Chapter 9). As a result, Cohort 2 was updated to Cohort 3 (Table 4.4) to match the PREDMORN study requirements. For instance, it was agreed that no minimum follow-up time threshold would be applied for the control group. Thus, the average follow-up time for the controls in Cohort 3 was 49.9 months (range 5.2-92.0) while the median time from the end of RT to diagnosis of ORN was 12.1 months (IQR 20.3). In the process, additional ORN cases were diagnosed during this time with a total of 92 cases treated between 2011 and 2022. The entire ORN group was reviewed and mislabelled ORN cases and cases that had ORN in the maxilla instead of the mandible were identified and excluded. During the time span considered, from a total of 1721 HNC patients radically treated, a total 142 patients (8.3%) were diagnosed with ORN, 50 of which were excluded because of unavailable RT dose and/or RT plan Digital Imaging and Communications in Medicine (DICOM) files (18), ORN region outside of the mandible (15), palliative or low prescribed dose (8), previous irradiation in the HN region (6) or two primary tumour sites (3). With regards to the updated control cohort, a well-defined selection process was followed as per the PREDMORN study. First, the primary site distribution was obtained from the ORN cohort. Primary tumour site groups considered included oral cavity, oropharynx, paranasal sinus/nasopharynx, larynx/hypopharynx, salivary glands and unknown primary (neck). All of the experiments in this thesis were performed used a class-balanced cohort (i.e., a 1:1 control-case ratio) to facilitate the prediction task in the DL methods. For the multi-institutional modelling

study, however, a 2:1 control-case ratio was agreed with all the centres. In both cases, the random control-case match based on treatment year was done for each primary tumour site. The ML and DL experiments published on Cohort 2 were repeated on Cohort 3 and the results for both Cohorts are discussed in Chapter 6.

4.1.2 Dosimetric comparison of Cohort 2 and Cohort 3

This subsection aims to analyse the dosimetric differences between Cohorts 2 and 3, which will have an effect on the results of the corresponding ML and DL experiments. As shown in Figure 4.1, there is a degree of separation in the curves for the median DVHs of the two groups, ORN and control, in both cohorts; whilst the largest separation in Cohort 2 (dotted line) occurs at high doses, for Cohort 3 (continuous line) they are clearly more separated in the intermediate dose region.

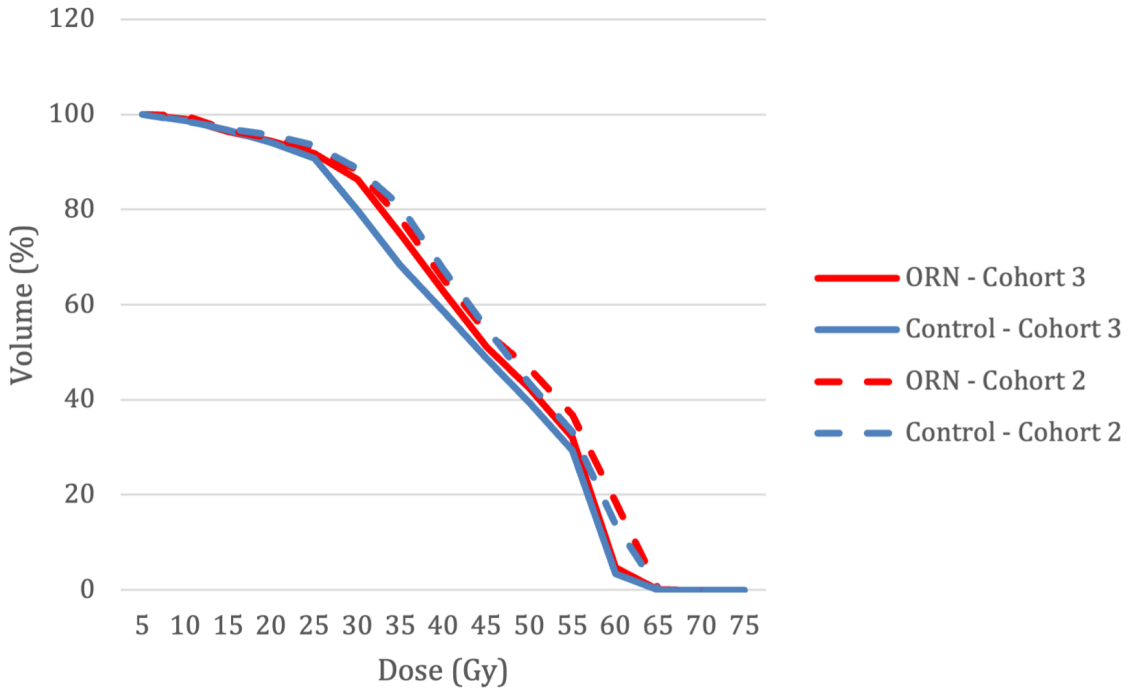


Fig. 4.1 Median DVH comparison between ORN and control groups for Cohort 2 and Cohort 3.

The Mann-Whitney U test [120] (MWU) is a 'distribution-free' alternative to the two-sample parametric t-test for comparing data from two independent groups as it doesn't assume normality. This test was used for all dosimetric variables even if some of them had a normal distribution. Whilst the p-value from the MWU test can inform whether there is an effect for a given variable, the Cohen's d [121] is a measure of the effect size for that

Table 4.1 Summary of the three cohorts used, the main changes between them and the related experiments they have been used in.

	ORN/Controls	Time range	Updates	Related experiments
Cohort 1	48/48	2011-2015	Original cohort	Predicting mandibular ORN from non-imaging data using ML (Chapter 5)
Cohort 2	70/70	2011-2019	a) Minimum follow-up time for control group extended from 13 months to 3 years; b) ORN cases reviewed	Comparison of DVH-based predictions using ML methods and dose map-based predictions using DL methods (Chapter 6)
Cohort 3	92/92	2011-2022	a) Control-case matching based on primary tumour-site and treatment year; b) PREDMORN inclusion/exclusion criteria; c) No minimum follow-up for control group; d) 'Current' alcohol and smoking status defined as that within 2 months prior to the start of RT	a) Comparison of DVH-based predictions using ML methods and dose map-based predictions using DL methods (Chapter 6); b) Combining image and tabular data (Chapter 7); c) Interpretable CNN methods (Chapter 8)

Table 4.2 Demographic and clinical variables characteristics of the ORN and control groups in **Cohort 1**.

	ORN	**Control**
Gender		
Male	34(71%)	38(79%)
Female	14(29%)	10(21%)
Age (median, (IQR))	64 (14)	59 (15)
Smoking		
Current	25(52%)	19(40%)
Previous	14(29%)	19(40%)
Alcohol		
Current	33(69%)	33(69%)
Previous	5(10%)	4(8%)
Chemotherapy	33(69%)	33(69%)
Pre-RT dental extractions	30(63%)	31(65%)
Pre-RT surgery		
Primary RT	38(79%)	30(63%)
PORT	10(21%)	18(38%)
Primary tumour site		
Oropharynx	28(58%)	28(58%)
Oral cavity	13(27%)	9(19%)
Larynx	3(6%)	7(15%)
Hypopharynx	0(0%)	2(4%)
Paranasal sinus	1(2%)	1(2%)
Unknown primary	3(6%)	1(2%)

Table 4.3 Demographic and clinical variables characteristics of the ORN and control groups in **Cohort 2**.

	ORN	**Control**
Gender		
Male	49(70%)	56(80%)
Female	21(30%)	20(20%)
Age (median, (IQR))	61(13)	61(15)
Smoking		
Never	20(29%)	17(24%)
Current	31(44%)	25(36%)
Previous	19(27%)	28(40%)
Alcohol		
Never	19(27%)	23(33%)
Current	44(63%)	43(61%)
Previous	7(10%)	4(6%)
Chemotherapy	46(66%)	42(60%)
Pre-RT dental extractions	45(64%)	47(67%)
Pre-RT surgery		
Primary RT	47(67%)	41(59%)
PORT	23(33%)	29(41%)
Primary tumour site		
Oropharynx	42(60%)	31(44%)
Oral cavity	21(30%)	16(23%)
Larynx	2(3%)	11(16%)
Hypopharynx	0(0%)	3(4%)
Salivary glands	1(1%)	4(6%)
Nasopharynx	0(0%)	0(0%)
Paranasal sinus	1(1%)	0(0%)
Unknown primary	3(4%)	3(4%)

Table 4.4 Demographic and clinical variables characteristics of the ORN and control groups in **Cohort 3**.

	ORN	**Control**
Gender		
Male	66(72%)	72(78%)
Female	26(28%)	20(22%)
Age (median, (IQR))	62 (13)	61 (15)
Smoking		
Never	19(21%)	25(27%)
Current	26(28%)	46(50%)
Previous	47(51%)	21(23%)
Alcohol		
Never	13(14%)	15(16%)
Current	8(9%)	14(15%)
Previous	71(77%)	63(69%)
Chemotherapy		
None	33(36%)	35(38%)
Cisplatin	50(54%)	52(57%)
Carboplatin	7(8%)	1(1%)
Cetuximab	2(2%)	4(4%)
Dental assessment pre-RT	89(97%)	80(87%)
Pre-RT dental extractions	55(60%)	50(54%)
Pre-RT surgery		
Primary RT	57(62%)	57(62%)
PORT	35(38%)	35(38%)
RT technique		
IMRT	53(58%)	53(58%)
VMAT	39(42%)	39(42%)
Primary tumour site		
Oropharynx	52(57%)	52(57%)
Oral cavity	28(30%)	28(30%)
Larynx / Hypopharynx	3(3%)	3(3%)
Salivary glands	3(3%)	3(3%)
Paranasal sinus/Nasopharynx	2(2%)	2(2%)
Unknown primary	4(4%)	4(4%)
Prescribed dose (Gy, fractions)		
71.5 in 30	2(2%)	0(0%)
70 in 35	0(0%)	1(1%)
67.2 in 28	1(1%)	0(0%)
66 in 33	10(11%)	4(4%)
65 in 30	54(59%)	59(64%)
60 in 30	20(22%)	26(28%)
55 in 20	1(1%)	0(0%)
50 in 20	4(4%)	2(2%)

variable. It measures the difference between two group means in terms of the number of standard deviations that the means differ (Equation 4.1). A minimum effect size of 0.80 is generally accepted as large [122] while sizes around 0.50 and 0.20 are considered medium and small, respectively.

$$Cohen's\ d = (Mean_{ORN} - Mean_{control})/SD_{pooled} \tag{4.1}$$

$$\text{where } SD_{pooled} = \sqrt{((SD^2_{ORN} + SD^2_{control})/2)}$$

Table 4.5 provides the results from a univariate analysis and the effect size for a set of DVH metrics for the two cohorts. From the DVH metrics considered, only the ones relating to maximum doses (Dmax and D2%) showed a medium effect size in Cohort 2. In Cohort 3, only the mean (p=0.031) and median (p=0.028) dose metrics showed some difference between the two groups; however, their corresponding effect size was small.

Table 4.5 Univariate analysis results and effect size of the DVH metrics for Cohorts 2 and 3. Values under the 0.5 significance level are in bold.

Cohort 2				
DVH metric (median (IQR))	ORN	Control	p-value (MWU)	Effect size (Cohen's d)
Dmax	68.5 (3.3)	67.6 (6.2)	**0.002**	0.441
D2%	65.2 (4.3)	64.3 (6.5)	**0.000**	0.540
Dmin	9.2 (6.8)	7.7 (6.8)	0.131	0.238
D98%	13.3 (12.2)	11.6 (9.1)	0.151	0.212
Dmean	46.6 (8.0)	46.5 (12.6)	0.159	0.270
D50%	49.9 (11.7)	48.6 (16.6)	0.078	0.294
Cohort 3				
DVH metric (median (IQR))	ORN	Control	p-value (MWU)	Effect size (Cohen's d)
Dmax	68.2 (4.7)	68.2 (6.1)	0.078	0.191
D2%	65.1 (5.3)	64.7 (5.8)	**0.035**	0.227
Dmin	8.1 (7.2)	7.3 (6.9)	0.490	0.059
D98%	12.1 (10.9)	12.2 (8.5)	0.564	0.051
Dmean	46.8 (9.3)	44.0 (10.7)	**0.031**	0.277
D50%	50.0 (14.0)	47.2 (16.3)	**0.028**	0.256

4.2 Data

The Head and Neck cnacer research database is held within the Guy's Cancer Cohort (REC reference 18/NW/0297 and IRAS Project ID 231443), which was reviewed by the

North-West Haydock REC Committee. For the work included in this thesis, data was retrospectively collected from this database under the Project Number 6333. Ethical clearance approval was granted by the Guy's Cancer Cohort Access Committee and Steering Committee on the 21st of April 2016. Figure 4.2 lists the data collected within each data type; a more detailed description is provided in Chapter 9 as part of the protocol for the PREDMORN multi-institutional study. Figure 4.3 summarises which data was used in each experiment performed in this thesis.

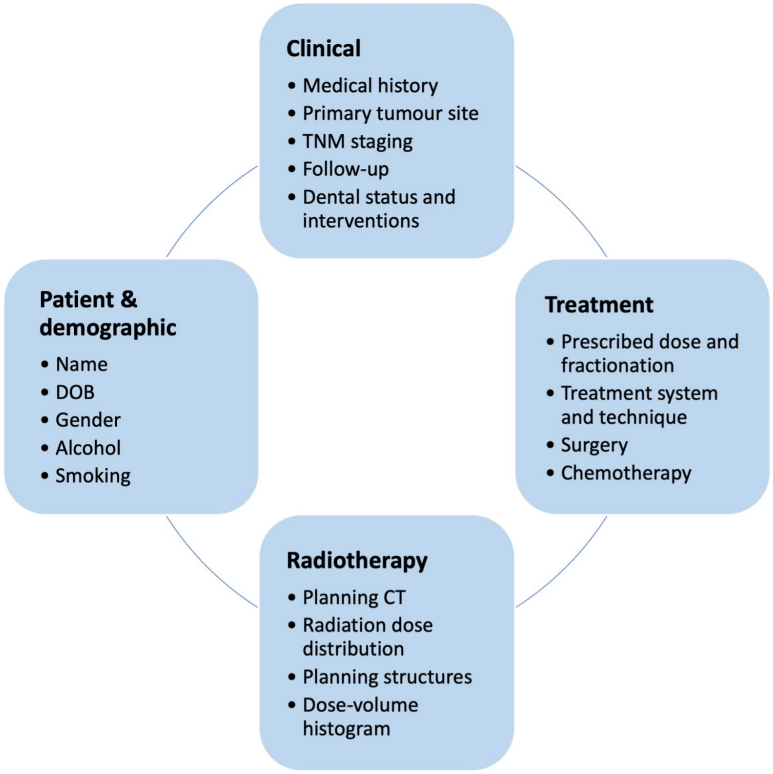


Fig. 4.2 Types of data with a list of items collected for each data type.

4.2.1 Patient, clinical and treatment data

When a patient is diagnosed with HNC, a well-established clinical protocol is followed at GSTT: 1) the patient status, medical history and disease stage are assessed, 2) a decision is made on the treatment type and schedule, 3) the treatment is delivered and 4) the patient is followed up over treatment and for up to five years post-treatment by oncology and surgical teams in order to assess treatment outcomes. Toxicity scoring using the NCI CTCAE v. 4.0/5.0 grading systems [49] is recorded at baseline, weekly during RT and at 6 weeks and 3, 6 and 12 months post-treatment and yearly thereafter prospectively at the point of care.

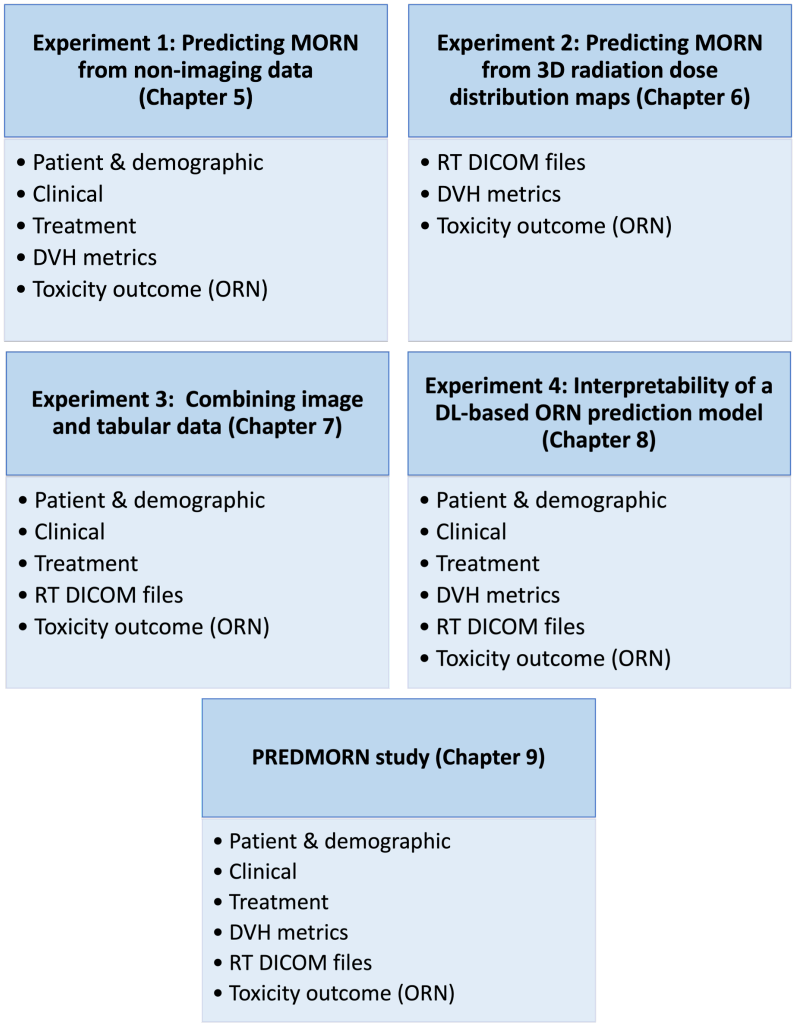


Fig. 4.3 Data used in each experiment and chapter.

Any clinically relevant information obtained at any of these steps is recorded using the Mosaiq (Elekta AB, Stockholm, Sweden) and the GSTT Electronic Patient Record (EPR) systems. In addition to this, and mostly for research purposes, the head and neck unit at GSTT maintains a database that contains patient, demographic and treatment data for all HNC patients treated since 2011.

4.2.2 Radiotherapy treatment planning data

During the radiotherapy treatment planning process patient-specific information is used and produced: planning CT images, OAR and target volume delineations, radiation dose distribution maps and DVH. All this information can be exported from the TPS as DICOM files which, in addition to images, may also contain sets of metadata such as patient information, image acquisition details, etc. Inverse-planned IMRT was introduced at GSTT in March 2011. Prior to this, the forward-planned IMRT and 3DCRT techniques were used for radical and palliative cases, respectively. The resulting differences in radiation dose distributions derived from both techniques (IMRT vs. 3DCRT) have been discussed in Chapter 2 (Section 2.2.1). Only patients treated with inverse-planned IMRT were included in the study in order to obtain a homogeneous cohort with respect to the treatment planning technique. Radical primary RT cases are prescribed a total dose of 65-70 Gy in 30-35 fractions and 55 Gy in 20 fractions in selected cases. Radical PORT cases are prescribed 60-66 Gy in 30-33 fractions and 50 Gy in 20 fractions in selected cases.

Within the time frame considered in this thesis (2011 to 2022), two different radiotherapy TPS were used to produce the clinical radiation dose treatment plans: the Monaco (Elekta AB, Stockholm, Sweden) TPS was used between 2011 and 2016 and the Eclipse (Varian Medical Systems, Milpitas, CA) TPS was used from 2016 onwards. In some of the patients planned with Monaco the absorbed dose had been calculated as $D_{w,m}$. I re-calculated the dose distribution for these patients as $D_{m,m}$ in order to maintain the same dose reporting method across the whole cohort. The two absorbed dose calculation methods, $D_{w,m}$ and $D_{m,m}$, are described in Chapter 2 (Section 2.2.3).

4.2.3 Dose-volume histogram (DVH)

Some of the experiments that were performed used DVH-based dosimetric data as input variables (Chapters 5 and 6). To obtain these I extracted the raw cumulative DVH for the mandible structure from the treatment TPS. I then used the 'DVHmetrics' package in the R software (R Foundation for Statistical Computing, Vienna, Austria) to obtain the DVH

metrics. Finally, I applied an EQD2 correction (Equation 2.1 in Chapter 2) for patients with a fraction dose different from 2 Gy.

4.2.4 ORN data

Patients who develop ORN after their RT course at GSTT are treated and closely monitored by a specialist oral surgical team in a dedicated clinic. The Notani [50] ORN grading system is used at GSTT (Section 2.4.1 in Chapter 2); however, for the purpose of binary classification in the experiments performed in this thesis, toxicity outcomes were dichotomised and any grade of ORN was considered as an event.

For ORN cases, the ORN region in the mandible was contoured on the RT planning CT images by an ORN expert oral surgeon (Dr Vinod Patel). This was done based on planar x-ray dental images, cone beam CT where available and dental follow-up clinical notes using cognitive transfer, i.e. obtaining the shape and localisation information of the ORN area from the dental images and manually contouring this region of the mandible on the RT planning CT accordingly. All contouring was done in the TPS used for the clinical treatment plan (Monaco or Eclipse). The RT Structure DICOM files for the ORN structure were exported from the TPS and processed in the same way as the mandible structure files as described in Section 4.3 below.

4.3 Image data processing

Several data processing steps were required for the image data to produce the mandible dose maps. Figure 4.3 shows a schematic of the data processing workflow, and the subsections below describe these steps in more detail.

4.3.1 Mandible segmentation

I manually segmented the mandible for all the patients in the cohort using the TPS contouring tools on the planning CT images. A subset of the manual segmentations were checked by Dr Teresa Guerrero Urbano and Dr Vinod Patel as a data curation quality control measure. The mandible contours included the whole mandible with mandible sockets and excluded the maxilla and the teeth [123, 41]. Patients (mostly oral cavity cases) who had undergone a mandibulectomy with flap reconstruction prior to RT were particularly challenging and I often required the support of clinical notes to visually identify

the reconstructed part in order to exclude it from the mandible contour. Figure 4.4 shows an example of the manual segmentation of the mandible in a case with a flap reconstruction after a bilateral mandibulectomy.

I exported the mandible segmentation from the TPS as an RT DICOM Structure file and converted it to a label map or binary mask (saved as a NIfTI file) using the *Segmentations* module with *Labelmap* as the output in 3D Slicer 5.0.2 [124]. The CT DICOM files were required to produce the mandible masks (Figure 4.5).

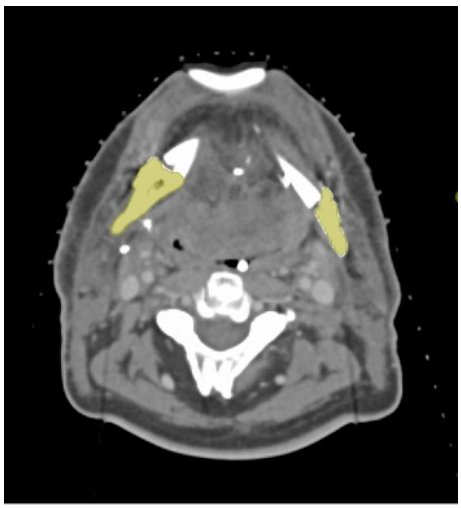


Fig. 4.4 CT slice from a patient who underwent flap reconstruction after a bilateral mandibulectomy. The yellow contours correspond to the mandible bone.

4.3.2 Image resampling

The CT images acquired for treatment planning purposes are required in the structure segmentation and dose calculation steps and can therefore be used as the geometrical reference system for both the RT Structure and RT Dose DICOM files. I resampled all the CT volumes (as NIfTI files) to a common slice thickness of 2 mm and slice size of 512 pixels x 512 pixels using the *Resample Scalar Volume* module in 3D Slicer using linear interpolation. I then resampled all the previously created mandible masks and the exported RT Dose DICOM files to the resliced CT using the *Resample Image (Brains)* module in 3D Slicer (again with linear interpolation).

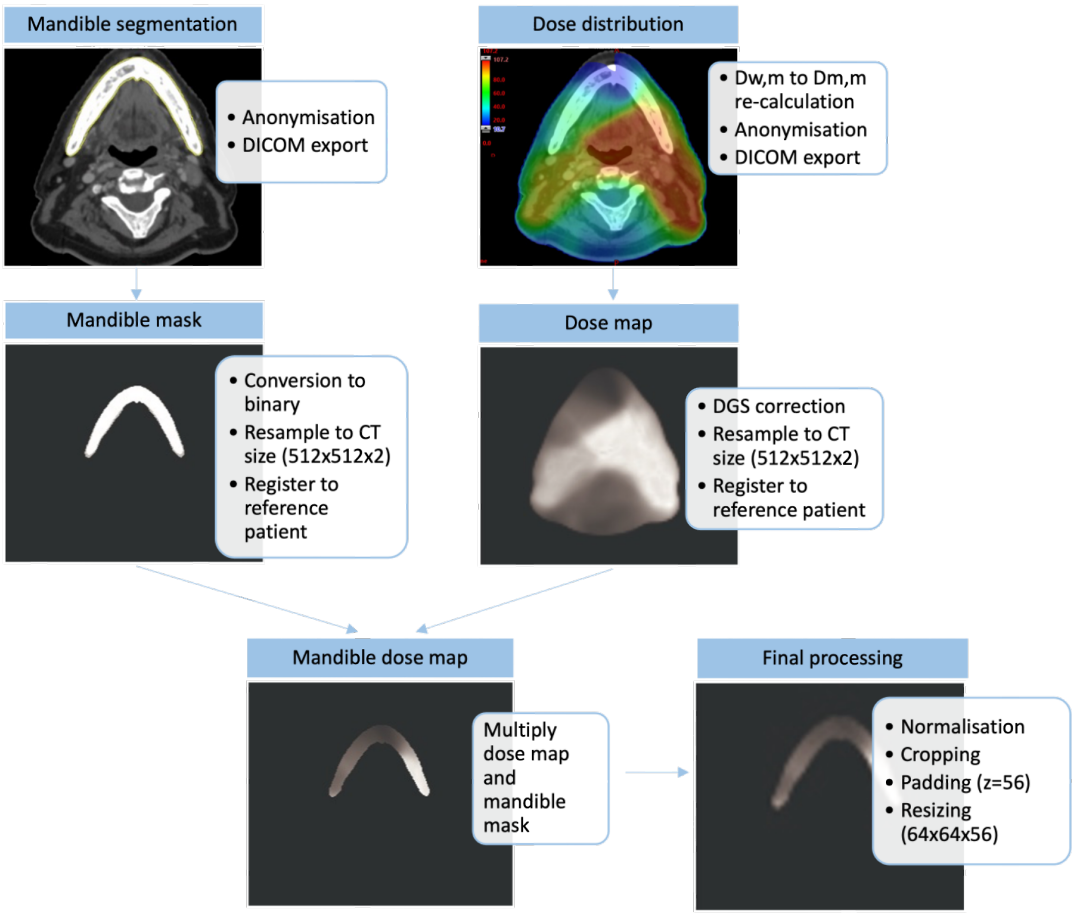


Fig. 4.5 Schematic of the image data processing workflow. Treatment data was anonymised and exported from the TPS. After the image registration and resampling steps, the mandible dose map was obtained by multiplying the masked mandible structure and the clinical dose distribution. Final processing steps were required before using it as input into the DL network.

4.3.3 Registration to a common reference space

Head and neck cancer patients are immobilised with a thermoplastic shell. At GSTT we originally used a 3-point head and neck shell with bear claws and then transitioned to a 5-point immobilisation shell. Patients are scanned with a comfortable neutral neck position; but as this may be different from patient to patient and there are anatomical differences there is unavoidable variation between patients that may result in large anterior-posterior rotations of the neck. Inspired by the methods described by Ibragimov et al. [102], I rigidly registered all mandible segmentations to a common reference space using ITK-SNAP [125] in order to minimise DL network to focus on dosimetric variations instead. I used the ITK-SNAP 6-degree (translation and rotation) rigid registration model with the Mutual Information image similarity metric. I selected the patient with the largest number of mandible slices (a total of 56) as the reference patient and added empty slices to smaller mandible volumes at a later processing step. I transformed the dose maps using the same rigid transformations to maintain alignment.

4.3.4 Mandible dose maps

To obtain the 3D mandible dose distribution maps I multiplied the whole dose distribution maps by the binary mandible segmentation masks. I then normalised the 3D mandible dose distribution maps to the voxel value range of the entire dataset and cropped the resulting volumes to reduce the empty voxels. I normalized the 3D mandible dose distribution maps as follows: $I_{normalised} = (I - D_{min})/(D_{max} - D_{min})$, where D_{min} and D_{max} are the global minimum and maximum intensities across the entire dataset.

Finally, I resized the final volumes to a size of 64 pixels x 64 pixels x 56 slices, where the number of slices was determined by the largest mandible in the cohort and any smaller mandible volumes were padded with empty slices.

4.4 Data protection and anonymisation

Strict data protection and anonymisation measures have been maintained. I only accessed patient, demographic, clinical and treatment data from a fully protected and approved Trust laptop and owned software. This data was then copied across into a password-protected Excel spreadsheet. In an anonymised copy of this spreadsheet the patient names, hospital

IDs and date of birth (DOB) were substituted by a study ID and the 'age at the start of RT' variable.

A study copy of all RT data was created. The anonymisation process of the RT data was different for the Eclipse and Monaco systems; while Eclipse allows for built-in anonymisation of all the RT data during the export step, an external software (DICOM Adjuster) was needed to anonymise the Monaco RT DICOM files.

4.5 Discussion

The experiments included in this thesis have used data that was obtained retrospectively. As a result of the learning curve during my PhD, the patient selection process has evolved, and the cohort has consequently changed. Thus, the different experiments described in Chapters 5, 6, 7, 8 and 9, have used slightly different cohorts. Chapter 6 in particular includes a comparison between cohorts 2 and 3 with regards to the resulting performance of the DL-based models.

One of the main differences between Cohorts 2 and 3 is the lack of minimum follow up time requirement in the control group for the latter. Excluding cases with a short follow-up time is common in previously published ORN studies [126, 66, 60] on the basis that ORN is a late toxicity and false negatives could be included in the cohort otherwise. Again, Chapter 6 explores the consequences of this difference with regards to the effect size of the cohort.

Chapter 5

Predicting MORN from non-imaging data using ML

Prediction of ORN incidence in the treatment of HNC may lead to risk-reduction measures (e.g., reduced mandibular radiation dose near extraction sites when possible) and/or a more dedicated follow-up for early detection and intervention of ORN. Chapter 2 describes the concept of NTCP modelling (Section 2.3.2) and provides a review of existing studies that have investigated the correlation between dosimetric, clinical and demographic factors and ORN incidence (Sections 2.4.2 and 2.4.3).

More complex supervised machine learning (ML) methods have been used to develop NTCP models for clinical decision support [31] in HNC RT [127, 128, 107, 102, 129]. This Chapter describes a comparison between different supervised classification machine learning methods (described in Chapter 3) for the prediction of ORN incidence. This study was published in the British Journal of Radiology in 2021 [67].

5.1 Data

This study was performed using Cohort 1 described in Chapter 4 (Table 4.1). Prescribed radiation doses ranged between 50 Gy in 20 fractions and 71.5 Gy in 30 fractions. Maximum (D_{max}) and mean (D_{mean}) mandible doses and relative dose–volume levels in the range V40 Gy to V70 Gy in 5 Gy increments were considered as the DVH-based dosimetric variables. All doses were converted to EQD2 as per Equation 2.1 in Chapter 2, using an α/β ratio of 3 to account for late toxicity of the mandible. The distribution of the age and DVH-based dosimetric variables is illustrated by the boxplots in Figure 5.1.

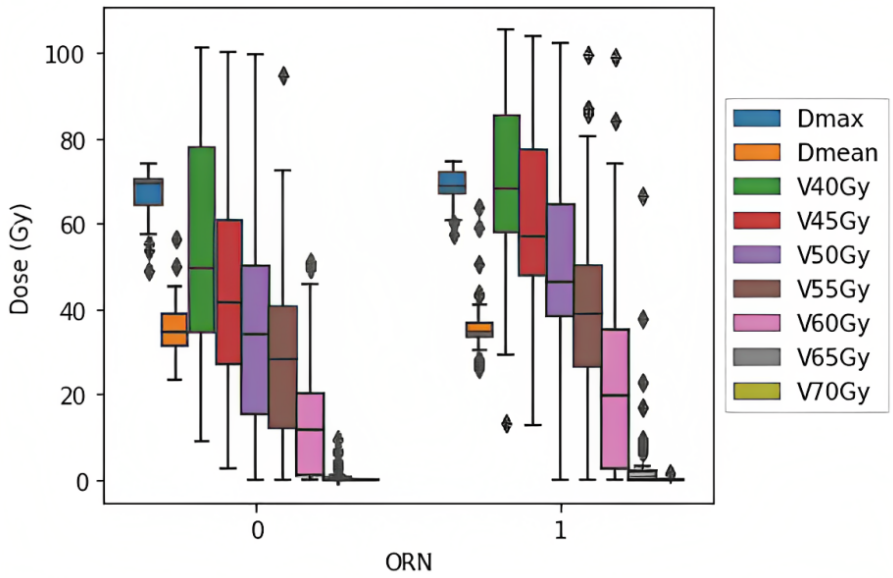


Fig. 5.1 Boxplots showing the distribution of the DVH-based variables used in this Chapter, where 0 and 1 in the x-axis correspond to the controls and cases, respectively.

5.2 Variable selection

As described in Chapter 2 (Table 2.4), there are several variables that are potentially associated with mandibular ORN and multifactorial models are a way of describing this association and/or predicting the toxicity outcome. Variable selection is commonly used as a statistical method to reduce the complexity of multifactorial models and to prevent over-fitting. Univariate analysis was applied as a variable selection method. The non-parametric Pearson's chi-squared test was used to compare the observed frequency distributions of the ORN and control groups for the categorical variables (e.g. smoking status, gender, dental extractions). Using a significance level of 0.3, dental extractions post-RT (p = 0.26) and surgery pre-RT (p = 0.13) were included as clinical variables. The discriminatory power of the continuous variables (i.e. age and DVH-based dosimetric variables) was assessed with the one-sided MWU test using the `scipy.stats.mannwhitneyu` module with a 'greater' alternative hypothesis (i.e. the median in the ORN group is greater than the median in the control group for all the tested variables). Using a significance level of 0.05, D_{max} (p = 0.045) was found to be the only dosimetric variables with class discriminating power (Table 5.1).

Table 5.1 Mann-Whitney U test results on the age and dosimetric variables for Cohort 1.

Metric	p-value
Dmax	0.045
Dmean	0.092
Age	0.115
V40	0.342
V45	0.225
V50	0.309
V55	0.223
V60	0.262
V65	0.395
V70	0.399

5.3 Model design and training

The predictive accuracy of five supervised ML methods (described in Chapter 3 section 3.1.3) - logistic regression (LR), support vector machine (SVM), random forest (RF), adaptive boosting (AdaBoost) and artificial neural network (ANN) - was tested using the SciKit-Learn (sklearn) package in Python 2.7.15.19. Stratified 5-fold nested cross-validation (described in Section 3.3.1 in Chapter 3) was used, with embedded model hyperparameter optimisation (Section 3.2.1 in Chapter 3) performed using the `model_Selection.GridSearchCV` module. For all models, at each of the 5 outer cross-validation folds, 20 out of the total 96 cases were kept unseen by the model for testing its performance. The remaining 76 cases were used for model training in the LR, SVM, RF and AdaBoost models. For the ANN model, the training data set was further split during training into 60 cases for model training and 16 for model validation (80:20). The average accuracy across all inner CV folds was calculated for each hyperparameter combination and the results reported are for the outer CV folds using the hyperparameter combination that obtained the highest average accuracy.

For the multivariate LR model, the module `linear_model.LogisticRegression` was implemented with a C parameter of 0.001 and the *l2* regularisation penalty. The SVM classifier was implemented using the Scikit-learn *SVC* class with an RBF kernel, a penalty parameter (C) of 100 and a γ parameter of 0.001. The RF classifier was implemented using the sklearn.ensemble.RandomForestClassifier module with a maximum number of estimators (i.e. number of built trees) of 10, a maximum tree depth of 50, a minimum number of samples at a leaf node of 1 and a minimum number of samples required to split an internal node of 0.5. The AdaBoost classifier was implemented using the

`ensemble.AdaBoostClassifier` module with a learning rate of 0.0001 and a maximum number of estimators of 10.

The ANN was implemented in Keras with Tensorflow as the backend and trained on a Nvidia Titan Xp GPU. It consisted of an input layer with the number of input nodes equal to the number of variables used, followed by a 200-node hidden dense (fully connected) layer with the ReLU activation function and a 1-node output layer with the sigmoid activation function for binary classification (ORN or not ORN). A dropout layer was added at the end of the network pipeline to reduce overfitting. The Binary Cross-Entropy loss function was used to train the ANN and the Adam optimiser was used to minimise the loss function. Hyperparameter optimisation was performed using a grid search strategy. Based on the grid search results, the model was trained for 2000 epochs with a batch size of 30, a dropout rate of 0.0 and a learning rate of 0.001. The best model was chosen based on the highest accuracy achieved with the validation data set during training.

5.4 Model performance

Model performance was assessed using the measures described in Section 3.2.2 in Chapter 3, and the results of these metrics are summarised in Table 5.2. Although no single model outperformed the rest in all measures considered, the ANN model (71%) had the highest overall prediction accuracy on the unseen test dataset, closely followed by the LR (70%), SVM (69%), AdaBoost (68%) and RF (66%) models. The performance of the models was generally enhanced when using only the most statistically significant variables as per the variable selection process.

The McNemar's statistical hypothesis test [120] was used to determine whether there were statistically significant differences in classifier model performance. A total of 10 pair-wise comparisons were thus performed in this study. Bonferroni correction [120] for multiple comparisons was applied, resulting in a corrected significance level of 0.05/10 = 0.005 for each comparison. Table 5.3 provides the results from the McNemar's test on all combinations of the models explored. Based on the corrected significance level, no statistically significant difference was observed between models.

Table 5.2 Model performance summary

Model (variables)	Accuracy	Sensitivity (TPR)	Specificity (TNR)	Precision (PPV)	NPV
LR (all/selected)	0.66/0.70	0.75/0.77	0.52/0.55	0.62/0.65	0.77/0.76
SVM (all/selected)	0.64/0.69	0.75/0.78	0.59/0.56	0.65/0.66	0.72/0.77
RF (all/selected)	0.63/0.66	0.64/0.65	0.59/0.65	0.63/0.65	0.65/0.65
AdaBoost (all/selected)	0.65/0.68	0.63/0.66	0.57/0.66	0.72/0.67	0.65/0.66
ANN (all/selected)	0.65/0.71	0.75/0.78	0.62/0.67	0.65/0.68	0.72/0.78

Table 5.3 Results from the McNemar's statistical test on all model pair combinations.

χ^2 p-value	LR	SVM	RF	AdaBoost
SVM vs.	0.628			
RF vs.	0.396	0.256		
AdaBoost vs.	0.984	0.995	0.382	
ANN vs.	0.658	1.000	0.211	0.825

5.5 Discussion

Most ORN-related published work has focused on finding correlations between ORN incidence and clinical and dosimetric variables based on population studies. While it is important to understand these associations, the ability to predict incidence on a case-by-case basis would be a more valuable clinical application. This study has compared different ML models for the task of predicting mandible ORN incidence, which is in essence a binary classification task, based on clinical, demographic and DVH-based dosimetric variables. The results presented show that ML-based methods can be used to assist clinical decision-making for HNC patients undergoing RT. We cannot recommend a specific model based on our prediction performance results, as these were not found to be statistically significantly different. It could be argued that the use of the Bonferroni correction when performing more than five model comparisons could lead to 'highly conservative' conclusions [120]. However, the McNemar test results showed no significant difference between models even with an uncorrected significance level of p=0.05. The ANN model showed the highest overall prediction accuracy. It could be argued that this is also the model with poorest interpretability and that a simpler and more transparent model (e.g. LR) might be preferred for clinical use if the prediction performance is not significantly compromised. The advantage of using ANNs over traditional ML methods is perhaps more obvious when using more complex input data, e.g. high-dimensional data such as images. Chapter 6 describes how the use of deep CNNs was explored in the task of predicting ORN incidence based on 3D images of the radiation dose distribution to the mandible.

As described in Chapter 3 (Section 3.3.1), other studies have previously exploited spatial dose information as input into a ML algorithm, usually together with clinical and demographic variables. The use of deep CNNs, although widely employed in medical imaging applications, has yet to be fully explored for toxicity prediction. A CNN is able to learn the image-based features that are most useful for the task being trained for. Reber et al. [115] have recently published a study with an unmatched cohort of 173 ORN cases and 1086 controls. They also explored the use of ML models (logistic regression, random forest, support vector machine and a random classifier reference) based on DVH parameters (D_{mean}, D_{min}, D_{max} and V65Gy) to predict ORN. They found that the logistic regression model was the overall best performing one with an accuracy of 0.64, which is similar to that obtained with our cohort, although no clinical variables were included. Chapter 6 discusses the results obtained from our CNN-based ORN prediction model. Chapter 7 explores multimodality fusion DL methods to combine radiation dose distribution maps and non-image data (i.e. clinical and demographic variables) for the prediction of ORN. Mandible ORN is a rare radiation-induced toxicity and the data sets available are naturally small. Low patient numbers make it difficult to attempt a multiclass prediction task where not only incidence but also ORN severity is predicted. The morbidity caused by ORN (at any grade) is such that the prediction of its incidence alone would already be an important clinical decision-support contribution.

Overfitting and poor generalisability are common problems when testing complex ML and DL models on independent data sets. We used a five-fold nested cross-validation scheme in our evaluation and none of the test data were used when training or tuning hyperparameters for the models that were applied to them. This represents a fair internal validation of the ML models. According to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement [96], prediction studies can be classified into three categories: model development, model validation or a combination of both. The work presented falls into the first category, where we describe the first steps towards the development of a mandible ORN incidence prediction model by comparing different ML models. Future work will include external validation of the model on an independent cohort from the Odense University Hospital group in Denmark.

The published study described in this Chapter aimed to open a new path towards personalised RT for HNC using ML to predict mandible ORN incidence. We proposed a new approach in the field of mandible ORN toxicity by using a prediction model for its incidence rather than just determining its potential contributing factors. We showed that this can be successfully done using ML methods and encouraged the transition to

ML-based prediction models for ORN as has already taken place for other HNC toxicity end points.

Chapter 6

Predicting ORN from radiation dose distribution maps

Chapter 3 (Section 3.3.2) described how deep CNNs can be used to automatically extract dosiomic features from radiation dose distribution maps. This Chapter describes a DL-based model trained on 3D radiation dose distribution maps of the mandible structure and compares it to a RF model trained on DVH parameters in the task of predicting the incidence of mandibular ORN. Further analysis is included on the effect of factors such as the choice of classification probability threshold or minimum follow-up time requirements for the control group on the model performance results.

The experiments described in this Chapter were first performed on Cohort 2 and then repeated on Cohort 3. Chapter 4 describes both cohorts (Section 4.1.1) and the methodology followed (Section 4.5) to obtain the data used here, i.e., the 3D mandible radiation dose distribution maps and the DVH data from the clinical RT treatment plan.

The results from some of the experiments on Cohort 2 described in this Chapter were presented as a poster at the Applications of Medical Artificial Intelligence (AMAI) workshop within the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022 conference and published in a subsequent proceedings book chapter [114].

6.1 DVH-based predictions

6.1.1 DVH metrics

For the work included in this Chapter, only the maximum dose (D_{max} and D2%), minimum dose (D_{min} and D98%), mean dose (D_{mean}) and median dose (D50%) DVH metrics for the mandible structure were used in the RF model.

6.1.2 Random Forest implementation

The `sklearn.ensemble.RandomForestClassifier` module was used to implement a RF classifier (Section 3.1.3 in Chapter 3) with a maximum number of estimators (i.e. number of built trees) of 10, a maximum tree depth of 50, a minimum number of samples at a leaf node of 1 and a minimum number of samples required to split an internal node of 0.5. Stratified 5-fold nested cross-validation (Section 3.3.1 in Chapter 3) was used, with embedded model hyperparameter optimisation (Section 3.2.1 in Chapter 3) performed using the `sklearn.model_Selection.GridSearchCV` module.

6.2 Dose map-based predictions

Different deep CNN models were trained on radiation dose distribution maps of the mandible for the binary classification of ORN vs. control cases. A 3D DenseNet121 [93] was trained on Cohorts 2 and 3; additionally, 3D DenseNet40 and 3D ShuffleNet [94] CNNs were also trained on Cohort 3 to assess performance with lower capacity networks. Chapter 3 (Section 3.1.5) describes the architecture of these networks. This section provides details on the data used and the implementation and training of the DL networks.

6.2.1 Mandible dose distribution maps

3D radiation dose distribution maps of the mandible were used as the input for the DL-based ORN prediction models. These were created by multiplying the clinical radiation dose distribution and the binary mask of the mandible structure, as described in Chapter 4 (Section 4.3).

6.2.2 CNN implementation

Implementation was performed using the Medical Open Network for Artificial Intelligence (MONAI) (https://monai.io/) Pytorch-based framework. The data were split into training, validation and test sets following a stratified 5-fold nested CV approach (Section 3.2.2 in Chapter 3). For the final training, an ensemble of models was trained to improve generalisation performance and to reduce the sensitivity of the model performance to stochastic noise of the training. In this work, the ensemble model was created by randomly initialising each model five times and each time, training the model on the training set of the outer fold. Due to the stochastic randomness of the weight initialisation and the selection of mini batches during training, this created five slightly different models for each outer fold. To calculate the prediction of this ensemble model, the predicted softmax probabilities of each of the five individual models were averaged for each class (i.e. soft voting). The Adam optimisation algorithm and the categorical cross entropy loss function (`torch.nn.CrossEntropyLoss`) were used in all models. A hyperparameter grid search was performed for each outer fold which included the following hyperparameters and values: dropout 0.6, 0.8; learning rate 0.01, 0.001, 0.0001; batch size 10, 16; weight decay 0.01, 0.001, 0001; epochs 50, 100, 300. Small 3D random rotation (-0.1 to 0.1 rad) and zoom (0.8 to 1.2) augmentations were applied to the training set. Based on the results of this approach, the DenseNet121 was trained for 300 epochs with dropout 0.8, batch size 10, learning rate 0.001 and weight decay 0.001. The DenseNet40 was trained for 50 (for CV folds 1, 2, 4 and 5) and 300 (for CV fold 3) epochs, dropout 0.8, batch size 10, learning rate 0.001 and weight decay 0.001. The ShuffleNet was trained for 50 epochs, dropout 0.8, batch size 16, learning rate 0.001 and weight decay 0.001.

6.3 Model performance

The predictive performance of the models was assessed in terms of their discriminative ability using the AUROC, sensitivity, specificity, and precision measures (Section 3.2.3). Tables 6.1 and 6.2 provide the ensemble and CV fold-specific model performance results, respectively. The ROC curves of the models (Figure 6.1) were compared with the DeLong nonparametric statistical test [130] using the pROC package [131] with the statistical software R (https://www.R-project.org/). The difference in AUROC between the RF and DenseNet121 models was not found to be statistically significant for either Cohort 2 or Cohort 3, with p-values of 0.24 and 0.60, respectively (significance level of 0.05). Table 6.3 provides the results from the DeLong's test on all combinations of the models explored

Table 6.1 Model discrimination performance for Cohort 2 and Cohort 3.

	Cohort 2		Cohort 3			
	RF	DN121	RF	DN121	DN40	ShN
AUROC (95% CI)	0.65 (0.57-0.73)	0.73 (0.65-0.80)	0.61 (0.53-0.69)	0.64 (0.56-0.72)	0.69 (0.63-0.76)	0.65 (0.59-0.72)
Accuracy	0.65	0.67	0.57	0.60	0.67	0.61
Sensitivity	0.66	0.53	0.64	0.62	0.71	0.70
Specificity	0.64	0.81	0.50	0.58	0.63	0.52
Precision	0.65	0.77	0.56	0.59	0.66	0.59

on Cohort 3. The AUROC differences between the DenseNet40 and the DenseNet121 and between the DenseNet40 and the ShuffleNet were the only ones found to be significant with a significance level of 0.05. However, after Bonferroni correction [120] for multiple comparisons was applied, resulting in a corrected significance level of 0.05/6 = 0.008 for each comparison, no statistically significant difference was observed between models' AUROC. It should be noted that although the predictive performance between models was not significantly different, the DenseNet40 and the ShuffleNet models have simpler architectures that result in significantly shorter network training times.

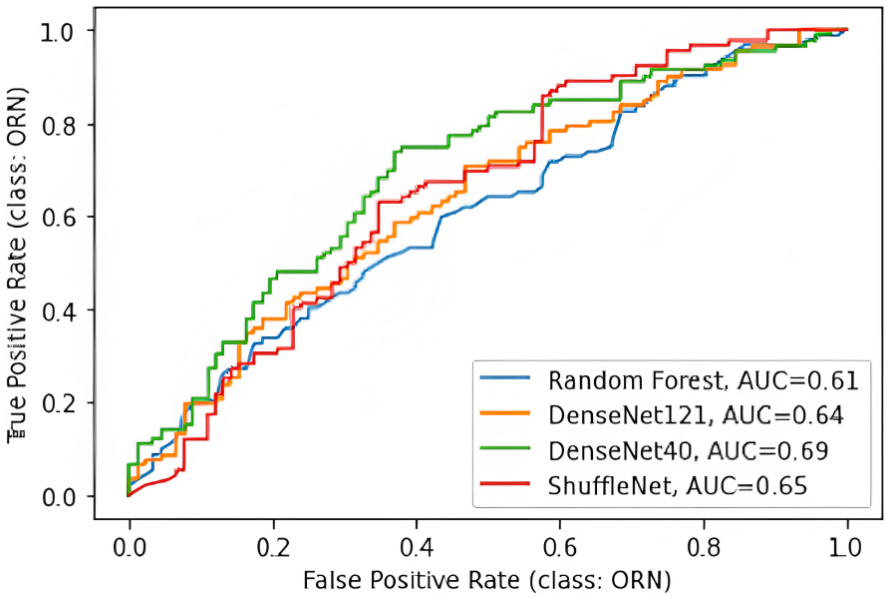


Fig. 6.1 ROC curves for the Random Forest, DenseNet121, DenseNet40 and ShuffleNet models.

6.4 Optimal classification probability threshold

The output of the binary classification CNN, for both the DenseNet121 and ShuffleNet models, is a predicted probability for each class (ORN and not-ORN) on the test dataset

Table 6.2 Model performance results for each of the 5 outer loops of the nested CV procedure obtained with the DenseNet40 CNN on Cohort 3.

		CV1	CV2	CV3	CV4	CV5
Accuracy	Ensemble	0.62	0.78	0.65	0.70	0.60
	Run1	0.65	0.78	0.59	0.68	0.53
	Run2	0.62	0.78	0.62	0.70	0.64
	Run3	0.62	0.73	0.70	0.70	0.64
	Run4	0.57	0.65	0.51	0.73	0.53
	Run5	0.57	0.73	0.59	0.62	0.53
Sensitivity	Ensemble	0.44	0.72	0.79	0.74	0.83
	Run1	0.44	0.78	0.79	0.63	0.78
	Run2	0.39	0.72	0.63	0.68	0.83
	Run3	0.56	0.67	0.74	0.84	0.89
	Run4	0.44	0.44	0.53	0.79	0.89
	Run5	0.44	0.61	0.74	0.58	0.89
Specificity	Ensemble	0.79	0.84	0.50	0.67	0.37
	Run1	0.84	0.79	0.39	0.72	0.28
	Run2	0.84	0.84	0.61	0.72	0.44
	Run3	0.68	0.79	0.67	0.56	0.39
	Run4	0.68	0.84	0.50	0.67	0.17
	Run5	0.68	0.84	0.44	0.67	0.17
Precision	Ensemble	0.67	0.81	0.62	0.70	0.56
	Run1	0.73	0.78	0.58	0.71	0.52
	Run2	0.70	0.81	0.63	0.62	0.60
	Run3	0.62	0.75	0.70	0.67	0.59
	Run4	0.57	0.73	0.53	0.71	0.52
	Run5	0.57	0.79	0.58	0.65	0.52

Table 6.3 Results from the DeLong statistical test on all model pair combinations for Cohort 3.

DeLong p-value	RF	DN121	DN40
DN121 vs.	0.60		
DN40 vs.	0.12	0.02	
ShN vs.	0.44	0.74	0.04

cases. A classification threshold of 0.5 (i.e. probability of 50%) is typically set for the final decision on the predicted class, where positive probabilities equal to or above 0.5 are predicted as positive cases and negative cases are predicted otherwise. The test accuracy results presented in Table 6.2 correspond to a 0.5 probability classification threshold. From a clinical perspective, however, a 0.5 probability classification threshold is not necessarily the most adequate one in the clinical decision-making process. For the task of making binary predictions on the incidence of ORN, correctly identifying ORN cases is as important as correctly identifying non-ORN cases for an efficient use of the clinical resources dedicated to the patients at a higher risk of developing ORN. Thus, in this case, the optimal probability threshold corresponds to that which results in equal sensitivity and specificity values, i.e. TPR = 1-FPR, which is the closest point to the upper-left corner of a ROC curve.

The ROC curves in Figure 6.2 plot the different TPR and FPR values for a range of probability threshold values for the ORN class for Cohort 3 predicted with the Random Forest, the 3D DenseNet121, the 3D DenseNet40 and the 3D ShuffleNet models, respectively. The optimal threshold was obtained by minimising the absolute difference between TPR and (1-FPR) as shown in Figure 6.2; the threshold values obtained were 0.51, 0.52, 0.53 and 0.62 for each of the four models considered. Table 6.4 includes the updated model performance metrics when these are re-calculated using the optimal classification probability thresholds.

Table 6.4 Model discrimination performance for Cohort 3 with the optimal classification probability threshold.

	RF	DN121	DN40	ShN
AUROC (95% CI)	0.61(0.53-0.69)	0.64 (0.56-0.72)	0.69(0.63-0.76)	0.65(0.59-0.72)
Classification threshold	0.51	0.52	0.53	0.62
Accuracy	0.56	0.60	0.66	0.63
Sensitivity	0.54	0.60	0.65	0.63
Specificity	0.58	0.61	0.66	0.63
Precision	0.56	0.60	0.66	0.63

6.5 Minimum follow-up time for controls

One of the main differences between Cohorts 2 and 3 is that only controls with at least 3 years of follow-up time were included in Cohort 2 whereas Cohort 3 does not have a minimum follow-up time for the control group. The median follow-up time for Cohort

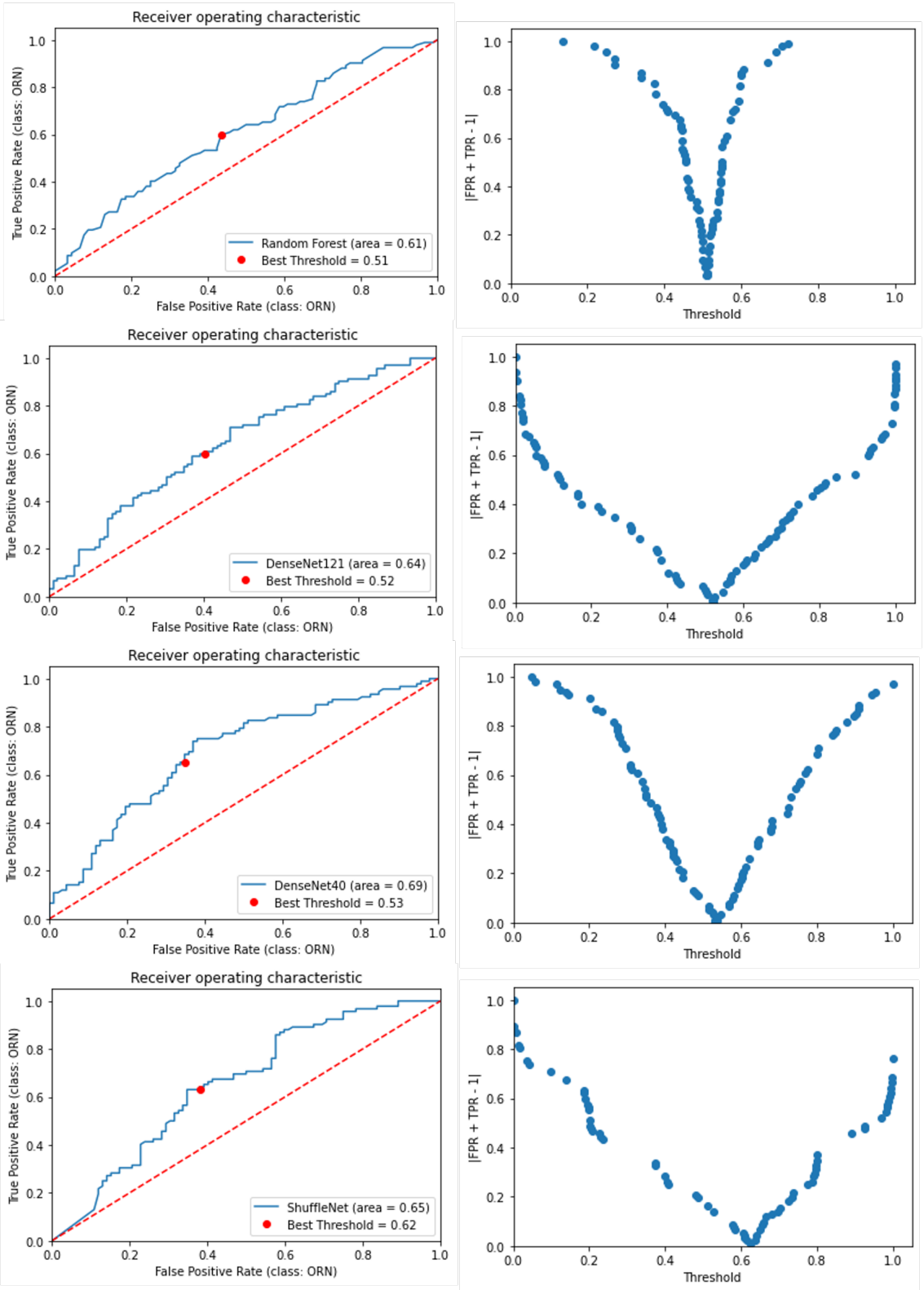


Fig. 6.2 ROC curves for the ORN class predicted probabilities with the (from top to bottom) Random Forest, DenseNet121, DenseNet40 and ShuffleNet models (left) and plots to find the optimal classification probability threshold for balanced sensitivity and specificity on Cohort 3 (right).

2 is 4.1 (IQR 3.0) and 4.3 (IQR 1.6) years for the ORN and control groups. The median follow-up time for Cohort 3 is 4.0 (IQR 3.2) and 4.3 (IQR 1.9) years for the ORN and control groups, respectively. For this same cohort, the median time from end of RT to ORN diagnosis for the ORN group was 1.0 (IQR 1.7) years. To investigate the effect of setting a minimum follow-up time, I repeated the univariate analysis and re-trained the 3D CNN for subsets of Cohort 3 with minimum follow-up times of 1, 2 and 3 years for the control group. The results are presented in Table 6.5. In each subset the corresponding number of ORN cases per primary tumour site was excluded to maintain the class balance and the primary site-based control-case matching. While the CNN prediction performance does not significantly vary with different minimum follow-up time thresholds on the control group for Cohort 3, the significance of the DVH metrics becomes more similar to that of Cohort 2 – which had a minimum follow-up time of 3 years – with maximum doses becoming the strongest discriminators between the two groups. However, the effect size of Cohort 2's maximum dose variables remains higher than that of Cohort 3.

Table 6.5 Univariate analysis, effect size and dose map-based prediction results for subsets of Cohort 3 with minimum follow-up times of 1, 2 and 3 years.

	Cohort 3, FU > 1 year		Cohort 3, FU > 2 years		Cohort 3, FU > 3 years	
	p-value (MWU)	Effect size (Cohen's d)	p-value (MWU)	Effect size (Cohen's d)	p-value (MWU)	Effect size (Cohen's d)
DVH metrics						
Dmax	0.043	0.250	0.040	0.255	0.025	0.285
D2%	0.014	0.303	0.017	0.296	0.009	0.339
Dmin	0.418	0.079	0.516	0.030	0.439	0.052
D98%	0.484	0.070	0.626	0.013	0.551	0.007
Dmean	0.021	0.316	0.077	0.228	0.060	0.240
D50%	0.023	0.292	0.064	0.230	0.060	0.230
ShuffleNet CNN (dosemaps)						
AUROC	0.67		0.63		0.66	
(95% CI)	0.61-0.74		0.56-0.71		0.58-0.73	
Accuracy	0.62		0.60		0.58	
Sensitivity	0.63		0.67		0.66	
Specificity	0.61		0.52		0.51	
Precision	0.62		0.59		0.57	

6.6 Discussion

In this Chapter, I have explored the use of CNN models to predict mandibular ORN incidence using clinical 3D radiation dose distribution maps. This is a novel approach to toxicity modelling for mandibular ORN as it uses the actual RT dose distribution rather than the more traditionally used DVH parameters.

The results presented here show that the CNN approach was able to discriminate between ORN and control cases based on the 3D mandible dose distribution maps (Table

6.1). However, as shown in Table 6.2, the performance of the model was highly dependent on the test-train data split (i.e. classification performance varied between the outer loop CV folds). This may be due to the high variability in the anatomical localisation of the radiation dose distribution of our cohort, suggesting that training using a larger cohort will lead to improved classification accuracy and robustness. Moreover, in some of the outer loop CV folds of the nested CV process, there was some variation between the ensemble models (Table 6.2), i.e. there was stochastic noise in the model training process for a given test-train data split.

The differences in AUROC between the RF and the CNN models were not found to be statistically significant based on the Bonferroni-corrected DeLong statistical test. However, the AUROC of the CNN models was 4.9%, 6.6% and 13.1% higher for the DenseNet121, ShuffleNet and DenseNet40 models, respectively, compared to the RF model. This could be due to the fact that the entire DVH was not available to the RF model, i.e. only maximum, minimum, mean and median doses were included as variables. However, the superior performance of the CNN models (even if not statistically significant) suggests that there may be more useful information for predicting ORN in the dose maps than just these DVH metrics. Precisely what this extra information is requires further analysis, and in Chapter 8 I will explore interpretability methods that aim to identify which areas of the dose distribution maps are getting more attention from the CNN model and are potentially contributing most to the final prediction. The inclusion of the spatial information in the dose maps may contribute to an improved performance as there are features such as the mandible volume that can be extracted from the mandible dose distribution maps but are not DVH dosimetric parameters that have previously been associated with ORN incidence [64, 65].

A recent study [115] concluded that DL methods based on 3D radiation dose distribution maps of the mandible and surrounding anatomy were outperformed by ML methods based on DVH data in predicting ORN development. Our earlier work [114], described in this Chapter, found no statistically significant difference in performance between ML and DL methods, although the DL models had slightly higher AUROC. There are several factors that could contribute to these seemingly contradictory results, and direct comparison between the two studies is not straightforward. A discussion on this is included in a letter to the editor accepted for publication in Advances in Radiation Oncology [132]. For instance, while only IMRT cases were included in Humbert-Vidan et al. [114], this was only true of a subset of the data used in Reber et al. [115]. Furthermore, even though data-level class imbalance handling was applied for the DL models in Reber et al. [115], further investigation would be required to assess how their results would compare if a

class-balanced matched cohort was used instead. In addition, different DVH parameters were used for the ML models in the two studies. There were also differences in terms of data preparation for the DL models. In Reber et al. [115] the 3D dose maps were cropped based on the mandible segmentation whereas in our work they were masked by the mandible segmentation. Further differences can be observed in model training (e.g. augmentation strategy) and possibly architecture (our work used a DenseNet with 121 layers but the number of layers for the models used in Reber et al. [115] was not specified). These differences can significantly impact model performance and further research is required to evaluate their specific impacts before clear conclusions can be drawn about the relative merits of traditional ML and DL models for this application. Additionally, when comparing dose map-based DL approaches to simpler DVH-based ML methods, it is important to also factor in the overall training time of the models, which depends on the number of trainable parameters. While the ML methods required between 10 and 20 minutes overall (including the hyperparameter optimisation, training and inference processes), the 3D CNN models required between 2 days (DenseNet40) and 5 days (DenseNet121).

Men et al. [103] included the CT images and the segmentation of the organ at risk along with the dose distribution maps as CNN inputs. By masking the dose map with the mandible segmentation, we are including information of its structure in the CNN while also simplifying the task by excluding potentially less relevant dosimetric information outside of the mandible. This study aimed at a direct comparison between DVH-based and dose map-based prediction models; future work will explore the effect of CT images as an additional CNN input.

Ibragimov et al. [102] and Zhen et al. [111] used transfer learning to pre-train their CNN on CT images. The training weights of the pre-trained CNN were then fine-tuned using the dose maps. Transfer learning may be used to enhance the model performance, especially when the study data set size is small. Dose distribution maps have very different image features to CT images, with smoother edges and contrast gradients. Future work will explore the effect of transfer learning on our results by pre-training on CT images, dose maps and a combination of both (i.e., pre-training with two inputs).

As discussed in Chapter 2 (Section 2.4.2), ORN has a multifactorial aetiology with radiation dose, clinical and demographic variables as potential risk factors. In this Chapter I have focused on radiation dose as the only risk factor but the inclusion of non-dosimetric clinical parameters into the model would be of great clinical value. Moreover, there are cases where ORN develops away from the high radiation dose region within the mandible and the correlation between ORN incidence and intermediate or low radiation doses is less obvious. Particularly in those cases, non-dosimetric parameters may play an important role

in the development of ORN. In Chapter 7 I will explore the inclusion of clinical variables into the DL-based prediction model to further improve its predictive performance.

Due to mandibular ORN being a rare toxicity, the case numbers are naturally low. In Chapter 9 I will describe the PREDMORN study, which is an initiative for obtaining a larger multi-institutional ORN population that will enable a more thorough evaluation of the potential of CNNs in ORN prediction as well as to perform an external validation of the models.

Although Cohort 2 has shown a larger effect size than Cohort 3 for its most highly discriminative DVH-based variables, a better control-case matching methodology was followed in the latter. Thus, the work described in subsequent Chapters is carried out using Cohort 3 only. Moreover, Cohort 3 will be included in the PREDMORN multi-centre study, the results of which I would like to compare to the work in this thesis.

Finally, to take full advantage of a DL model using dose distribution maps, future work will include knowledge of the actual ORN region or at least the ORN localisation within the mandible in the training data of the model.

Chapter 7

Combining image and tabular data

In Chapter 2 I reviewed the existing work on using traditional multivariate analysis to combine the clinical and demographic risk factors that may potentially contribute to an increased probability of developing mandibular ORN, in addition to the risk caused by radiation dose. In Chapter 3 I discussed how deep CNNs can automatically extract dosiomic features that can then be used, instead of the DVH metrics, in the prediction of radiation-induced toxicities. This Chapter explores DL multimodality fusion strategies for the prediction of mandibular ORN using the 3D radiation dose maps (image data) and clinical and demographic variables (tabular data) from Cohort 3.

7.1 Single-modality predictions

Using the data from Cohort 3 (Table 4.4), two separate single-modality models were trained on the image and tabular data, respectively. For the first, the 3D DenseNet40 CNN trained on radiation dose distribution maps in the experiments described in Chapter 6 is utilised for the work described in this Chapter. For the second, a random forest was trained on the corresponding clinical and demographic variables using the same model training methodology described for the RF model included in Chapter 5 (Section 5.3). For the grid search CV procedure, the following hyperparameters were considered: bootstrap (True, False), maximum depth (1, 2, 10, 20, None), maximum features (auto, sqrt), minimum samples per leaf (1, 2, 4), minimum samples per split (2, 5, 10) and number of estimators (200, 500, 1000, 1300, 1700, 2000).

Table 7.1 provides a summary of the clinical and demographic variables considered for Cohort 3 and the corresponding results from a univariate analysis (Chi squared or Mann-

Whitney U tests for categorical and continuous variables, respectively). Control-case matching for Cohort 3 was based on primary tumour site and treatment year; thus, primary tumour site and RT technique (which is largely dependent on treatment year and was equal for both groups anyway) are not considered as variables here. Even though only smoking status showed a significant difference between the two groups in the univariate analysis, all variables shown in Table 7.1 were included in the RF model. The two groups are similarly differentiated by the 'previous' and 'current' variables for smoking and alcohol status. Smoking status and alcohol consumption were considered as positive if reported as such within two months of the start of the radiotherapy treatment course. Categorical variables were dichotomised (0 or 1) and the variable 'age' was normalised to values between 0 and 1.

Table 7.1 Univariate analysis results for clinical and demographic variables for Cohort 3.

	ORN	Control	p-value
Age (median (IQR))	62 (13)	61 (15)	0.455
Gender: male/female	66(72%)/26(28%)	72(78%)/20(22%)	0.395
Smoking	47(51%)	21(23%)	0.000
Alcohol	71(77%)	63(69%)	0.246
Pre-RT extraction	55(60%)	50(54%)	0.551
Pre-RT surgery (PORT)	35(38%)	35(38%)	1.000
Chemotherapy	59(64%)	57(62%)	0.879

7.2 Multimodality fusion

The fusion of pixel data (images) with tabular data has already shown improved performance over single modality models in several studies [133] within the clinical prediction and diagnosis fields using radiological images. The different strategies followed by most of these studies have been classed as *early or feature level* fusion, *joint or intermediate* fusion and *late or decision level* fusion, based on when the fusion of data takes place [134, 133]. Figure 7.1 by Huang et al. [133] describes these three fusion strategy groups and further splits early and joint fusion into types I and II depending on whether the fusion is with original or extracted features. In early fusion, the inputs from different modalities, some of which may be features extracted by a ML algorithm, are combined into a single vector that is then fed into one single ML model. In joint fusion, in at least one of the modalities combined, the features are learned using a feature extraction neural network model. The combined features are input into a final neural network, the loss of which is

backpropagated to the first feature extraction neural network model(s). Finally, in late fusion, the final predictions from multiple models are combined to make a final decision. Both early and joint fusion strategies are able to model the interactions between features from different modalities. Joint fusion is thought to result in better feature representations due to the backpropagation of the combined model loss to the feature extraction neural networks during training. However, joint fusion can result in a more complex network design than early or late fusion. The following sections describe the different fusion strategies considered in this work and the prediction performance results with each method are summarised in Table 7.2. Given the small dataset size available for this work, I have only considered early and late fusion strategies [133].

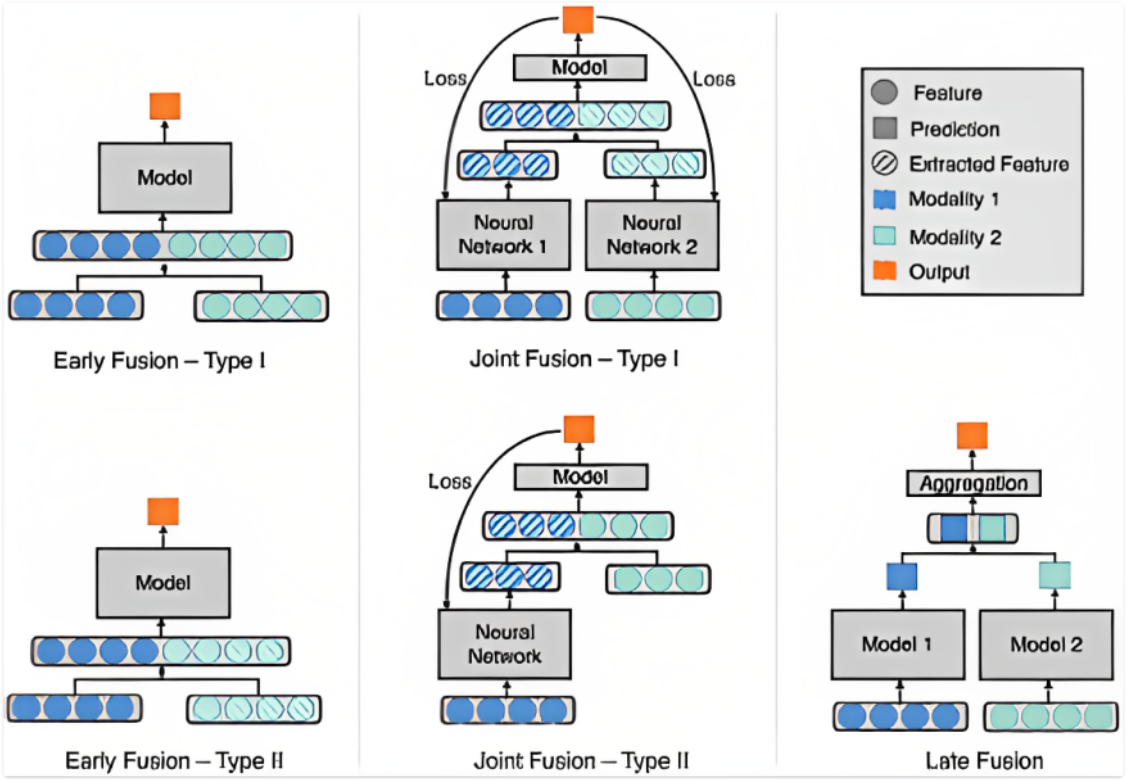


Fig. 7.1 Early (feature level), joint (intermediate) and late (decision level) multimodality fusion strategies using deep learning. In both early and joint fusion strategies, the features may be original or extracted with a machine learning algorithm (e.g. image features extracted with a CNN). When the inputs from at least one modality are extracted or learned features (e.g. predicted probabilities), the fusion strategy is considered type II; if the inputs are original features (i.e. not extracted), it is considered type I. Figure source: Huang et al. [133].

7.2.1 Type II early fusion

A 3D CNN was designed based on the 3D DenseNet40 to train on the dose maps and then concatenate the extracted image features with the tabular data into one single feature vector before feeding it into a final linear layer (Figure 7.2) following the type II early fusion strategy described in Figure 7.1. The early fusion 3D CNN was trained following a stratified 5-fold nested CV approach for 50 (CV1, CV2, CV4 and CV5) and 300 epochs (CV3), with a dropout rate of 0.8, a weight decay of 0.001, a batch size of 10 and a learning rate of 0.001.

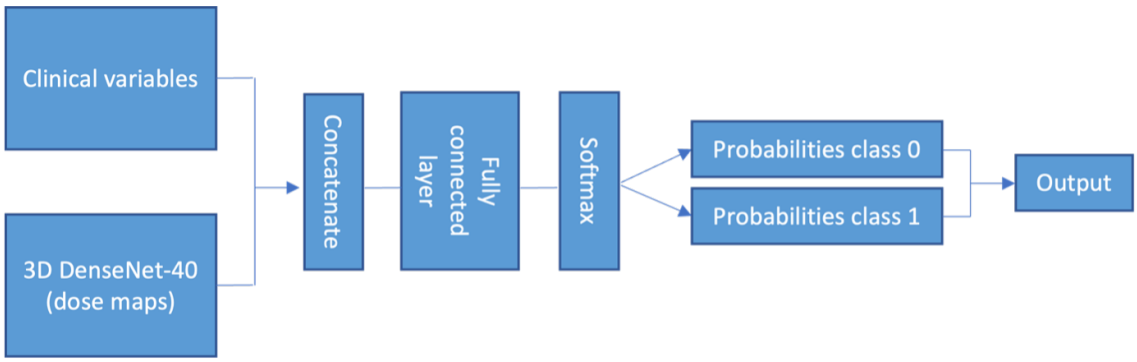


Fig. 7.2 The image features extracted from the radiation dose maps using a 3D DenseNet40 were concatenated with the clinical variables into one single vector using a type II early fusion strategy. The combined feature vector was input into a fully connected layer for classification of ORN vs. controls. A final softmax activation layer was added to obtain the class predicted probabilities.

7.2.2 Late fusion

A soft-voting ensemble approach was taken following the late fusion strategy illustrated in Figure 7.1: I combined the outputs from the 3D DenseNet40 and the RF models by averaging the predicted classification probabilities for each of the two classes (ORN and no ORN) to obtain the final class decision on a case-by-case basis for the test dataset (Figure 7.3). The 3D DenseNet40 CNN was trained on dose distribution maps of the mandible and implemented as described in Chapter 6, Section 6.2.2. The RF was trained on clinical and demographic variables and implemented as described in Section 7.1 of this Chapter.

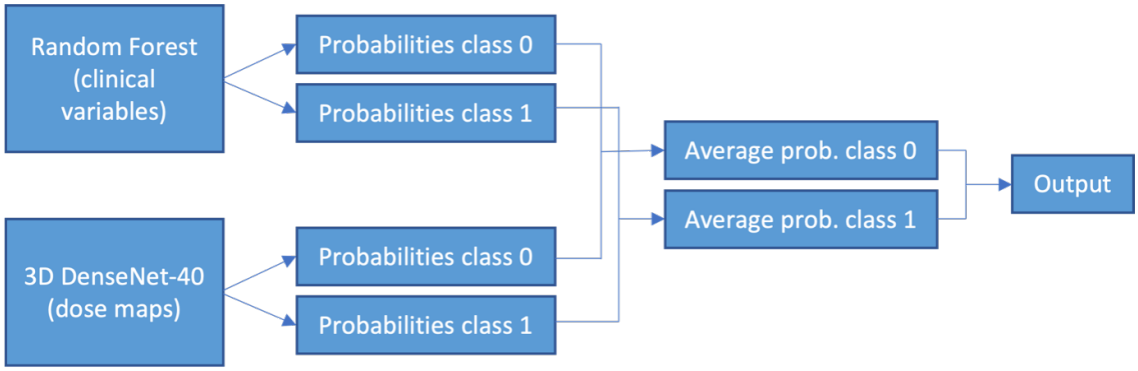


Fig. 7.3 Schematics of the late fusion strategy followed for ORN prediction from dose maps and clinical/demographic variables.

7.3 Model comparison

Table 7.2 below provides the results from the RF and DenseNet40 single-modality models and from the early and late multimodality fusion methods in the binary prediction of ORN. Figure 7.4 shows a comparison of the models' AUROC for the ORN classification. The ROC curves for all the model pair combinations were compared with the DeLong nonparametric statistical test [130]; the results are shown in Table 7.3. The other model performance metrics considered (accuracy, sensitivity, specificity, precision and F1 score) were calculated based on a classification probability threshold of 0.5.

Table 7.2 Summary of model performance results for ORN classification for the single modality models and the early and late multimodality fusion strategies considered.

	Random Forest (clinical variables)	DenseNet40 (dose maps)	Early fusion	Late fusion
AUROC (95% CI)	0.60 (0.53-0.67)	0.69 (0.63-0.76)	0.68 (0.61-0.75)	0.70 (0.64-0.77)
Accuracy	0.59	0.67	0.69	0.67
Sensitivity	0.73	0.71	0.65	0.73
Specificity	0.45	0.63	0.72	0.61
Precision	0.57	0.66	0.70	0.65
F1 score	0.64	0.68	0.67	0.69

Table 7.3 Results from the DeLong statistical test.

DeLong p-value	RF	DN40	Early fusion
DN40 vs.	0.09		
Early fusion vs.	0.09	0.72	
Late fusion vs.	0.03	0.36	0.44

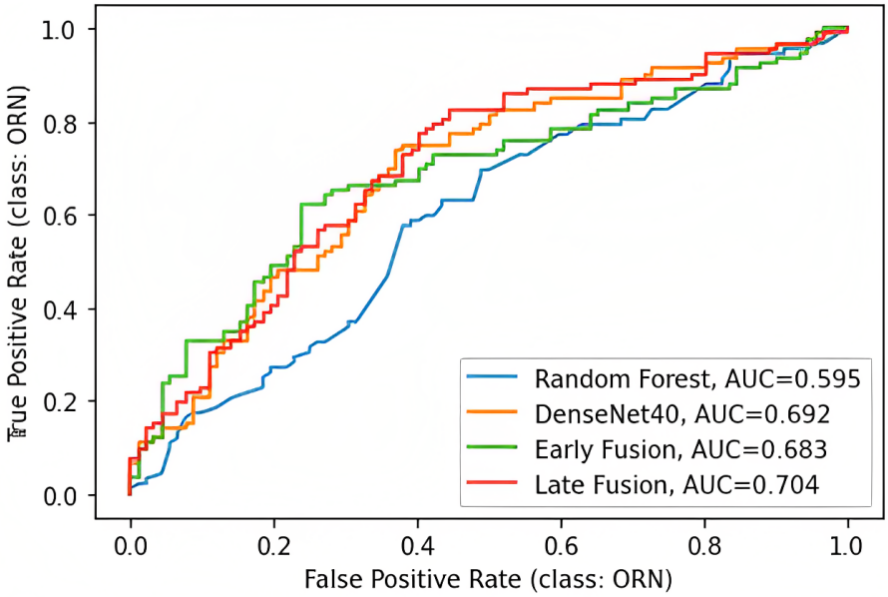


Fig. 7.4 ROC curves for the random forest and DenseNet40 single-modality models and the early and late multimodality fusion strategies considered for ORN classification.

7.4 Discussion

The development of radiation-induced toxicities is a multifactorial process. Existing DVH-based prediction models use traditional multivariate analysis to combine all the potential risk factors. However, with an image based NTCP modelling approach, the combination of dosimetric information with the other potential risk factors is perhaps not as trivial as in a multivariate analysis. Multimodality fusion is the next natural step in the process of implementing DL methods in the field of radiation-induced toxicity prediction based on radiation dose distribution maps. As recommended by Huang et al. [133], multiple fusion strategies should be compared and reported. In this thesis I have explored early and late fusion strategies, both recommended for small datasets [133], for combining radiation dose maps and clinical and demographic variables in the prediction of mandibular ORN incidence.

Ibragimov et al. [102] reported improved predictive performance with a late fusion multimodality DL strategy for the prediction of hepatobiliary toxicity after liver stereotactic body RT. They trained a CNN on 3D portal vein dose plans analysis and a fully connected neural network (FcNN), a RF and a SVM on clinical variables. They compared the prediction accuracy of the single-modality models (AUROC 0.79 for the CNN, 0.84 for the FcNN, 0.79 for the RF and 0.76 for the SVM) to that of a weighted sum model. They obtained the best performance with the combination of the CNN and the FcNN (AUROC

0.85), with a weight of 0.5 each. Although they did not report on the statistical significance of the AUROC differences, they concluded a superior performance with the CNN+FcNN fusion model.

In our study, the highest AUROC was observed with the late fusion approach, which was statistically significantly different to that of the RF single-modality model with a significance level of 0.05. However, after Bonferroni correction [120] for multiple comparisons was applied, resulting in a corrected significance level of 0.05/6 = 0.008 for each comparison, no statistically significant difference was observed between models' AUROC. No statistically significant differences in AUROC were observed between fusion strategies and between both fusion models and the DenseNet40 model trained on dose maps only. This is most likely due to the lack of discriminative contribution observed from clinical variables, which in turn resulted in a poorly predictive RF model. Although not intentionally, the clinical and patient characteristics of the ORN and control groups are very similar in Cohort 3. Further work will repeat this analysis on a larger and more diverse cohort (see Chapter 9 for a description of the PREDMORN multi-centre study). Additionally, with a larger dataset, joint fusion strategies will also be explored.

The fusion approaches explored in this Chapter are static, i.e. the same trained fusion network is used for inference regardless of the potential inherent variations in input datasets. However, the inherent noise in multimodal medical data may result in differences in informativeness from each modality and feature. The concept of dynamic multimodal fusion has been recently introduced [135, 136] to adaptively fuse multimodal input data during inference. The informativeness is modelled for each modality and feature and used to adjust the importance of the features/modalities in the final fusion step. Future work will explore joint fusion strategies for the prediction of ORN as well as the implementation of dynamic multimodal fusion methods.

To my knowledge, no previous work has been published on the use of multimodal fusion DL methods to combine dose distribution maps and clinical variables in the prediction of mandibular ORN. The work presented in this Chapter demonstrates the potential of DL in the prediction of the multifactorial side effects resulting from radiotherapy treatments.

Chapter 8

Interpretability of a deep CNN-based ORN prediction model

In Chapter 6 I explored the use of CNNs to classify 3D radiation dose distribution maps into ORN or control cases, thus predicting the probability of ORN incidence. CNNs are generally considered to be non-interpretable ML methods [137] and often referred to as 'black boxes' because interpreting the representations and intermediate outputs of the inner (hidden) layers of their architecture is not as straightforward as with other ML methods such as decision trees. Being able to interpret the different steps of the decision process of a model allows the user to identify its limitations and gain the users' trust. Moreover, a certain degree of explainability is required by the General Data Protection Regulation (GDPR) law in a DL-based decision-making tool before it can be used clinically [138, 139]. Finally, model interpretability methods might provide more in-depth information such as hitherto unknown feature associations, potentially leading to knowledge discovery. This Chapter provides a quantitative interpretability analysis of the predictions made by the DenseNet40 model (Chapter 6), including an analysis of potential associations between high attribution regions and the different dose levels within the radiation dose distribution maps of the mandible. First, an analysis is included on the laterality correlation between attribution maps, high dose region and ORN region. Second, the spatial overlap between attribution maps and dose regions is assessed. Finally, a dose level-based pixel attribution analysis is performed to assess how much importance the model is giving to the high and low/intermediate dose regions separately.

8.1 3D GradCAM voxel attribution maps

Interpretability and explainability are two closely related but different concepts. Interpretable models inherently provide transparency on how the decisions were made (e.g., in a decision tree, the user can follow the node splitting process to understand which variables contributed most to the final decision). Explainable models are not only interpretable but also able to provide an insight into the causes of the algorithm decisions [137].

Interpretability can either be local or global [139]. Local interpretability is obtained on an image-by-image (or case-by-case) level whereas global interpretability highlights the common image features that the DL model considers more relevant across the entire dataset.

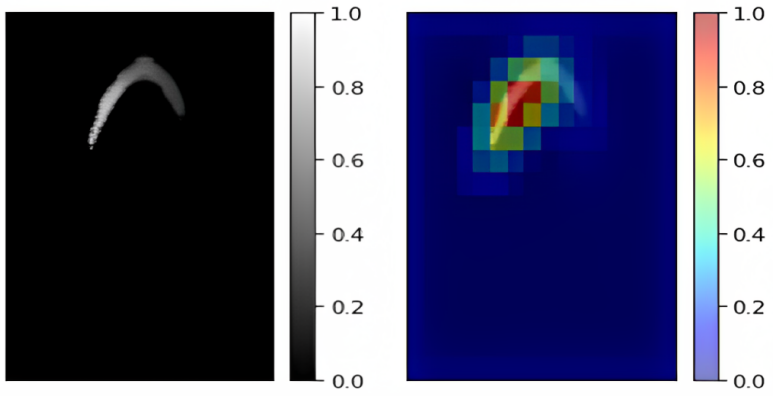
When a model is not intrinsically interpretable (e.g. random forests or deep CNNs), we can apply post-hoc interpretability methods to achieve transparency in its predictions [139, 140]. There are several post-hoc model interpretability methods described in the literature; the interested reader is directed to a recent review by Salahuddin et al. [139] for a more extensive description of these. This section focuses on attribution maps using the 3D CNN based Gradient-weighted Class Activation Mapping (3D Grad-CAM) method [141]. In this Chapter, the 3D Grad-CAM post-hoc interpretability method is applied locally, and an average of the case-by-case interpretability metrics is used to extract conclusions that can contribute to the explainability of the DL model. Pixel attribution maps provide post-hoc explanations of a model by highlighting the regions of the input image that the model considers most relevant for its predictions [139]. Selvaraju et al. [141] developed a novel 3D CNN-based Gradient-weighted Class Activation Mapping method (3D-GradCAM) that learns local geometric features within the input image based on the classification CNN's cost function. In this method, the global class discriminative localisation map for a given input image, $L_{3DGradCAM}$, is computed as per Equation 8.1 [141], where A_l are the feature maps in the last convolutional layer of the CNN and α_l designates the spatial importance for each feature map. The upsampled heatmap of the localisation map $L_{3DGradCAM}$ can be overlaid with the input image for easier interpretation (Figure 8.1).

$$L_{3DGradCAM} = ReLU(\sum_{l} \alpha_l A_l) \tag{8.1}$$

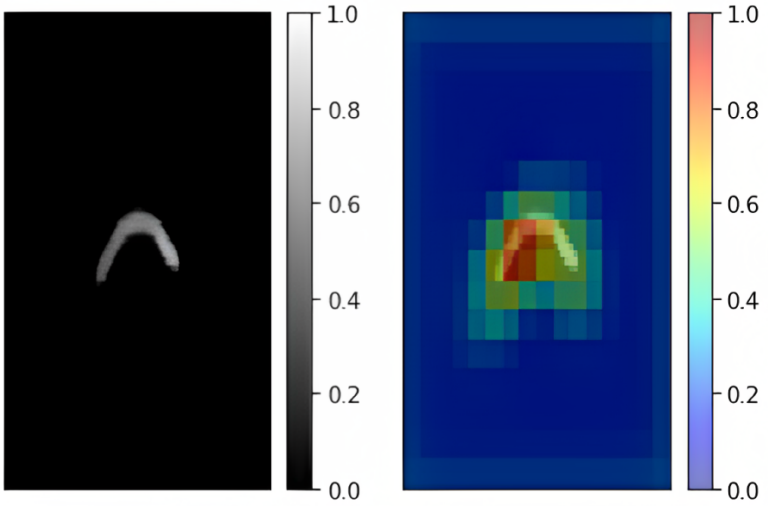
I implemented the 3D GradCAM pixel attribution method with the Captum [142] PyTorch-based library. Note that 3D GradCAM does not alter the training of the model but rather adds an extra layer during inference, which was applied at the end of the first

dense block of the 3D DenseNet40 CNN architecture, just after the last convolution layer. Attribution maps were normalised to within a pixel value range between 0 and 1 as per Equation 8.2, where PV_{min} and PV_{max} are the global minimum and maximum attribution values across the entire dataset, respectively.

$$AttributionMap_{normalised} = (PV - PV_{min})/(PV_{max} - PV_{min}) \tag{8.2}$$



(a)



(b)

Fig. 8.1 Axial slice of a radiation dose distribution map of the mandible (left) and a visual overlay of the corresponding pixel-attribution map obtained with the 3D GradCAM method (right). Both the dose and the attribution maps have been normalised to pixel values between 0 and 1, with high pixel values (white on the dose map intensity scale and red on the attribution map intensity scale) corresponding to high doses and high attribution values, respectively. Figure *a* shows an example where the high dose region received high attribution whereas figure *b* shows an example where the high attribution was given to the lower doses.

8.2 Laterality association between attribution maps, ORN region and high dose region

Amongst other risk factors (described in Chapter 2, Section 2.4.2), mandibular ORN is associated with radiation dose. In this section I investigate if the ORN region coincides with the higher radiation dose regions for Cohort 3. I then compare the laterality of the high attention regions to that of the high dose regions to investigate whether the DL network is giving more attention to the high dose regions to correctly predict ORN.

The laterality of the high dose region within the mandible dose distribution maps was defined based on the mean voxel intensity of the right and left halves of the 3D dose map (Figure 8.2). The same method was followed for defining the laterality of the high attention region within the attribution maps. Common space registration was applied when creating the mandible dose maps (Chapter 4, Section 4.3.3) and thus the image centre (both for the dose maps and the attribution maps) coincides with the anatomical centre of the mandible in most cases. However, because the same image splitting method has been used for dose maps and attribution maps, the comparison between the two is valid on a patient-by-patient basis even in the cases with slight deviations from the anatomical mandible centre.

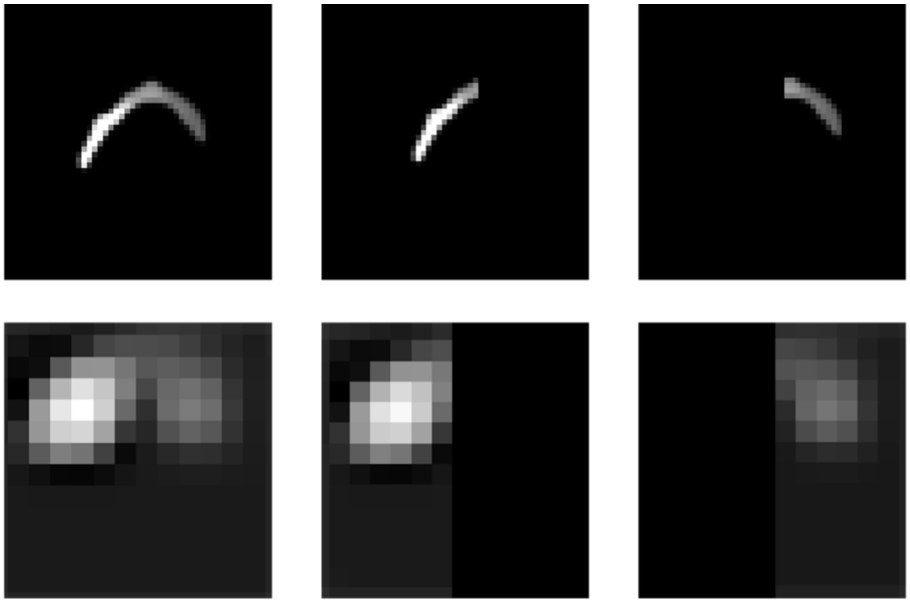


Fig. 8.2 Axial slice of the whole (left), right half (middle) and left half (right) mandible radiation dose distribution maps (top) and corresponding pixel-attribution maps (bottom).

8.2.1 ORN region vs. dose maps

The laterality of the ORN region was determined from the clinical dental notes and was left in 46.2% of cases and right in 36.3%. In 13.2% of cases the ORN region was bilateral and in 4.4% of the cases the ORN developed in the anterior region of the mandible. In 71.4% of ORN cases, the ORN region (or part of it for the bilateral cases) developed in the higher dose region.

8.2.2 Pixel-attribution vs. dose maps

The high attention region was found ipsilateral to the higher dose region in 73.2% of all subjects and in 72.5% of ORN cases. When considering the correctly (true positives and true negatives) and incorrectly (false positives and false negatives) predicted cases separately, this association between high attention and high doses was observed in 74.0% and 71.7% of cases, respectively.

8.2.3 Pixel-attribution vs. ORN region

As a result of the ORN-to-dose and attention-to-dose associations described above, the high attention region was found ipsilateral to the ORN region in 68.1% of the ORN cases.

8.3 Spatial overlap

This section aims to quantify the spatial overlap between the pixel attribution maps and the radiation dose maps. Different overlap metrics were computed: the percentage overlap and Dice similarity coefficient (DSC) [143]. For this analysis, the normalised attribution maps were thresholded based on the case-by-case mean attribution value. The thresholded pixel attribution maps were then converted into a binary attribution mask. The dose distribution maps were also converted into binary masks (with a 1 Gy threshold). Finally, the dose distribution and pixel attribution masks were overlaid and the overlap metrics were calculated (Figure 8.3). Table 8.1 provides the results from the overlap analysis. A sensitivity analysis was included by exploring different thresholding levels: the mean attribution map voxel level and 25% below and above the mean voxel value (Figure 8.4).

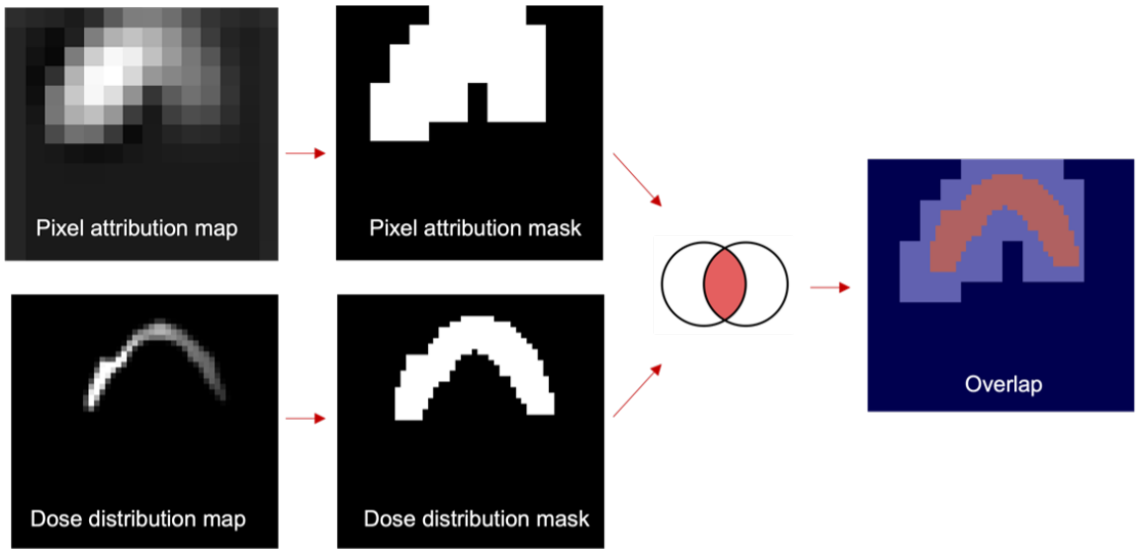


Fig. 8.3 Schematics of the overlap analysis workflow, where the dose distribution and pixel attribution maps are converted into binary masks that are then overlaid and overlap metrics are calculated.

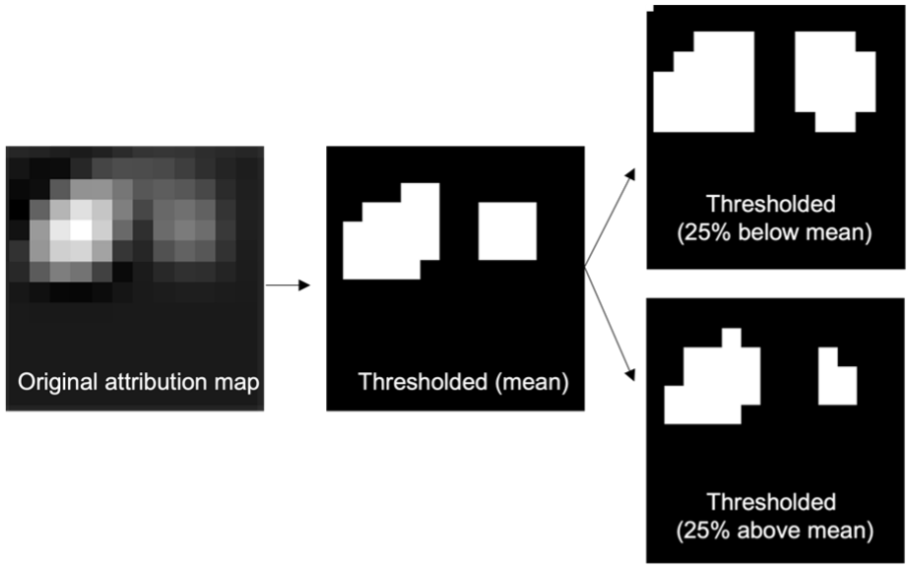


Fig. 8.4 Pixel attribution map (left) and corresponding binary mask with mean attribution value thresholding (middle). A sensitivity analysis was performed for all overlap metrics using a thresholding value of 25% below (right top) and above (right bottom) the mean attribution value.

8.3.1 Percentage overlap

The percentage overlap was calculated as the mean percentage of voxels within the overlap of each dose map mask / attribution map mask combination with respect to the total number of voxels in the smallest volume between the dose map and the attribution map masks as per Equation 8.3.

$$Overlap(DoseMap, AttMap) = \frac{|DoseMap \cap AttMap|}{min(|DoseMap|, |AttMap|)} \tag{8.3}$$

8.3.2 Dice similarity coefficient

Additionally, the spatial overlap accuracy was quantified in terms of the Dice similarity coefficient (DSC) calculated based on the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) values obtained from the overlap between the dose map and attention map binary masks (Equation 8.4).

$$DSC(DoseMap, AttMap) = \frac{2|DoseMap \cap AttMap|}{|DoseMap| + |AttMap|} = \frac{2TP}{2TP + FP + FN} \tag{8.4}$$

Table 8.1 Results from the spatial overlap between the pixel attribution maps (at different threshold levels) and the radiation dose maps. The metrics were calculated on a case-by-case basis and the median over the entire cohort is reported here.

Threshold value	% overlap (median, IQR)	DSC (median, IQR)
Mean attribution	97.43 (5.05)	0.927 (0.016)
25% below mean	98.88 (3.48)	0.913 (0.021)
25% above mean	95.89 (6.07)	0.940 (0.017)

8.4 Dose level-based pixel attribution analysis

This section provides two different approaches to a quantitative analysis on how much attention the 3D DenseNet40 ORN prediction model is giving to the different dose levels in the mandible radiation dose distribution map. First, the attribution maps were masked with low, intermediate and high dose regions. Second, the dose maps were masked with attribution maps thresholded at different attention levels. The methodology and results for both approaches are provided below.

8.4.1 Dose-based masked pixel attribution maps

For this analysis, the radiation dose distribution maps were separated into two binary masks according to the following dose thresholds: $D \geq 45Gy$ (high doses) and $1Gy \leq D < 45Gy$ (low/intermediate doses). The normalised attribution maps were thresholded based on the case-by-case 50^{th} percentile (i.e., the median) attribution value. Then, the pixel attribution map was then masked by the high and low/intermediate dose masks (Figure 8.5) and the maximum attribution value per dose level was calculated (Figure 8.6) for the overall dataset and the ORN and control groups separately.

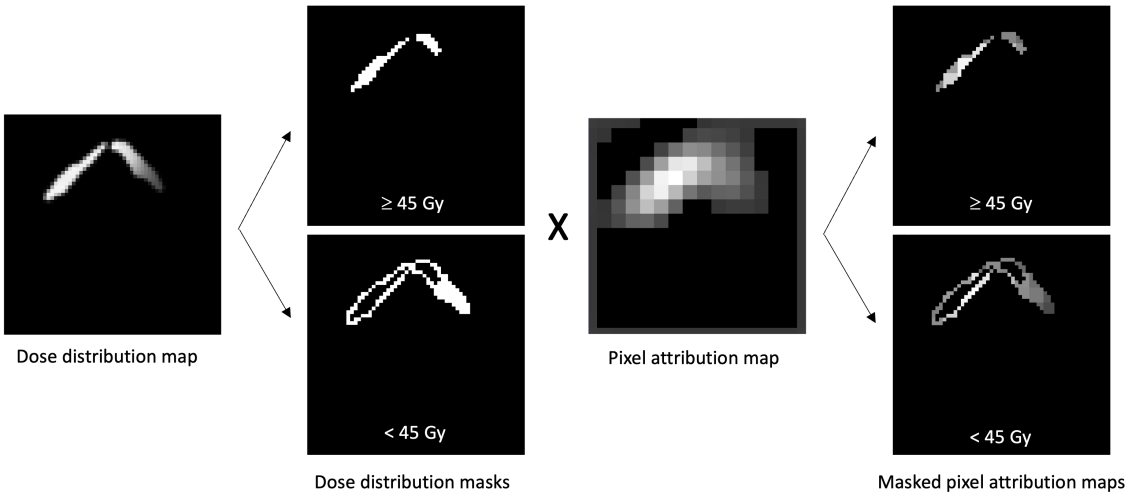


Fig. 8.5 Workflow for masking the pixel attribution map with the high and low radiation dose regions. The dose map is converted into high and low dose masks which, when multiplied by the attribution map, result in the dose-based masked attribution maps. High intensity pixels are shown in white and correspond to high dose and high attribution values in the dose distribution and pixel attribution maps, respectively. In the case shown in this image, the dose-based masked pixel attribution maps show that the low dose region is receiving the highest pixel attribution values.

8.4.2 Attention-based masked dose distribution maps

For this analysis, the pixel attribution maps were masked into the low attribution level (i.e., thresholded to contain the region with attribution values between the 50^{th} and 98^{th} percentile) and the high attribution level (i.e., thresholded to contain the region with attribution values above the 98^{th} percentile) (Figure 8.7). The dose distribution maps were then masked by these attribution masks and the maximum dose in the masked dose maps was calculated per attribution level (Figure 8.8) for the entire dataset and the ORN and control groups separately.

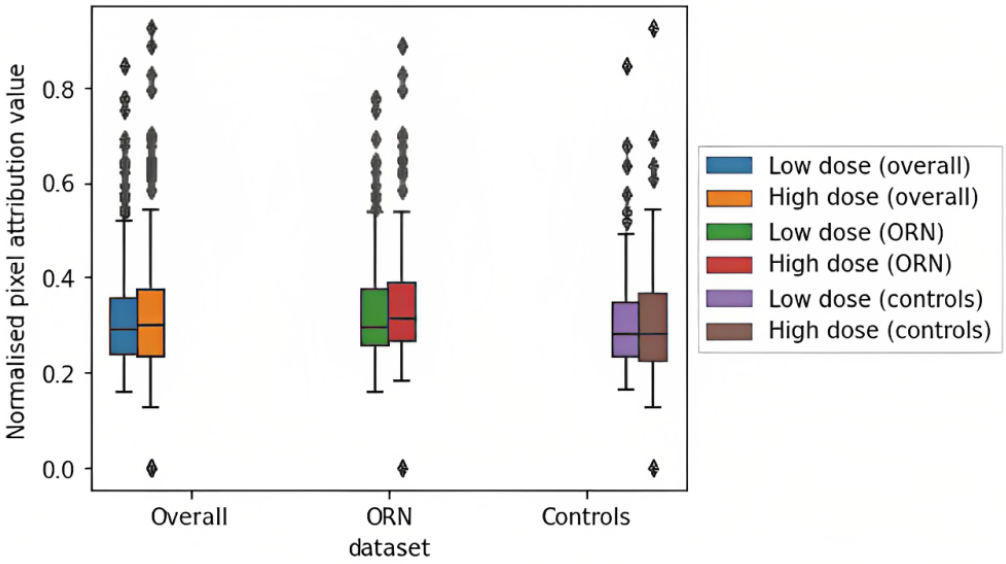


Fig. 8.6 Boxplots of the maximum attribution value in the pixel attribution maps masked with low and high dose regions for the entire cohort and the ORN and control groups separately. High dose regions generally contain higher maximum attention values, especially in the ORN group.

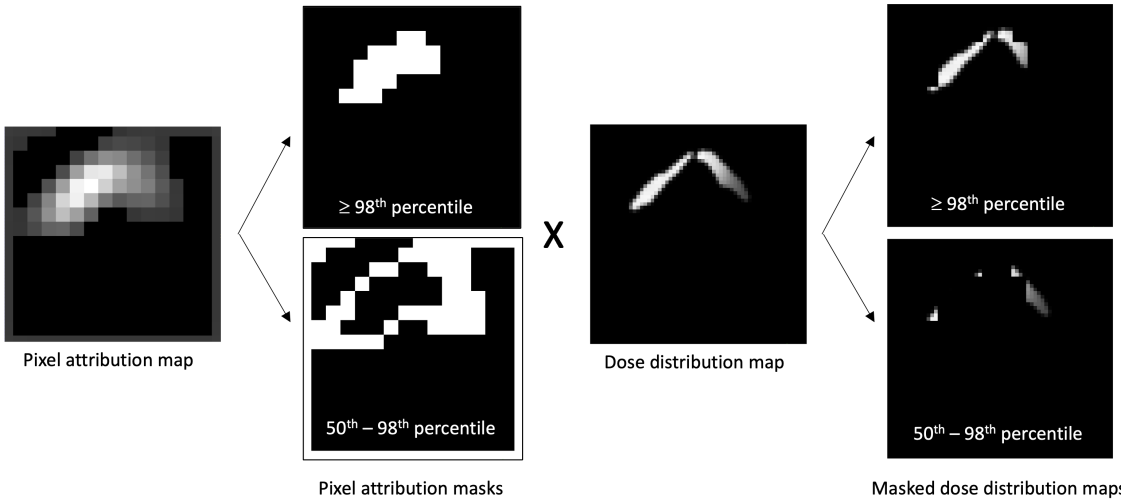


Fig. 8.7 Workflow for masking the dose distribution maps with the different levels of pixel attribution. The attribution map is converted into high and low attribution masks which, when multiplied by the dose map, result in the attribution-based masked dose maps. High intensity pixels are shown in white and correspond to high dose and high attribution values in the dose distribution and pixel attribution maps, respectively.

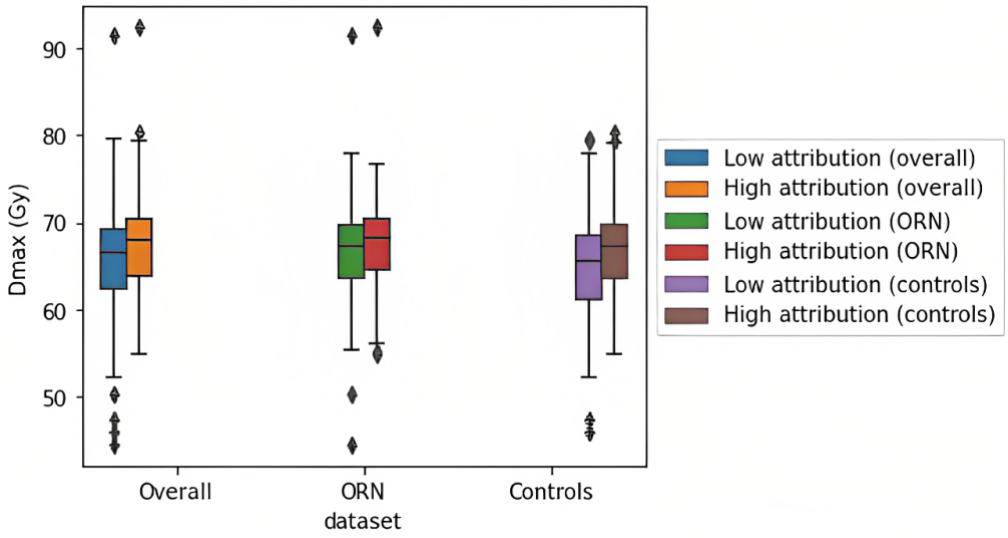


Fig. 8.8 Boxplots of the maximum radiation dose value in the dose distribution maps masked with low and high attribution levels for the entire cohort and the ORN and control groups separately. High attribution regions generally contain higher doses. The ORN group shows generally higher doses than the controls group but the difference in median D_{max} between high and low attribution masks is larger in the controls group.

8.5 Discussion

Explainability is key for achieving transparency in a DL-based toxicity prediction model and gaining trust for its clinical implementation as well as potentially facilitating its transferability to other domains [137]. Interpretability methods, both quantitative and qualitative, are a first step towards model explainability.

Previous head and neck cancer studies have investigated associations between spatial dose patterns and radiation-induced toxicities such as dysphagia [100] and trismus [106] to identify the most critical anatomical regions. These studies, however, used a traditional voxel-based approach to manually extract the highly associated voxel clusters. In other anatomical regions more sophisticated interpretability methods have been used. For example, Elhaminia et al. [144] implemented the DL-based Grad-CAM method for attention analysis and observed associations between abdominal spatial dose patterns and patient-reported toxicity. Liang et al. [145] also used Grad-CAM to interpret a 3D CNN thoracic toxicity prediction model and obtained associations between low and high grades of radiation pneumonitis and spatial dose patterns within the lungs. These studies, however, only provide qualitative analysis of the Grad-CAM results. Our work is, to our knowledge, the first study that analyses spatial dose associations with mandibular ORN incidence using the DL-based 3D Grad-CAM voxel attribution method. We also explore quantitative

analysis methods for the Grad-CAM results that have not yet been used in existing RT toxicity studies.

In the majority of ORN cases of our cohort, ORN developed on the side of the mandible with a higher mean planned dose. The 3D DenseNet40 model was able to capture this dosimetric association based on the radiation dose distribution maps of the mandible by giving high attention to the more highly irradiated mandible half in the majority of cases. Based on the high percentage overlap obtained between the radiation dose maps and the pixel attribution maps, it can be concluded that the DL model is focusing on the mandible dose for its predictions. This is perhaps not surprising as the DL model was trained on a dose distribution that was highly limited to the mandible structure. Based on the dose level-based pixel attribution analysis (Section 8.4, the high dose regions ($\geq$ 45 Gy) may generally be playing a more important role as it receives the highest attention, especially for the ORN group (as demonstrated by the boxplots in Figures 8.6 and 8.8). However, in a number of subjects (e.g., the case shown in Figure 8.1b), a higher attribution was given to the low dose regions.

An earlier statistical analysis by De Felice et al. [41] on a cohort of 36 patients from our centre found that a majority of ORN cases had mean and maximum doses above 60 Gy. A recent study by Möring et al. [63] also found that V60 Gy was significantly associated with an increased risk of ORN. Kubota et al. [62] found V60Gy >14% to be an independent risk factor for ORN, with V30Gy-V70Gy being significantly higher in the ORN group. Aarup-Kirstensen et al. [61] obtained significant differences between the ORN and control groups for intermediate and low doses between 30 Gy and 60 Gy.

Our results have shown that importance to high doses is more associated to ORN cases than controls. Although further work is needed towards a fully explainable model, this represents an important step towards gaining trust for the clinical implementation of a DL-based ORN prediction model. Furthermore, the relatively high laterality association between high attention regions and ORN regions implies that voxel attribution maps from the 3D DenseNet40 could be used to produce an ORN risk map. Future work will explore this possibility by utilising actual ORN region segmentations, either in the training or evaluation of the DL models.

The dose maps on which the 3D DenseNet40 was trained were highly cropped and excluded the dosimetric context outside the mandible. This was a limiting factor to the GradCAM analysis. Future work will repeat this experiment with a broader dose distribution combined with separate mandible masks and assess whether the 3D DenseNet40

is able to identify the dose in the mandible as the most important information for its predictions.

An alternative to attribution maps could be to perform a sensitivity analysis [137] where, for instance, the voxel intensity of the radiation dose maps could be manually altered to observe the resulting changes in the DL-based predictions. This approach was followed by Ibragimov et al. [113] to identify critical-to-spare anatomical regions of the liver during stereotactic body radiation therapy and produce a hepatobiliary toxicity risk map. Similarly to comparing different CNNs for model development (Chapter 6), future work will consider comparing a range of voxel-wise interpretability methods and analyse their adequacy for this task [146].

Chapter 9

The PREDMORN multi-centre study

One of the limitations in my experiments has been the small dataset size that results from the low prevalence of a rare toxicity such as mandibular ORN. On the other hand, while external validation of a model is a recognised [96] method for assessing how well the model would perform on unseen and independent data, using a more diverse dataset to train the model would certainly contribute to its generalisability. Moreover, unlike clinical trials, multi-centre studies using real world data can benefit from larger patient diversity.

Consequently, I initiated collaborations with five other teaching hospitals with well-maintained ORN databases and designed a multi-institutional study to develop, train and validate robust and generalisable NTCP models for mandibular ORN using the largest ORN dataset worldwide. Combining datasets from different institutions will not only improve the generalisability of the models but also highlight potential correlations between clinical practice and toxicity outcome. The study protocol has been registered in the OSF registries under the DOI: 10.17605/OSF.IO/V9JKR (https://osf.io/v9jkr).

This Chapter aims to describe the protocol itself and how it was designed as well as the challenges encountered during this collaborative effort. The actual protocol has been published in the Radiotherapy & Oncology journal as supplementary material in a Protocol Letter [147]. This study is expected to run beyond the time frame of my PhD; thus, the data analysis results will not be included in this Chapter but rather in subsequent Journal publications.

9.1 Study design

The PREDMORN study involves six teaching hospitals (see Appendix A): Guy's and St Thomas' NHS Foundation Trust (GSTT, UK) in collaboration with King's College London (KCL, UK), Odense University Hospital (OUH, DK), Catalan Institute of Oncology Girona (ICO, ES), University Medical Centre Groningen (UMCG, NL), MD Anderson Cancer Centre (MDACC, US) and Erasmus MC Cancer Institute (EMC, NL). Most of these centres already had a well-curated ORN database but only OUH, UMCG and MDA were able to provide full datasets for their entire HNC population. Therefore, a retrospective observational-analytical case-control study design [148] was agreed.

9.2 Patient selection

Since inclusion of the entire HNC population was not feasible for all centres, a 2:1 ratio of controls to ORN cases was agreed as a compromise to increase statistical power. For each centre, the distribution for the ORN group was obtained with regards to a) primary tumour sites (oral cavity, oropharynx, paranasal sinus/nasopharynx, larynx/hypopharynx, salivary glands and unknown primary (neck)) and b) treatment year for each primary tumour site. A 2:1 subset of controls was then randomly selected for each primary tumour site from the same treatment year (Figure 9.1). All other confounders are considered variables for the subsequent correlation and modelling analysis. A summary of the patient inclusion/exclusion criteria agreed for the ORN and control groups can be found in Table 9.1 below.

9.3 Data collection

The methods in this study were designed to only consider DVH data and clinical and demographic variables in the first instance. A second phase of this study will carry out further analysis and modelling with the inclusion of radiation dose distribution maps from all the institutions involved as a continuation of the DL-based work carried out in this thesis. The PREDMORN protocol benefited from the contributions and criticism by leaders from different disciplines across several organisations. After extensive discussions among the different experts involved (oncologists, oral medicine specialists, physicists, epidemiologists, etc.), relevant dosimetric, clinical and demographic variables were agreed (see Appendix B). To facilitate the data collection process, I provided each participating

Table 9.1 Inclusion and exclusion criteria for the PREDMORN study.

Inclusion criteria	• HNC cases treated radically with RT or chemo-radiation (CRT) and post-operative RT (PORT)+/- chemotherapy (C-PORT) using IMRT or VMAT. *Specific to ORN cases:* • Confirmed diagnosis of ORN (any grade). ORN defined clinically as 'an area of exposed bone in the mandible that had been present for at least 8 weeks in a previously irradiated field, in the absence of recurrent tumour' [19]. NB: cases with recurrences outside of the HN region are included. *Specific to control cases:* • Any histology except for T1/2N0 Larynx cases, as these do not receive significant dose to the mandible.
Exclusion criteria	• Datasets without available CT volume, RT structure or RT dose DICOM files. • ORN outside the mandible (e.g., maxilla). NB: if multiple ORN sites, cases are included if at least one of the sites is the mandible. • Cases with re-irradiation to the HN region. • Cases treated with less than a radical dose of radiation and/or those with life expectancy less than 12 months.

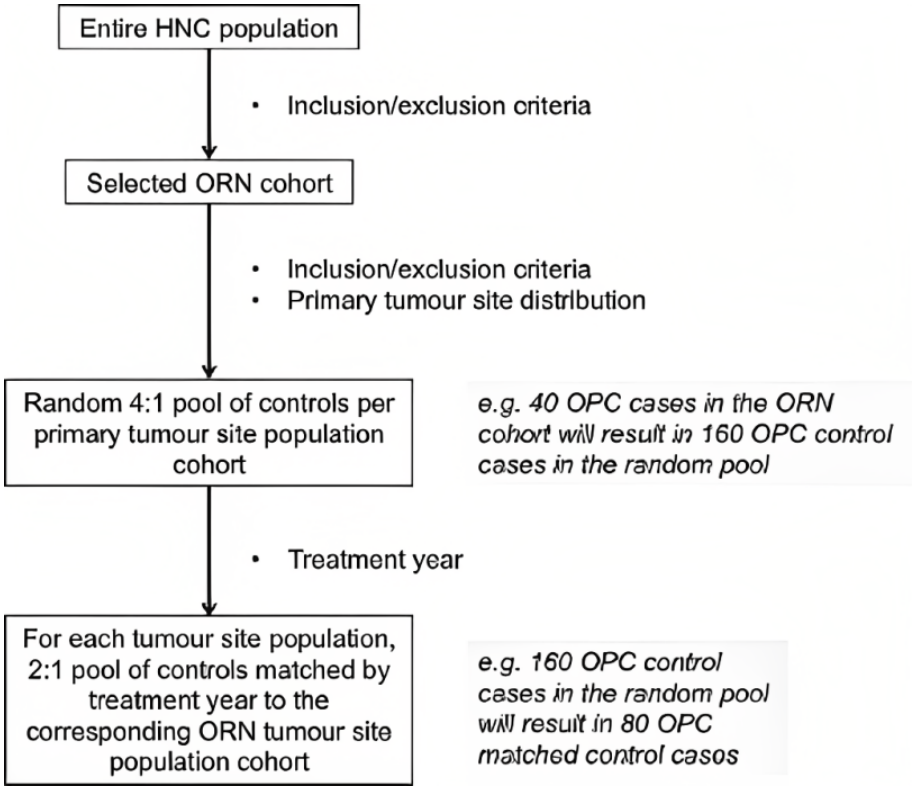


Fig. 9.1 Flow diagram of the control-case cohort selection process.

centre with an Excel database template that included all these variables with drop-down lists with the agreed options for each variable, where appropriate. For the dosimetric variables, the anonymised raw DVHs were exported by each centre and the R package 'DVHmetrics' was used to extract the agreed metrics. These included the following: Dmean, Dmax, Dmin, D2%, D5% - D95% in 5% steps, D98%, V5Gy – V75Gy in 5Gy steps in the first instance with smaller steps also being considered for future analysis.

9.4 Data transfer

The data transfer process was initiated in parallel to the protocol development and data collection processes at each individual organisation. The data transfer process is entirely dependent on obtaining a data transfer agreement (DTA). We initially explored a 'Joint Controllers' data transfer approach in order to facilitate joint use of the study data by all parties. However, a 'Controller to Controller' approach was considered more legally efficient by the organisations' legal teams. Thus, individual DTAs were agreed between KCL/GSTT and the other institutions individually, where each institution could use the data for its own purposes and KCL/GSTT requested access to that data. Ethic approvals were

already in place at each institution for the use of the clinical data shared with GSTT/KCL. A provision for the transfer of DICOM files for future image-based analysis collaborations was included in the DTAs agreed. The legal and procedural challenges encountered in setting up data transfer agreements and data transfer process of the PREDMORN study are described in an e-poster that will presented at the ESTRO 2023 conference [149].

9.5 Data modelling

Based on published literature I proposed three different ML-based data modelling approaches: a LASSO logistic regression (LASSO-LR) [150], a regularised greedy forest (RGF) [151] and a Bayesian neural network (BNN) classifier [152, 153]. In the LASSO-LR, the most relevant variables are 'automatically' selected by shrinking their coefficients within the model to zero (L1 regularization penalty). The model coefficients can be retrieved to obtain the importance assigned to each variable. In an RGF, relative importance scores are obtained for each input variable. For the BNN method, however, a prior variable selection step is required using methods such as recursive feature elimination or stepwise selection [154]. Although I have worked with EQD2-corrected doses (both DVH and dose map data) in the experiments included in this thesis, the agreement for this study was to include the fractionation schedule (i.e., total prescribed dose and number of fractions) as an additional confounder instead. Given the intrinsic correlation of the different dosimetric variables extracted from a DVH, a parallel principal component analysis (PCA) [155, 156] was included in the protocol as a method to create a new, independent and uncorrelated subset of dosimetric variables (principal components, PC) based on the DVH-based variables. One important disadvantage of this method is that PCs are not directly interpretable from the clinical perspective. This study explores the use of PCs into NTCP models in parallel with the more traditionally used DVH metrics, and a total of six different models are compared.

9.6 Model evaluation

Internal validation and external validation of a model are necessary steps [157, 158] to correctly assess the model performance and identify potential overfitting issues. In this study, both types of validation will be included, i.e. as a combination of the study types 1b and 3 described by Collins et al. [96]. Cases from one institution will be selected as the external validation dataset. Data from the other institutions will be combined into a

single dataset for model development and internal validation. This combined dataset will be split into training (60%), internal validation (20%) and test (20%) sets maintaining the original class ratios (i.e. ORN vs. control case numbers) following a stratified nested cross-validation (CV) approach (described in Chapter 3, Section 3.2.2).

Model discrimination performance will be assessed on the test and external validation sets using the AUROC, the proportion of variance [159] and the discrimination slope [160]. Model calibration will be assessed using the Hosmer-Lemeshow test [161]. Models will be ranked according to their parsimony based on the Bayesian Information Criterion (BIC) statistic [162].

9.7 Discussion

Recently published ORN prediction studies [114, 115] have concluded that their results are limited by small datasets. Other recent ORN studies [63] have focused on a particular sub-group of ORN patients, which does not facilitate extrapolation of their results to other centres. Multi-centre collaborations can enable more generalisable and statistically robust toxicity prediction models. Unlike clinical trials, multi-centre studies using real world data benefit from a larger patient diversity. However, the data collection process may be more challenging due to ethical or consent related issues, especially in late toxicities such as mandibular ORN.

The methodology described in the study protocol, especially the data modelling methods, is an initial suggestion that was accepted by all participating centres with the acknowledgement that further or alternative methods might be explored depending on the characteristics of the final dataset available.

The diversity of expertise in the members involved in the study was in turn the cause of delayed consensus in several aspects of the protocol. Initiating the legal process for the DTAs at an early stage while data was still being collected can reduce delays to the start of the data analysis. However, a more standardised and streamlined data sharing process to facilitate multi-centre collaborations would further limit unnecessary delays and costly resources and promote more robust clinical research studies, thus resulting in improved quality of patient care.

Future work will include DL-based NTCP models for ORN using the PREDMORN study dataset. Prospective data collection (e.g., in the form of a clinical trial) may also be considered as a continuation of this study in the future. Collaborations have already been

initiated with Newcastle upon Tyne Hospitals NHS Foundation Trust and the Clatterbridge Cancer Centre NHS Foundation Trust for potential external validation of the results and conclusions that will be obtained from the PREDMORN study.

Chapter 10

Conclusion

This chapter contains the final discussion and conclusions of this thesis. First, it summarises the novel scientific contributions of this thesis and their clinical impact. Second, it discusses the limitations of this work and possible future directions to address these limitations. Finally, overall conclusions are drawn.

10.1 Summary

Existing prediction models for mandibular ORN rely solely on DVH data as a surrogate for the dosimetric risk factors. Clinically, population-based generic dose-volume constraints are used during the treatment planning process to limit the radiation dose received by the mandible. The bone composition and vascularisation variations within the mandible result in a heterogeneous radiobiological response, with some regions more prone to ORN development than others. Moreover, this radiobiological response also varies between patients. As opposed to a DVH, radiation dose distribution maps preserve the spatial dose information that can be used to capture these radiobiological and anatomical heterogeneities when predicting radiation-induced toxicities. The use of spatial dose metrics in NTCP modelling has increasingly gained research interest over recent years.

The primary aim of the work presented in this thesis was exploring how ML and DL methods can contribute to personalised and explainable NTCP modelling of mandibular ORN using spatial dose information. The major contributions of this thesis are summarised below:

- The first major contribution of this thesis was the use of ML methods in the context of mandibular ORN incidence prediction.
- The second major contribution of this thesis was the implementation of a DL-based pipeline to predict mandibular ORN incidence. A 3D deep CNN was trained to automatically obtain spatial dose information from radiation dose distribution maps that was used to classify ORN vs. no ORN, with comparable predictive performance to more traditional DVH-based ML approaches.
- The third major contribution of this thesis was the implementation of multimodality DL fusion strategies for combining the image-based spatial dose information with clinical and demographic variables into a more comprehensive ORN NTCP model.
- The fourth major contribution of this thesis was the implementation of interpretability methods with in-depth quantitative analysis as a first step towards gaining trust for the clinical implementation of a DL-based ORN prediction model.
- Finally, the fifth major contribution of this thesis was the design and development of the multi-institutional PREDMORN study to develop, train and validate robust and generalisable NTCP models for mandibular ORN using the largest ORN dataset worldwide.

10.2 Current limitations and future directions

This section discusses the main limitations of the work presented in this thesis and possible solutions that could be explored in future work. It also analyses the assumptions made and how these may affect the proposed methods.

Radiation dose. A large amount of the work included in this thesis is based on radiation dose information extracted retrospectively from the clinical treatment planning systems. Below are some of the limitations related to this process:

- The radiation dose distributions were obtained from the clinical radiotherapy treatment plans, which are a simulation of the dose distribution based on a pre-radiotherapy CT scan rather than the actual dose delivered during treatment. The actual absorbed dose distribution during treatment was not available for the patients included in this study. This is a common limitation in existing retrospective studies. Future work should assess the differences between planned and delivered doses and the resulting uncertainties in the prediction of ORN.

- Head and neck cancer patients may experience anatomical changes due to weight loss over the radiotherapy treatment course. These often result in the need to fit a new immobilisation mask and to replan the radiation distribution accordingly. In these cases, a plan and a replan will be available, each with a fraction of the total number of treatment sessions delivered. The clinical approach to obtaining the overall planned dose distribution is to perform a weighted sum of both plans based on the number of treatment fractions they were used for. This approach, however, was not possible in the Monaco TPS as the summed plan could not be exported. Thus, the plan with the largest amount of delivered fractions was considered as delivered for the entire treatment course. It was considered that excluding these patients from the study would have a more detrimental effect than the dosimetric uncertainties introduced by this approximation. The department is currently transferring all the patients planned with the Monaco TPS to the Eclipse TPS, which will enable correctly summing the plan and replan dose distributions in future studies.

- Chapter 2 (Section 2.2.3) describes the conversion from $D_{w,m}$ to $D_{m,m}$ for some of the RT plans created with the Monaco TPS using the stopping power ratios of water and the different tissue types. However, this conversion may introduce uncertainties as the tissue densities might not be well defined in the TPS [163]. We are currently participating in a study led by ICO that aims to assess how the dosimetric correlations with ORN are affected by the differences between the two dose calculation modes, $D_{w,m}$ and $D_{m,m}$.

Mandible segmentation. In some oral cavity cancer cases, part of the mandible is removed during surgery, a procedure known as mandibulectomy, sometimes with subsequent reconstruction typically using flap bone. In this work only the mandible bone was included during manual segmentation of the mandible structure and any external bone was excluded as it strictly considered ORN developing on the mandible bone only. However, it would be of high clinical interest to investigate how the inclusion of the reconstructed bone as part of the mandible structure could potentially modify our findings.

Mandible dose map. The processing steps for the image data described in Chapter 4 (Section 4.3.4) include the masking of the dose map to obtain a mandible dose map. This approach, largely inspired by Dean et al. [128] and Ibragimov et al. [102], was followed to aid the network to focus on the mandible structure. Reber et al. [115] did not mask the dose map but used the mandible structure to crop the dose map thus minimising its size. Further work will include an assessment of whether the exclusion of radiation dose and

anatomical information surrounding the mandible has a significant effect on the prediction of ORN with DL methods.

ORN severity classification. Mandible ORN is a rare toxicity, and the data sets available are naturally small. Low patient numbers make it difficult to attempt a multiclass prediction task where not only incidence but also ORN severity is predicted. Although the morbidity caused by ORN (at any grade) is such that the prediction of its incidence alone would already be an important clinical decision-support contribution, future work will aim at increasing the size of the study cohort to allow for ORN severity prediction.

Multimodality fusion. Chapter 7 (Section 7.2) described how dose maps and clinical variables were combined using early and late fusion strategies, which are the recommended types for small datasets [133] but may be outperformed by joint fusion strategies. Future work will explore the development of more complex joint strategies with a larger dataset.

ORN laterality analysis. Chapter 8 (Section 8.2) described the laterality association analysis between high attention regions, ORN regions and high dose regions. While the laterality of the high attention and high dose region was assessed by dividing the corresponding maps into two equal parts, this was not possible with the ORN region because the ORN region segmentation was not available for all subjects. The laterality of the ORN region was obtained from the clinical notes. Although common space registration was applied to all cases, it is possible that in some cases the mandible was not perfectly central. While this did not introduce any errors when comparing the laterality of the high dose and high attention regions, this is not necessarily the case with respect to the ORN region. Future work will aim at obtaining the manual segmentation of all the ORN regions on the planning CT image so that the laterality of the ORN region can be determined in the same way as for the high dose and high attention regions.

External validation. The work presented in this thesis was based on a single-institution dataset. Although a number of methods were applied to enhance its generalisability (nested cross-validation, dropout, etc.), external validation of the methods and results is essential. A collaboration established with the Odense University Hospital in Denmark has allowed the transfer of the DICOM files for their cohort. I am currently processing these files and the results of the external validation of my work with their data will be published at a later stage.

Multi-toxicity modelling. As discussed in Chapter 2 (Section 2.1.4), HNC toxicities are often associated. Although this thesis has focused on mandibular ORN, planned future work will investigate the associations between ORN and other HNC toxicities, with the ultimate aim of developing a DL-based HNC multi-toxicity prediction model.

Model calibration. In the experiments included in this thesis, a binary classification of ORN vs. no ORN has been explored based on the output of models. Without model calibration, however, the output of the models cannot be interpreted as true probabilities. Thus, for the clinical implementation of the methods explored, the models would need to be calibrated in order to calculate the actual toxicity risk at an individual patient level.

10.3 Conclusions

This thesis has explored the potential of DL methods in the prediction of mandibular ORN incidence. Previous ORN prediction models have used DVH data as the dosimetric risk factor for ORN. However, DVH data excludes any spatial dose information from the clinical RT treatment plan. In this thesis, 3D radiation dose distribution maps have been used instead of DVH data, which allow the inclusion of anatomical and radiobiological heterogeneities within the mandible. In addition, multimodality fusion strategies have been explored to combine the dose maps with other clinical risk factors. Finally, the DL-based predictions have been analysed with interpretability methods that confirmed spatial dose associations with the incidence of mandibular ORN. The promising results reported in this work might stimulate further work on the use of DL methods for toxicity prediction in head and neck cancer as a comprehensive clinical decision-support tool.

References

[1] L.V. van Dijk, A.A. Abusaif, J. Rigert, M.A. Naser, K.A. Hutcheson, S.Y. Lai, C.D. Fuller, and A.S.R. Mohamed. Normal tissue complication probability (ntcp) prediction model for osteoradionecrosis of the mandible in patients with head and neck cancer after radiation therapy: Large-scale observational cohort. *International Journal of Radiation Oncology*Biology*Physics*, 111:549–558, 10 2021.

[2] Cancer research uk. www.cancerresearchuk.org.

[3] M.D. Mody, J.W. Rocco, S.S. Yom, R.I. Haddad, and N.F. Saba. Head and neck cancer. *The Lancet*, 398:2289–2299, 12 2021.

[4] A.K. Dhull, R. Atri, R. Dhankhar, A.K. Chauhan, and V. Kaushal. Major risk factors in head and neck cancer: A retrospective analysis of 12-year experiences. *World Journal of Oncology*, 9:80–84, 2018.

[5] D.E. Johnson, B. Burtness, C.R. Leemans, V.W.Y. Lui, J.E. Bauman, and J.R. Grandis. Head and neck squamous cell carcinoma. *Nature Reviews Disease Primers*, 6:92, 11 2020.

[6] American academy of otolaryngology – head and neck surgery foundation. Quick reference guide to tnm staging in head and neck cancer and neck dissection classification (fourth edition). 2014.

[7] W. Lydiatt, B. O'Sullivan, and S. Patel. Major changes in head and neck staging for 2018. *American Society of Clinical Oncology Educational Book*, pages 505–514, 5 2018.

[8] M.B. Amin, F.L. Greene, S.B. Edge, C.C. Compton, J.E. Gershenwald, R.K. Brookland, L. Meyer, D.M. Gress, D.R. Byrd, and D.P. Winchester. The eight edition ajcc cancer staging manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin*, 67:93–99, 3 2017.

[9] H.P. van der Laan, L. Van den Bosch, E. Schuit, R.J.H.M. Steenbakkers, A. van der Schaaf, and J.A. Langendijk. Impact of radiation-induced toxicities on quality of life of patients treated for head and neck cancer. *Radiotherapy and Oncology*, 160:47–53, 7 2021.

[10] L. van den Bosch, A. van der Schaaf, H.P. van der Laan, F.J.P. Hoebers, O.B. Wijers, J.G.M. van den Hoek, K.G.M. Moons, J.B. Reitsma, R.J.H.M. Steenbakkers, E. Schuit, and J.A. Langendijk. Comprehensive toxicity risk profiling in radiation therapy for head and neck cancer: A new concept for individually optimised treatment. *Radiotherapy and Oncology*, 157:147–154, 4 2021.

[11] F. Siddiqui and B. Movsas. Management of radiation toxicity in head and neck cancers. *Seminars in Radiation Oncology*, 27:340–349, 10 2017.

[12] S. Nabil and N. Samman. Incidence and prevention of osteoradionecrosis after dental extraction in irradiated patients: a systematic review. *International Journal of Oral and Maxillofacial Surgery*, 40:229–243, 3 2011.

[13] A.T.T. Wong, S.Y. Lai, G.B. Gunn, B.M. Beadle, C.D. Fuller, M.P. Barrow, T.M. Hofstede, M.S. Chambers, E.M. Sturgis, A.S.R. Mohamed, J.S. Lewin, and K.A. Hutcheson. Symptom burden and dysphagia associated with osteoradionecrosis in long-term oropharynx cancer survivors: A cohort analysis. *Oral Oncology*, 66:75–80, 3 2017.

[14] E.B. Podgorsak. *Radiation Oncology Physics: A Handbook for Teachers and Students*. International Atomic Energy Agency (IAEA), 2005.

[15] X. Wang and A. Eisbruch. Imrt for head and neck cancer: reducing xerostomia and dysphagia. *Journal of Radiation Research*, 57:i69–i75, 8 2016.

[16] B. O'Sullivan, R.B. Rumble, and P. Warde. Intensity-modulated radiotherapy in the treatment of head and neck cancer. *Clinical Oncology*, 24:474–487, 9 2012.

[17] D.I. Rosenthal, M.S. Chambers, C.D. Fuller, N.C.S. Rebueno, J. Garcia, M.S. Kies, W.H. Morrison, K.K. Ang, and A.S. Garden. Beam path toxicities to non-target structures during intensity-modulated radiation therapy for head and neck cancer. *International Journal of Radiation Oncology*Biology*Physics*, 72:747–755, 11 2008.

[18] J. van der Veen and S. Nuyts. Can intensity-modulated-radiotherapy reduce toxicity in head and neck squamous cell carcinoma? *Cancers*, 9:135, 10 2017.

[19] T. De Maesschalck, N. Dulguerov, F. Caparrotti, P. Scolozzi, C. Picardi, N. Mach, N. Koutsouvelis, and P. Dulguerov. Comparison of the incidence of osteoradionecrosis with conventional radiotherapy and intensity-modulated radiotherapy. *Head Neck*, 38:1695–1702, 11 2016.

[20] M.C. Joiner and A. van der Kogel. *Basic Clinical Radiobiology*. CRC Press, 4th edition, 2009.

[21] ICRU. The international commission on radiation units and measurements (icru) report 83. *Journal of the ICRU. Oxford University Press*, 10, 2010.

[22] The Royal College of Radiologists (RCR). Head and neck cancer rcr consensus statements. 2 2022.

[23] C.L. Brouwer, R.J. Steenbakkers, J. Bourhis, W. Budach, C. Grau, V. Grégoire, M. van Herk, A. Lee, P. Maingon, C. Nutting, B. O'Sullivan, S.V. Porceddu, D.I. Rosenthal, N.M. Sijtsema, and J.A. Langendijk. Ct-based delineation of organs at risk in the head and neck region: Dahanca, eortc, gortec, hknpcsg, ncic ctg, ncri, nrg oncology and trog consensus guidelines. *Radiotherapy and Oncology*, 117:83–90, 10 2015.

[24] T.D. DenOtter and J. Schubert. Hounsfield unit. *https://www.ncbi.nlm.nih.gov/books/NBK547721/*, 2022.

[25] F. De Martino, S. Clemente, C. Graeff, G. Palma, and L. Cella. Dose calculation algorithms for external radiation therapy: An overview for practitioners. *Applied Sciences*, 11:6806, 7 2021.

[26] N. Dogan, J.V. Siebers, and P.J. Keall. Clinical comparison of head and neck and prostate imrt plans using absorbed dose to medium and absorbed dose to water. *Physics in Medicine and Biology*, 51:4967–4980, 10 2006.

[27] M.A Ebert, S.L. Gulliford, O. Acosta, R. de Crevoisier, T. McNutt, W.D. Heemsbergen, M. Witte, G. Palma, T. Rancati, and C. Fiorino. Spatial descriptions of radiotherapy dose: normal tissue complication models and statistical associations. *Physics in Medicine Biology*, 66:12TR01, 6 2021.

[28] D.S. Chang, F.D. Lasley, I.J. Das, M.S. Mendonca, and J.R. Dynlacht. *Basic Radiotherapy Physics and Biology. Cham: Springer, 2014*. Springer, 2014.

[29] J.T. Lyman. Complication probability as assessed from dose-volume histograms. *Radiation research. Supplement*, 8:S13–9, 1985.

[30] B. Emami. Tolerance of normal tissue to therapeutic radiation. 1:35, 2013.

[31] P. Lambin, R.G.P.M. van Stiphout, M.H.W. Starmans, E. Rios-Velazquez, G. Nalbantov, H.J.W.L. Aerts, E. Roelofs, W. van Elmpt, P.C. Boutros, P. Granone, V. Valentini, A.C. Begg, D. De Ruysscher, and A. Dekker. Predicting outcomes in radiation oncology—multifactorial decision support systems. *Nature Reviews Clinical Oncology*, 10:27–40, 1 2013.

[32] N.P. Brodin, R. Kabarriti, M.K. Garg, C. Guha, and W.A. Tomé. Systematic review of normal tissue complication models relevant to standard fractionation radiation therapy of the head and neck region published after the quantec reports. *International Journal of Radiation Oncology Biology Physics*, 100:391–407, 2 2018.

[33] R.G.J. Kierkels, E.W. Korevaar, R.J.H.M. Steenbakkers, T. Janssen, A.A. van't Veld, J.A. Langendijk, C. Schilstra, and A. van der Schaaf. Direct use of multivariable normal tissue complication probability models in treatment plan optimisation for individualised head and neck cancer radiotherapy produces clinically acceptable treatment plans. *Radiotherapy and Oncology*, 112:430–436, 9 2014.

[34] H.S. Gabryś, F. Buettner, F. Sterzing, H. Hauswald, and M. Bangert. Design and selection of machine learning methods using radiomics and dosiomics for normal tissue complication probability modeling of xerostomia. *Frontiers in Oncology*, 8, 3 2018.

[35] S.M. Bentzen, L.S. Constine, J.O. Deasy, A. Eisbruch, A. Jackson, L.B. Marks, R.K. Ten Haken, and E.D. Yorke. Quantitative analyses of normal tissue effects in the clinic (quantec): An introduction to the scientific issues. *International Journal of Radiation Oncology Biology Physics*, 76, 3 2010.

[36] S. Stieb, A. Lee, L.V. van Dijk, S. Frank, C.D. Fuller, and P. Blanchard. Ntcp modeling of late effects for head and neck cancer: A systematic review. *International Journal of Particle Therapy*, 8:95–107, 6 2021.

[37] L.B. Marks, E.D. Yorke, A. Jackson, R.K. Ten Haken, L.S. Constine, A. Eisbruch, S.M. Bentzen, J. Nam, and J.O. Deasy. Use of normal tissue complication probability models in the clinic. *International Journal of Radiation Oncology*Biology*Physics*, 76:S10–S19, 3 2010.

[38] I. El Naqa. *Machine and Deep Learning in Oncology, Medical Physics and Radiology*. Springer Nature Switzerland AG, 2022.

[39] J. W. Hopewell and K-R. Trott. Volume effects in radiobiology as applied to radiotherapy. *Radiotherapy and Oncology*, 56:283–288, 9 2000.

[40] A.J. Frankart, M.J. Frankart, B. Cervenka, A.L. Tang, D.G. Krishnan, and V. Takiar. Osteoradionecrosis: Exposing the evidence not the bone. *International Journal of Radiation Oncology Biology Physics*, 109:1206–1218, 4 2021.

[41] F. De Felice, C. Thomas, V. Patel, S. Connor, A. Michaelidou, C. Sproat, J. Kwok, M. Burke, D. Reilly, M. McGurk, R. Simo, A. Lyons, R. Oakley, J.P. Jeannon, M. Lei, and T. Guerrero Urbano. Osteoradionecrosis following treatment for head and neck cancer and the effect of radiotherapy dosimetry: the guy's and st thomas' head and neck cancer unit experience. *Oral Surg Oral Med Oral Pathol Oral Radiol*, 122:28–34, 7 2016.

[42] D. Patel, S. Haria, and V. Patel. Oropharyngeal cancer and osteoradionecrosis in a novel radiation era: a single institution analysis. *Oral Surgery*, 14:113–121, 5 2021.

[43] S. Habib, I. Sassoon, I. Thompson, and V. Patel. Risk factors associated with osteoradionecrosis. *Oral Surgery*, 14:227–235, 8 2021.

[44] V. Patel, M. Fenlon, L. Di Silvio, and M. McGurk. Osteoradionecrosis in the current era of radiation treatment. *Dental Update*, 49:64–67, 1 2022.

[45] A. Chronopoulos, T. Zarra, M. Ehrenfeld, and S. Otto. Osteoradionecrosis of the jaws: definition, epidemiology, staging and clinical and radiological findings. a concise review. *International Dental Journal*, 68:22–30, 2 2018.

[46] R.E. Marx. Osteoradionecrosis: A new concept of its pathophysiology. *Journal of Oral and Maxillofacial Surgery*, 41:283–288, 5 1983.

[47] S. Delanian and J-L. Lefaix. The radiation-induced fibroatrophic process: therapeutic perspective via the antioxidant pathway. *Radiotherapy and Oncology*, 73:119–131, 11 2004.

[48] V. Patel, L. Ormondroyd, A. Lyons, and M. McGurk. The financial burden for the surgical management of osteoradionecrosis. *British Dental Journal*, 222:177–180, 2 2017.

[49] National Cancer Institute (NCI). U.S. Department of Health and Human Services. Common terminology criteria for adverse events (ctcae) v5.0. 2017.

[50] K.I. Notani, Y. Yamazaki, S. Moriya, N. Sakakibara, H. Nakamura, M. Watanabe, and H. Fukuda. Osteoradionecrosis of the mandible - factors influencing severity. *Asian Journal of Oral and Maxillofacial Surgery*, 14:5–9, 2002.

[51] T. Reuther, T. Schuster, U. Mende, and A. Kübler. Osteoradionecrosis of the jaws as a side effect of radiotherapy of head and neck tumour patients—a report of a thirty year retrospective review. *International Journal of Oral and Maxillofacial Surgery*, 32:289–295, 6 2003.

[52] A.T. Suryawanshi, V. Pawar, M. Singh, R.S. Dolas, R. Khindria, and S.N.S. Kumar. Maxillofacial osteoradionecrosis. *Journal of Dental Research and Review*, 1:42, 2014.

[53] F. De Felice, V. Tombolini, D. Musio, and A. Polimeni. Radiation therapy and mandibular osteoradionecrosis: State of the art. *Current Oncology Reports*, 22:89, 9 2020.

[54] D.R. Gomez, C.L. Estilo, S.L. Wolden, M.J. Zelefsky, D.H. Kraus, R.J. Wong, A.R. Shaha, J.P. Shah, J.G. Mechalakos, and N.Y. Lee. Correlation of osteoradionecrosis and dental events with dosimetric parameters in intensity-modulated radiation therapy for head-and-neck cancer. *International Journal of Radiation Oncology*Biology*Physics*, 81:e207–e213, 11 2011.

[55] C.J. Tsai, T.M. Hofstede, E.M. Sturgis, A.S. Garden, M.E. Lindberg, Q. Wei, S.L. Tucker, and L. Dong. Osteoradionecrosis and radiation dose to the mandible in patients with oropharyngeal cancer. *International Journal of Radiation Oncology*Biology*Physics*, 85:415–420, 2 2013.

[56] J.D. Raguse, J. Hossamo, I. Tinhofer, B. Hoffmeister, V. Budach, B. Jamil, K. Jöhrens, N. Thieme, C. Doll, S. Nahles, S.T. Hartwig, and C. Stromberger. Patient and treatment-related risk factors for osteoradionecrosis of the jaw in patients with head and neck cancer. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, 121:215–221.e1, 3 2016.

[57] F. Caparrotti, S.H. Huang, L. Lu, S.V. Bratman, J. Ringash, A. Bayley, J. Cho, M. Giuliani, J. Kim, J. Waldron, A. Hansen, L. Tong, W. Xu, B. O'Sullivan, R. Wood, D. Goldstein, and A. Hope. Osteoradionecrosis of the mandible in patients with oropharyngeal carcinoma treated with intensity-modulated radiotherapy. *Cancer*, 123:3691–3700, 10 2017.

[58] A.A. Owosho, C.J. Tsai, R.S. Lee, H. Freymiller, A. Kadempour, S. Varthis, A.Z. Sax, E.B. Rosen, S.K. Yom, J. Randazzo, E. Drill, E. Riedel, S. Patel, N.Y. Lee, J.M. Huryn, and C.L. Estilo. The prevalence and risk factors associated with osteoradionecrosis of the jaw in oral and oropharyngeal cancer patients treated with intensity-modulated radiation therapy (imrt): The memorial sloan kettering cancer center experience. *Oral Oncology*, 64:44–51, 1 2017.

[59] A.S.R. Mohamed, B.P. Hobbs, K.A. Hutcheson, M.S. Murri, N. Garg, J. Song, G.B. Gunn, V. Sandulache, B.M. Beadle, J. Phan, W.H. Morrison, S.J. Frank, P. Blanchard, A.S. Garden, H. El-Halawani, M. Kamal, M.S. Chambers, J.S. Lewin, R. Ferrarotto, X.R. Zhu, X. Zhang, T.M. Hofstede, R.C. Cardoso, A.M. Gillenwater, E.M. Sturgis, R.S. Weber, D.I. Rosenthal, C.D. Fuller, and S.Y. Lai. Dose-volume correlates of mandibular osteoradionecrosis in oropharynx cancer patients receiving intensity-modulated radiotherapy: Results from a case-matched comparison. *Radiotherapy and Oncology*, 124:232–239, 8 2017.

[60] D.H. Moon, S.H. Moon, Kyle Wang, Mark C. Weissler, T.G. Hackman, A.M. Zanation, B.D. Thorp, S.N. Patel, J.P. Zevallos, L.B. Marks, and B.S. Chera. Incidence of, and risk factors for, mandibular osteoradionecrosis in patients with oral cavity and oropharynx cancers. *Oral Oncology*, 72:98–103, 9 2017.

[61] S. Aarup-Kristensen, C.R. Hansen, L. Forner, C. Brink, J.G. Eriksen, and J. Johansen. Osteoradionecrosis of the mandible after radiotherapy for head and neck cancer: risk factors and dose-volume correlations. *Acta Oncologica*, 58:1373–1377, 10 2019.

[62] H. Kubota, D. Miyawaki, N. Mukumoto, T. Ishihara, M. Matsumura, T. Hasegawa, M. Akashi, N. Kiyota, H. Shinomiya, M. Teshima, K-I. Nibu, and R. Sasaki. Risk factors for osteoradionecrosis of the jaw in patients with head and neck squamous cell carcinoma. *Radiation Oncology*, 16:1, 12 2021.

[63] M.M. Möring, H. Mast, E.B. Wolvius, G.M. Verduijn, S.F. Petit, N.D. Sijtsema, B.P. Jonker, R.A. Nout, and W.D. Heemsbergen. Osteoradionecrosis after postoperative radiotherapy for oral cavity cancer: A retrospective cohort study. *Oral Oncology*, 133:106056, 10 2022.

[64] V. Patel, L. Humbert-Vidan, C. Thomas, I. Sassoon, M. McGurk, M. Fenlon, and T. Guerrero Urbano. Radiotherapy quadrant doses in oropharyngeal cancer treated with intensity modulated radiotherapy. identifying dental regions at risk of osteoradionecrosis from post-radiotherapy events provides invaluable information. 11, 2020.

[65] L. Humbert-Vidan, V. Patel, R.H. Begum, M. McGovern, D. Eaton, A. Kong, I. Petkar, M. Reis Ferreira, M. Lei, A.P. King, and T. Guerrero Urbano. Ph-0387 mandible osteoradionecrosis: a dosimetric study, radiotherapy and oncology. *Radiotherapy and Oncology*, Volume 161:285–286, 2021.

[66] M.A. Ben-David, M. Diamante, J.D. Radawski, K.A. Vineberg, C. Stroup, C.A. Murdoch-Kinch, S.R. Zwetchkenbaum, and A. Eisbruch. Lack of osteoradionecrosis of the mandible after intensity-modulated radiotherapy for head and neck cancer: Likely contributions of both dental care and improved dose distributions. *International Journal of Radiation Oncology Biology Physics*, 68:396–402, 6 2007.

[67] L. Humbert-Vidan, V. Patel, I. Oksuz, A.P. King, and T. Guerrero Urbano. Comparison of machine learning methods for prediction of osteoradionecrosis incidence in patients with head and neck cancer. *The British Journal of Radiology*, 94:20200026, 4 2021.

[68] S.N. Rogers, J.J. D'Souza, D. Lowe, and A. Kanatas. Longitudinal evaluation of health-related quality of life after osteoradionecrosis of the mandible. *British Journal of Oral and Maxillofacial Surgery*, 53:854–857, 11 2015.

[69] T.M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[70] L.P. Kaelbling, M.L. Littman, and A.W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 5 1996.

[71] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, 2017.

[72] H-H. Tseng, Y. Luo, R.K. Ten Haken, and I. El Naqa. The role of machine learning in knowledge-based response-adapted radiotherapy. *Frontiers in Oncology*, 8, 7 2018.

[73] K. Harrison, H. Pullen, C. Welsh, O. Oktay, J. Alvarez-Valle, and R. Jena. Machine learning for auto-segmentation in radiotherapy planning. *Clinical Oncology*, 34:74–88, 2 2022.

[74] L.J. Isaksson, M. Pepa, M. Zaffaroni, G. Marvaso, D. Alterio, S. Volpe, G. Corrao, M. Augugliaro, A. Starzyńska, M.C. Leonardi, R. Orecchia, and B.A. Jereczek-Fossa. Machine learning-based models for prediction of toxicity outcomes in radiotherapy. *Frontiers in Oncology*, 10, 6 2020.

[75] M. Maalouf. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies*, 3:281, 2011.

[76] C. Cortes, V. Vapnik, and L. Saitta. Support-vector networks editor. *Machine Leaming*, 20:273–297, 1995.

[77] P Kubben, M Dumontier, and A Dekker, editors. *Fundamentals of Clinical Data Science*. Springer, 2019.

[78] C. Sammut and G.I. Webb. *Encyclopedia of Machine Learning*. Springer, 2011.

[79] L. Breiman. Random forests. 45:5–32, 2001.

[80] T. Chengsheng, L. Huacheng, and X. Bing. Adaboost typical algorithm and its application research. volume 139. EDP Sciences, 12 2017.

[81] F. Rosenblatt. *Principles of neurodynamics; perceptrons and the theory of brain mechanisms*. Spartan Books, 1962.

[82] S. Haykin. *Neural networks and learning machines*. Prentice Hall, 2nd edition, 2008.

[83] I.H. Sarker. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2:420, 11 2021.

[84] J. Lederer. Activation functions in artificial neural networks: A systematic overview. 1 2021.

[85] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 5 2015.

[86] F. Chollet. *Deep Learning with Python. Manning Publications*. 2018.

[87] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Proceedings of Advances in Neural Information Processing Systems (NIPS) 30*, 2017.

[88] M.J. Willemink, H.R. Roth, and V. Sandfort. Toward foundational deep learning models for medical imaging in the new era of transformer networks. *Radiology: Artificial Intelligence*, 4, 11 2022.

[89] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 12 1989.

[90] I. Vasilev, D. Slater, G. Spacagna, P. Roelants, and V. Zocca. *Python Deep Learning: Exploring Deep Learning Techniques and Neural Network Architectures with Pytorch, Keras, and Tensorflow*. Packt Publishing, 2019.

[91] V.H. Phung and E.J. Rhee. A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, 9:4500, 10 2019.

[92] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. 9 2014.

[93] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger. Densely connected convolutional networks. 2017.

[94] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. 7 2017.

[95] Y. Yang, L. Zhang, M. Du, J. Bo, H. Liu, L. Ren, X. Li, and M.J. Deen. A comparative analysis of eleven neural networks architectures for small datasets of lung images of covid-19 patients toward improved clinical decisions. *Computers in Biology and Medicine*, 139:104887, 12 2021.

[96] G.S. Collins, J.B. Reitsma, D.G. Altman, and K. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *BMC Medicine*, 13:1, 2015.

[97] G.C. Cawley and N.L.C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

[98] S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 12 2006.

[99] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 7.

[100] S. Monti, G. Palma, V. D'Avino, M. Gerardi, G. Marvaso, D. Ciardo, R. Pacelli, B.A. Jereczek-Fossa, D. Alterio, and L. Cella. Voxel-based analysis unveils regional dose differences associated with radiation-induced morbidity in head and neck cancer patients. *Scientific Reports*, 7:7220, 8 2017.

[101] J.A. Dean, K.H. Wong, L.C. Welsh, A.B. Jones, U. Schick, K.L. Newbold, S.A. Bhide, K.J. Harrington, C.M. Nutting, and S.L. Gulliford. 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. *Radiotherapy and Oncology*, 120:21–27, 7 2016.

[102] B. Ibragimov, D. Toesca, D. Chang, Y. Yuan, A. Koong, and L. Xing. Development of deep neural network for individualized hepatobiliary toxicity prediction after liver sbrt. *Medical Physics*, 45:4763–4774, 10 2018.

[103] K. Men, H. Geng, H. Zhong, Y. Fan, A. Lin, and Y. Xiao. A deep learning model for predicting xerostomia due to radiation therapy for head and neck squamous cell carcinoma in the rtog 0522 clinical trial. *International Journal of Radiation Oncology*Biology*Physics*, 105:440–447, 10 2019.

[104] F. Buettner, A.B. Miah, S.L. Gulliford, E. Hall, K.J. Harrington, S. Webb, M. Partridge, and C.M. Nutting. Novel approaches to improve the therapeutic index of head and neck radiotherapy: An analysis of data from the parsport randomised phase iii trial. *Radiotherapy and Oncology*, 103:82–87, 4 2012.

[105] R. Munbodh, A. Jackson, J. Bauer, C.R. Schmidtlein, and M.J. Zelefsky. Dosimetric and anatomic indicators of late rectal toxicity after high-dose intensity modulated radiation therapy for prostate cancer. *Medical Physics*, 35:2137–2150, 4 2008.

[106] W. Beasley, M. Thor, A. McWilliam, A. Green, R. Mackay, N. Slevin, C. Olsson, N. Pettersson, C. Finizia, C. Estilo, N. Riaz, N.Y. Lee, J.O. Deasy, and M. van Herk. Image-based data mining to probe dosimetric correlates of radiation-induced trismus. *International Journal of Radiation Oncology Biology Physics*, 102:1330–1338, 11 2018.

[107] J.A. Dean, K. Wong, H. Gay, L. Welsh, A.B. Jones, U. Schick, J.H. Oh, A. Apte, K. Newbold, S. Bhide, K. Harrington, J. Deasy, C. Nutting, and S. Gulliford. Incorporating spatial dose metrics in machine learning-based normal tissue complication probability (ntcp) models of severe acute dysphagia resulting from head and neck radiotherapy. *Clinical and Translational Radiation Oncology*, 8:27–39, 1 2018.

[108] W. Jiang, P. Lakshminarayanan, X. Hui, P. Han, Z. Cheng, M. Bowers, I. Shpitser, S. Siddiqui, R.H. Taylor, H. Quon, and T. McNutt. Machine learning methods uncover radiomorphologic dose patterns in salivary glands that predict xerostomia in patients with head and neck cancer. *Advances in Radiation Oncology*, 4:401–412, 4 2019.

[109] M.L. Welch, C. McIntosh, A. McNiven, S.H. Huang, B-B. Zhang, L. Wee, A. Traverso, B. O'Sullivan, F. Hoebers, A. Dekker, and D.A. Jaffray. User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions. *Physica Medica*, 70:145–152, 2 2020.

[110] A.L. Appelt, B. Elhaminia, A. Gooya, A. Gilbert, and M. Nix. Deep learning for radiotherapy outcome prediction using dose data – a review. *Clinical Oncology*, 34:e87–e96, 2 2022.

[111] X. Zhen, J. Chen, Z. Zhong, B. Hrycushko, L. Zhou, S. Jiang, K. Albuquerque, and X. Gu. Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. *Physics in Medicine Biology*, 62:8246–8263, 10 2017.

[112] B. Ibragimov, D.A.S. Toesca, Y. Yuan, A.C. Koong, D.T. Chang, and L. Xing. Neural networks for deep radiotherapy dose analysis and prediction of liver sbrt outcomes. *IEEE Journal of Biomedical and Health Informatics*, 23:1821–1833, 9 2019.

[113] B. Ibragimov, D.A.S. Toesca, D.T. Chang, Y. Yuan, A.C. Koong, L. Xing, and I.R. Vogelius. Deep learning for identification of critical regions associated with toxicities after liver stereotactic body radiation therapy. *Medical Physics*, 47:3721–3731, 8 2020.

[114] L. Humbert-Vidan, V. Patel, R. Andlauer, A.P. King, and T. Guerrero Urbano. Prediction of mandibular orn incidence from 3d radiation dose distribution maps using deep learning. *Applications of Medical Artificial Intelligence, AMAI 2022. Lecture Notes in Computer Science*, 13540:49–58, 2022.

[115] B. Reber, L.V. van Dijk, B. Anderson, A.S.R. Mohamed, C.D. Fuller, S. Lai, and K. Brock. Comparison of machine-learning and deep-learning methods for the prediction of osteoradionecrosis resulting from head and neck cancer radiation therapy. *Advances in Radiation Oncology*, 8:101163, 7 2022.

[116] D. Huang, H. Bai, L. Wang, Y. Hou, L. Li, Y. Xia, Z.Y., W. Chen, L. Chang, and W. Li. The application and development of deep learning in radiotherapy: A systematic review. *Technology in Cancer Research Treatment*, 20:153303382110163, 1 2021.

[117] P. Meyer, V. Noblet, C. Mazzara, and A. Lallement. Survey on deep learning for radiotherapy. *Computers in Biology and Medicine*, 98:126–146, 7 2018.

[118] R.F. Thompson, G. Valdes, C.D. Fuller, C.M. Carpenter, O. Morin, S. Aneja, W.D. Lindsay, H.J.W.L. Aerts, B. Agrimson, C. Deville, S.A. Rosenthal, J.B. Yu, and C.R. Thomas. Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation? *Radiotherapy and Oncology*, 129:421–426, 12 2018.

[119] C.L. Brouwer, A.M. Dinkla, L. Vandewinckele, W. Crijns, M. Claessens, D. Verellen, and W. van Elmpt. Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. *Physics and Imaging in Radiation Oncology*, 16:144–148, 10 2020.

[120] D.G. Altman. Practical statistics for medical research. douglas g. altman, chapman and hall, london, 1991. no. of pages: 611. price: £32.00. *Statistics in Medicine*, 10:1635–1636, 10 1999.

[121] D. Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in Psychology*, 4, 2013.

[122] J. Cohen. Statistical power analysis for the behavioral sciences. 1988.

[123] Y. Sun, X-L. Yu, W. Luo, A.W.M. Lee, J.T.S. Wee, N. Lee, G-Q. Zhou, L-L. Tang, C-J. Tao, R Guo, Y-P. Mao, R. Zhang, Y. Guo, and J. Ma. Recommendation for a contouring method and atlas of organs at risk in nasopharyngeal carcinoma patients receiving intensity-modulated radiotherapy. *Radiotherapy and Oncology*, 110:390–397, 3 2014.

[124] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J.V. Miller, S. Pieper, and R. Kikinis. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, 30:1323–1341, 11 2012.

[125] P.A. Yushkevich, J. Piven, H.C. Hazlett, R.G. Smith, S. Ho, J.C. Gee, and G. Gerig. User-guided 3d active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, 31:1116–1128, 7 2006.

[126] I.J. Lee, W.S. Koom, C.G. Lee, Y.B. Kim, S.W. Yoo, K.C. Keum, G.E. Kim, E.C. Choi, and I.H. Cha. Risk factors and dose-effect relationship for mandibular osteoradionecrosis in oral and oropharyngeal cancer patients. *International Journal of Radiation Oncology Biology Physics*, 75:1084–1091, 11 2009.

[127] H-M. Ting, T-F. Lee, M-Y. Cho, P-J. Chao, C-M. Chang, L-C. Chen, and F-M. Fang. Comparison of neural network and logistic regression methods to predict xerostomia after radiotherapy. *World Academy of Science, Engineering and Technology International Journal of Biomedical and Biological Engineering*, 7, 2013.

[128] J.A. Dean, L.C. Welsh, K.J. Harrington, C.M. Nutting, and S.L. Gulliford. Predictive modelling of toxicity resulting from radiotherapy treatments of head and neck cancer. 12 2014.

[129] J.P. Reddy, W.D. Lindsay, C.G. Berlind, C.A. Ahern, A. Holmes, B.D. Smith, J. Phan, S.J. Frank, G.B. Gunn, D.I. Rosenthal, W.H. Morrison, A.S. Garden, G.M. Chronowski, S.J. Shah, L.L. Mayo, and C.D. Fuller. Applying a machine learning approach to predict acute radiation toxicities for head and neck cancer patients. *International Journal of Radiation Oncology*Biology*Physics*, 105:S69, 9 2019.

[130] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837, 9 1988.

[131] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J-C. Sanchez, and M. Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 12 2011.

[132] A.P. King T. Guerrero Urbano L. Humbert-Vidan, V. Patel. Letter to the editor regarding the paper entitled "comparison of machine learning and deep learning methods for the prediction of osteoradionecrosis resulting from head and neck cancer radiation therapy" by reben et al. (2022). *Accepted for publication in Advances in Radiation Oncology*, 2023.

[133] S-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M.P. Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3:136, 10 2020.

[134] D. Ramachandram and G.W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34:96–108, 11 2017.

[135] Z. Han, F. Yang, J. Huang, C. Zhang, and J. Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. 2022.

[136] Z. Xue and R. Marculescu. Dynamic multimodal fusion. *arXiv:2204.00102v1 [cs.CV]*, 2022.

[137] Y. Luo, H-H. Tseng, S. Cui, L. Wei, R.K. Ten Haken, and I. El Naqa. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR|Open*, 1, 11 2019.

[138] M. Temme. Algorithms and transparency in view of the new general data protection regulation. *European Data Protection Law Review*, 3:473–485, 2017.

[139] Z. Salahuddin, H.C. Woodruff, A. Chatterjee, and P. Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine*, 140:105111, 1 2022.

[140] C. Molnar. *Interpretable machine learning. A guide for making black box models explainable*. 2022.

[141] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. 10 2016.

[142] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. 2020.

[143] K.H. Zou, S.K. Warfield, A. Bharatha, C.M.C. Tempany, M.R. Kaus, S.J. Haker, W.M. Wells, F.A. Jolesz, and R. Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index1. *Academic Radiology*, 11:178–189, 2 2004.

[144] B. Elhaminia, A. Gilbert, J. Lilley, M. Abdar, A.F. Frangi, A. Scarsbrook, A.L. Appelt, and A. Gooya. Toxicity prediction in pelvic radiotherapy using multiple instance learning and cascaded attention layers. *IEEE Journal of Biomedical and Health Informatics*, 2023.

[145] B. Liang, Y. Tian, X. Chen, H. Yan, L. Yan, T. Zhang, Z. Zhou, L. Wang, and J. Dai. Prediction of radiation pneumonitis with dose distribution: A convolutional neural network (cnn) based model. *Frontiers in Oncology*, 9, 1 2020.

[146] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. 10 2018.

[147] L. Humbert-Vidan, C.R. Hansen, C.D. Fuller, S. Petit, A. Van der Schaaf, L.V. van Dijk, G.M. Verduijn, H. Langendijk, C. Muñoz-Montplet, W. Heemsbergen, M. Witjes, A.S.R. Mohamed, A.A. Khan, J. Marruecos Querol, I. Oliveras Cancio, V. Patel, A.P. King, J. Johansen, and T. Guerrero Urbano. Protocol letter: A multi-institutional retrospective case-control cohort investigating PREDiction models for Mandibular OsteoRadioNecrosis in head and neck cancer (PREDMORN). *Radiotherapy and Oncology*, 176:99–100, 11 2022.

[148] M.F. Fathalla, M.M.F. Fathalla, and World Health Organization. Regional Office for the Eastern Mediterranean. *A practical guide for health researchers*. World Health Organization, Regional Office for the Eastern Mediterranean, 2004.

[149] L. Humbert-Vidan, E. Blackmore, C.R. Hansen, C.D. Fuller, S. Petit, A. Van der Schaaf, L.V. van Dijk, G.M. Verduijn, H. Langendijk, C. Muñoz-Montplet, W. Heemsbergen, M. Witjes, A.S.R. Mohamed, A.A. Khan, J. Marruecos Querol, I. Oliveras Cancio, V. Patel, A.P. King, J. Johansen, and T. Guerrero Urbano. Challenges in international real world evidence research collaboration. the predmorn experience (e23-1669). 2023.

[150] R. Tibshirani. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Methodol*, 58:267–288, 1996.

[151] R. Johnson and T. Zhang. Learning nonlinear functions using regularized greedy forest. 9 2011.

[152] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[153] E. Goan and C. Fookes. Bayesian neural networks: An introduction and survey. pages 45–87, 2020.

[154] M.Z.I. Chowdhury and T.C. Turin. Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, 8, 2 2020.

[155] A. Maćkiewicz and W. Ratajczak. Principal components analysis (pca). *Computers Geosciences*, 19:303–342, 3 1993.

[156] C.R. Hansen, A. Bertelsen, R. Zukauskaite, L. Johnsen, U. Bernchou, D.I. Thwaites, J.G. Eriksen, J. Johansen, and C. Brink. Prediction of radiation-induced mucositis of hn cancer patients based on a large patient cohort. *Radiotherapy and Oncology*, 147:15–21, 6 2020.

[157] E.W. Steyerberg, A.J. Vickers, N.R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M.J. Pencina, and M.W. Kattan. Assessing the performance of prediction models. *Epidemiology*, 21:128–138, 1 2010.

[158] E.W. Steyerberg and F.E. Harrell. Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology*, 69:245–247, 1 2016.

[159] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

[160] M.J. Pencina, J.P. Fine, and R.B. D'Agostino. Discrimination slope and integrated discrimination improvement - properties, relationships and impact of calibration. *Statistics in Medicine*, 36:4482–4490, 12 2017.

[161] D.W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16:965–980, 5 1997.

[162] A.E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111, 1995.

[163] N. Reynaert, F. Crop, E. Sterpin, I. Kawrakow, and H. Palmans. On the conversion of dose to bone to dose to water in radiotherapy treatment planning systems. *Physics and Imaging in Radiation Oncology*, 5:26–30, 1 2018.

Appendix A. PREDMORN participating institutions and study collaborators

- **Guy's and St Thomas' NHS Foundation Trust (GSTT, London, UK) / King's College London (KCL, London, UK)** - *Leading institution*
  - Teresa Guerrero Urbano: Department of Clinical Oncology (GSTT)
  - Laia Humbert-Vidan: Department of Medical Physics (GSTT) / School of Cancer and Pharmaceutical Sciences (KCL)
  - Vinod Patel: Department of Oral Surgery (GSTT)
  - Andrew P. King: School of Engineering and Biomedical Sciences (KCL)
- **Odense University Hospital (OUH, DK)**
  - Jørgen Johansen
  - Christian Rønn Hansen
  - Abdul Ahad Khan
- **Catalan Institute of Oncology Girona (ICO, ES)**
  - Carles Muñoz Montplet
  - Jordi Marruecos Querol
  - Irene Oliveras Cancio
- **University Medical Centre Groningen (UMCG, NL)**
  - Hans Langendijk
  - Max Witjes
  - Arjen van der Schaaf
  - Lisanne V van Dijk

- **MD Anderson Cancer Centre (MDACC, USA)**
  - Clifton Dave Fuller
  - Abdallah Sherif Radwan Mohamed
- **Erasmus Medical Centre (EMC, NL)**
  - Gerda Verduijn
  - Steven Petit
  - Wilma Heemsbergen

Appendix B. Demographic and clinical variables included in the PREDMORN study

Variable	Units/format/options	Comments
Gender	male/female	
Date/year[1] of birth	dd/mm/yyyy	This is needed to calculate the age at the start of RT
RT start date/year	dd/mm/yyyy	
RT end date	dd/mm/yyyy	
Last follow-up date/year	dd/mm/yyyy	This is needed to calculate the follow-up time since the end of RT
Primary site group	• oral cavity (OCC) • oropharynx (OPC) • paranasal sinus/nasopharynx (PNS/NP) • larynx/hypopharynx • salivary gland • unknown primary	
TNM stage	• T: X, 0, 1, 2, 3, 4, 4a, 4b • N: 0, 1, 2, 2a, 2b, 2c, 3 • M: 0, 1, M	Use of TNM7 notation
HPV	• HPV+ • HPV- • not available	If available
Smoking status at start of RT	• Current: smoking at the start or within 3 months of the start of RT • Previous: stopped smoking at least more than 3 months prior to the start of RT • Never: the patient has never smoked	
Smoking amount at start of RT	• None • < 10 pack years • > 10 pack years • not available	If available
Alcohol status at start of RT	• Current: alcohol intake at the start or within 3 months of the start of RT. If available, also state amount: < 21 u/w or > 21 u/w. • Previous: stopped alcohol intake at least more than 3 months prior to the start of RT • Never: no alcohol intake at all	If available
ECOG performance status at start of RT	• 0: Able to carry normal activity without restriction • 1: Restricted in physically strenous activity but ambulatory and able to carry out light work	If available

[1] Year of birth can be used instead of date of birth. All other dates (e.g. RT start, follow-up) will then need to be adapted accordingly (e.g. 1/1/1950) in order to obtain the actual time difference.

	• 2: Ambulatory and capable of self-care but unable to carry out any work. Up and about for more than 50% of waking hours • 3: Capable only of limited self care; confined to bed or chair more than 50% of waking hours • 4: Completely disabled; cannot carry out any self care; totally confined to bed or chair • Not available	
Xerostomia at baseline	• G0: None • G1 (Mild): Symptomatic (e.g., dry or thick saliva) without significant dietary alteration; unstimulated saliva flow >0.2ml/min • G2 (Moderate): Oral intake alterations (e.g., copious water, other lubricants, diet limited to purees or soft food); unstimulated saliva flow 0.1-0.2 ml/min • G3 (Severe): Inability to adequately aliment orally; tube feedings or TPN indicated, unstimulated saliva flow <0.1 ml/min • not available	If available. Use the NCI CTCAE v4/v5 grading system (both versions are the same)
Xerostomia (grade≥2 at ≥1year post-RT)	Use NCI CTCAE v4/v5 grading system described above or the RTOG grading system for late toxicities described for ORN below.	If available. NCI CTCAE v4/v5 or RTOG (both versions are the same)
Pre-RT dental assessment	Yes / No	
Pre-RT dental extraction	Yes / No	
Teeth extacted	If any extractions performed, number the teeth extracted according to the FDI system	Use the FDI (World Dental Federation) numbering system
Date of pre-RT dental extraction	dd/mm/yyyy	This is needed to calculate the time from extraction to RT start
Pre-RT surgery	Primary RT (no pre-RT surgery) / Post-operative RT (PORT)	
Pre-RT surgery type	If available, describe the type and site of surgery	
RT technique	IMRT / VMAT	Cases treated with 3D conformal RT are excluded from the study

TPS	Specify radiotherapy treatment planning system and algorithm used to produce the clinical treatment plan (e.g. Monaco, Monte Carlo, Eclipse-Acuros XB, Eclipse-AAA, etc.)	This is needed in order to anticipate variations in exported data details between TPS
Dose calculation and reporting	Dm,m / Dw,m / Dw,w	This is needed to identify potential systematic biases in reported absorbed dose
Prescribed dose	in Gy	
Total number of fractions		This is needed to calculate the fraction size
Treatment schedule	Please specify number of fractions per week and how these are distributed (e.g. 5 fractions/week, 1 fraction/day)	
Delivered dose and fractions		If different from prescribed
Mandible volume	in cc	
Dmean parotid glands	in Gy	
Dmean submandibular glands	in Gy	
Chemotherapy type	• None • Carboplatin • Cisplatin • Cetuximab • Other	
ORN	Yes / No	The info below is only relevant to ORN cases
Date or ORN diagnosis	dd/mm/yyyy	This is needed to calculate the time to ORN diagnosis since the end of RT
ORN site	Specify using the notation based upon equivalent teeth	
Primary ORN cause	• Spontaneous • Induced - pre-RT dental extraction • Induced - post-RT dental extractions • Induced - dental infection • Induced - dental implant	

	• Induced – denture • Induced - mandible plate infection	
ORN grade (Notani) at diagnosis	• Stage I: ORN confined to alveolar bone • Stage II: ORN limited to the alveolar bone and/or above the level of the inferior alveolar canal • Stage III: ORN under the lower part of the inferior alveolar canal, with fistula or bone fracture	Use the Notani grading system
Peak ORN grade (Notani)	(same as above)	Use the Notani grading system
Date of peak grade (Notani)	dd/mm/yyyy	
ORN grade (CTCAE v4) at diagnosis	• Grade 1: Asymptomatic; clinical or diagnostic observations only; intervention not indicated • Grade 2: Symptomatic; medical intervention indicated (e.g., topical agents); limiting instrumental ADL • Grade 3: Severe symptoms; limiting self-care ADL; elective operative intervention indicated • Grade 4: Life-threatening consequences; urgent intervention indicated • Grade 5: Death	Use the CTCAE v4 grading system
Peak ORN grade (CTCAE v4)	(same as above)	Use the CTCAE v4 grading system
Date of peak grade (CTCAE v4)	dd/mm/yyyy	